DOCUMENT RESUME

ED 310 156	TM 013 772
TITLE	Planning Papers for the National Longitudinal Study of Chapter 1.
INSTITUTION	Policy Studies Associates, Inc., Washington, DC.
SPONS AGENCY	Department of Education, Washington, DC. Office of Planning, Budget, and Evaluation.
PUB DATE	May 89
CONTRACT	300-85-0103
NOTE	199p.
PUB TYPE	Collected Works - General (020) -- Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC08 Plus Postage.
DESCRIPTORS	*Compensatory Education; *Educational Planning; Elementary Secondary Education; *Longitudinal Studies; Long Range Planning; *National Programs; *Program Evaluation; *Research Design; Research Methodology
IDENTIFIERS	*Elementary Secondary Education Act Title I

ABSTRACT

The Elementary and Secondary School Improvement Amendments of 1988 (P.L. 100-297) require the U.S. Department of Education to conduct a national longitudinal study of Chapter 1 of the Elementary Secondary Education Act (ESEA). The department commissioned the selection of experts qualified to provide design suggestions and advice for the national longitudinal study and related activities. This collection of 12 planning papers was developed in association with the longitudinal evaluation of the compensatory education program mandated by the ESEA, and the papers represent the experts' reviews of their planning briefings. The papers include: (1) "Synthesis of Planning Papers" (Elizabeth R. Reisner); (2) "Designing a Chapter 1 Study: Implications from Research on Preschool Education" (Steve Barnett); (3) "Adjoining Randomized Experiments to Longitudinal Surveys in Chapter I Evaluation: Satellite Policy" (Robert F. Boruch); (4) "Use of Comparison Groups in the Evaluation of Chapter 1" (Edward C. Bryant); (5) "Design Proposals for Study of Chapter 1 Programs and Their Effects" (James S. Coleman); (6) Design for a National Longitudinal Study of Chapter 1" (William W. Cooley); (7) "Issues in Designing a National Study of Compensatory Education" (Gary Echternacht); (8) "A Discussion of Some Statistical Sampling Issues Related to the Proposed Chapter 1 Longitudinal Study" (Martin R. Frankel); (9) "Issues in Longitudinal Analyses of Chapter 1 Data" (Joy A. Frechtling); (10) "Longitudinal Analysis of Student Achievement Data: Issues for Chapter 1 Evaluation" (David Rogosa); (11) "The National Longitudinal Study of Chapter 1: Design Considerations in Promoting Study Usefulness to Practitioners" (Robert E. Slavin); and (12) "Thoughts on the Chapter I Longitudinal Evaluation Design" (Marshall Smith). (TJH)

Planning Papers for the National Longitudinal Study of Chapter 1

Prepared Under Contract by:

Policy Studies Associates, Inc.
Washington, D.C.

Contract No. 300-85-0103

U.S. DEPARTMENT OF EDUCATION • OFFICE OF PLANNING, BUDGET & EVALUATION

POLICY STUDIES ASSOCIATES, INC.
1718 CONNECTICUT AVENUE, N. W. • SUITE 400 • WASHINGTON, D. C. 20009 • (202) 939-9780

PLANNING PAPERS FOR THE NATIONAL LONGITUDINAL STUDY OF CHAPTER 1

May 1989

Prepared under U.S. Department of Education Contract 300-85-0103
Office of Planning, Budget and Evaluation

TABLE OF CONTENTS

Page

Synthesis of Planning Papers
Elizabeth R. Reisner 1

Designing a Chapter 1 Study: Implications from Research on Preschool Education
Steve Barnett 17

Adjoining Randomized Experiments to Longitudinal Surveys in Chapter 1 Evaluation: Satellite Policy
Robert F. Boruch 39

Use of Comparison Groups in the Evaluation of Chapter 1
Edward C. Bryant 83

Design Proposals for Study of Chapter 1 Programs and Their Effects
James S. Coleman 95

Design for a National Longitudinal Study of Chapter 1
William W. Cooley 107

Issues in Designing a National Study of Compensatory Education
Gary Echternacht 121

A Discussion of Some Statistical Sampling Issues Related to the Proposed Chapter 1 Longitudinal Study
Martin R. Frankel 131

Issues in Longitudinal Analyses of Chapter 1 Data
Joy A. Frechtling 143

Longitudinal Analysis of Student Achievement Data: Issues for Chapter 1 Evaluation
David Rogosa 151

The National Longitudinal Study of Chapter 1: Design Considerations in Promoting Study Usefulness to Practitioners
Robert E. Slavin 181

Thoughts on the Chapter I Longitudinal Evaluation Design
Marshall Smith 191

Synthesis of Planning Papers

Elizabeth R. Reisner

The Augustus F. Hawkins - Robert T. Stafford Elementary and Secondary School Improvement Amendments of 1988 (P.L. 100-297) require the U.S. Department of Education (ED) to conduct a national longitudinal study of Chapter 1 of the Elementary and Secondary Education Act. The statutory provisions are as follows (Section 1462):

> (a) National Longitudinal Study--The Secretary shall contract with a qualified organization or agency to conduct a national longitudinal study of eligible children participating in programs under this chapter. The study shall assess the impact of participation by such children in Chapter 1 programs until they are 18 years of age. The study shall compare educational achievement of those children with significant participation in Chapter 1 programs and comparable children who did not receive Chapter 1 services. Such study shall consider the correlations between participation in programs under this chapter and academic achievement, delinquency rates, truancy, school dropout rates, employment and earnings. and enrollment in postsecondary education. They study shall be conducted throughout the country in urban, rural, and suburban areas and shall be of sufficient size and scope to assess and evaluate the effect of the program in all regions of the nation.
>
> (b) Follow-up--The agency or organization with which the Secretary has entered a contract under subsection (a) shall conduct a follow-up of the initial survey which shall include a periodic update on the participation and achievement of a representative group of children who participated in the initial study. Such follow-up shall evaluate the effects of participation until such children are 25 years of age.
>
> (c) Report--A final report summarizing the findings of the study shall be submitted to the appropriate committees of the Congress not later than January 1, 1997; an interim report shall be so submitted not later than January 1, 1993.

Because this mandate presents certain technical challenges, officials of the ED Planning and Evaluation Service (PES) invited experts in research design and educational evaluation to present briefings and prepare issue papers on selected topics related to the implementation of this study. This paper describes the process used to obtain the contributions of these experts and summarizes their key observations and suggestions.

Arrangements for the Briefings

To select and retain experts qualified to provide design suggestions for the national longitudinal study and to conduct other related activities, PES issued a task order in June 1988 to Policy Studies Associates, Inc., (PSA) under the Data Analysis Support Center contract (300-85-0103). The first activity under the task order was to assist PES in identifying areas in which ED needed assistance before developing a research plan for the national longitudinal study. The topics that PES selected were as follows:

- Alternative approaches to the collection and analysis of data on Chapter 1 services;
- Measurement of the outcomes of compensatory education;
- Longitudinal analysis of program effects on students;
- Sampling issues, including procedures concerned with regional representativeness, attrition, and dispersion;
- Identification of suitable comparison groups;
- Collection and analysis of retrospective data;
- Use of synthetic cohorts (e.g., the use of comparable groups of children who differ mainly by age, in order to create an artificially "longitudinal" view of student change across time);
- Feasibility of implementing a demonstration strategy to complement the longitudinal study;
- Design considerations in promoting the usefulness of the study's findings to Chapter 1 practitioners; and
- Relevant lessons from long-term longitudinal evaluations of preschool interventions.

Working with PSA, PES selected experts qualified to provide advice in these areas; PSA then contacted each of them and arranged for their participation in the planning process. Twelve individuals agreed to provide assistance. The list below presents the experts and the areas in which PES expected each to provide special assistance. The date of each briefing is also indicated.[1]

Steve Barnett, Utah State University (December 14)
Longitudinal analysis of program effects
Lessons from preschool studies

Robert F. Boruch, Northwestern University (November 10)
Longitudinal analysis of program effects
Comparison groups

Edward C. Bryant, Westat, Inc. (September 27)
Sampling issues
Comparison groups

James S. Coleman, University of Chicago (December 12)
General design issues
Synthetic cohorts

William W. Cooley, University of Pittsburgh (September 30)
Use of data on differing programmatic services
Measurement of compensatory education outcomes

Martin R. Frankel, National Opinion Research Corporation
(September 29)
Sampling issues
Synthetic cohorts

Joy A. Frechtling, Montgomery County (MD) Public Schools (August 8)
Collection and analysis of retrospective data

Craig Ramey, University of North Carolina (December 19)
Longitudinal analysis of program effects
Lessons from preschool studies

David Rogosa, Stanford University (September 23)
Longitudinal analysis of program effects

[1] Experts who provided briefings later in this process had the opportunity to read both the minutes of the earlier briefings and the issue papers that had already been submitted. The later presentations and papers could thus build on what had already been accomplished in the planning process.

Robert E. Slavin, The Johns Hopkins University (August 1)
Usefulness of the study to Chapter 1 practitioners

Marshall Smith, Stanford University (September 23)
Uso of data on differing programmatic services
General design issues

Michael Timpane, Teachers College of Columbia University (September 23)
Design of a demonstration strategy to complement the longitudinal study

Each briefing lasted about two hours. In addition to PES, other federal offices represented in at least some of the briefings were (1) the ED Office of Compensatory Education Programs (which administers the Chapter 1 program), (2) ED Budget Service, (3) ED National Center for Education Statistics, (4) Administration on Children, Youth and Families of the U.S. Department of Health and Human Services (which administers Head Start), and (5) Office of Management and Budget. Within a month or so after the briefing, each of the experts submitted an issue paper that reviewed the topics on which the briefing had focused; these papers are presented in this volume.

In addition to the papers obtained through this process, PSA also received a paper prepared by Gary Echternacht of Educational Testing Service. Because the paper addresses several of the topics reviewed here, it has been included in this volume.

The next sections of this paper summarize the experts' suggestions and observations on five issues, identified by PES as particularly important in the design of the study. The review draws mainly from the issue papers, except in the case of Ramey and Timpane, whose presentation remarks are the basis for our references here.

Objectives of the National Longitudinal Study

An overriding reaction of several researchers (including Coleman, Smith, and Timpane) was the importance of the research opportunity presented by the legislative mandate. According to these commenters, the mandated research offers the chance to expand our collective knowledge about compensatory education and. more generally, to learn how to improve the educational services delivered to disadvantaged children.

Most of the experts advised ED to set two major objectives for the Chapter 1 study. The first objective, which closely reflects the language of the mandate, would be to demonstrate whether and how Chapter 1-supported services make a difference in the educational and social development of program participants. Smith links this objective to the legislative interest generated by the longitudinal studies of preschool programs (especially the Perry Preschool Study); he interprets the underlying intent of the current mandate as being, in part, to determine whether significant participation in Chapter 1 has the same positive long-term influence on beha ior as does participation in early childhood programs.

The experts vary in their recommendations for accomplishing this objective. Their preferred methods include the following:

- A large-scale longitudinal survey (recommended by Barnett, Boruch, Bryant, Coleman, and Frankel);

- A series of small-scale longitudinal studies in selected school systems (Rogosa and Slavin);

- A single longitudinal study implemented in a small number of sites (Ramey); and

- An "explanatory observational study" (Cooley), which would use an ethnographic approach to identify the problems that Chapter 1 participants experience in school and the ways that Chapter 1 interventions address those problems.

In considering these alternatives, Coleman cautions that the research must be fully defensible on a "hard" scientific basis, which, he believes, argues for a quantitative survey approach. Even though good qualitative methods are available, they may not be as readily defended in the political debate that this research is likely to generate, according to Coleman.

With either a quantitative or qualitative approach, several problems arise in implementing this objective. One problem is the difficulty of comparing service outcomes across students, projects, and curricular approaches. This problem arises because (1) Chapter 1 services do not constitute a uniform treatment (Echternacht, Frechtling, and Rogosa) and (2) students who receive Chapter 1 services experience varying patterns of program participation (Echternacht, Frechtling, Rogosa, and Smith).

The second objective of the research, according to these experts, would be to determine the types of Chapter 1 interventions that generate the largest positive effects on the outcomes of greatest interest to the program (Boruch, Coleman, Cooley, Echternacht, Frechtling, Ramey, Rogosa, Slavin, Smith, and Timpane). This objective suggests that the research should explore why particular interventions are effective with certain populations and under certain conditions. To facilitate such analyses, Coleman and Echternacht suggest that an early step in the research design be development of a scheme for classifying types of educational interventions.

Ramey, Slavin, and Timpane suggest that this objective be addressed through federal sponsorship of what Timpane calls "developmental demonstration" projects, which would be assessed to determine their implementation requirements and costs and their effects on participants. If this strategy were adopted, ED would select a set of promising Chapter 1 approaches, based on clear evidence of previous success and sound theoretical underpinnings, and would then provide support for the program developers to implement their service approaches in several locations under stringent experimental conditions. The government's evaluation contractor would assess and report on the implementation and effects of each of the approaches. Coleman criticizes this strategy on the grounds that "whatever Chapter 1 programs are feasible already exist" and that evaluation of existing programs will provide more valid information than will evaluation of programs implemented for this study, due to the Hawthorne-effect problems of the latter. Ramey notes that his experience with two such studies has not indicated a Hawthorne effect, however.

Rogosa suggests that ED can learn about effective compensatory-education strategies through a design that consists of systematic longitudinal tests of proven practices (or "strategic variations," to use Coleman's terminology) in a small number of sites. The accumulation of findings across sites would provide the needed external validity for this research.

Boruch provides detailed suggestions for a strategy of adjoining prospective experimental tests of relevant variations of Chapter 1 educational approaches to a national longitudinal survey. Smith endorses this strategy and urges that services in the experimental substudy be carefully planned, monitored, and documented as well as evaluated for effectiveness. Boruch states that combining a longitudinal survey with experimental substudies "capitalizes on the strongest merits of each" and permits causal inferences to be drawn regarding activities and outcomes observed in the longitudinal data. He cites examples of possible experimental tests, which include examination of strategies for sustaining parental involvement, decreasing the incidence of student failure, improving achievement through tutoring, and retaining students in school until graduation.

Comparison of "Children with Significant Participation in Chapter 1" and Similar Children

The legislative provision mandating the longitudinal study requires that it compare the "educational achievement of those children with significant participation in Chapter 1 programs and comparable children who did not receive Chapter 1 services." This provision is intended to permit conclusions about the amount of achievement growth that Chapter 1 students experience that is specifically attributable to program participation--and not, for example, to participation in regular instruction or normal maturation. The legislation anticipates that the researchers will draw these conclusions by comparing students with "significant" Chapter 1 participation and comparable students who did not receive Chapter 1 services.

The experts pointed out the difficulty of identifying students who are comparable to Chapter 1 participants but who do not receive Chapter 1 (or similar) services. Given the broad coverage of compensatory education services nationwide, nonparticipating students (especially those in the elementary grades) with "comparable" levels of educational deprivation will almost always either (1) receive compensatory education services funded from state or local sources or (2) attend schools whose average achievement is relatively high. In either case, these students would not be suitable for comparison purposes.

Several alternatives were identified to remedy this problem:

- Randomly assign students to treatment and control status and provide unrelated services to control students.

  Boruch recommends that small experimental studies be adjoined to the larger longitudinal survey. The small "satellite" studies would randomly assign students to control and treatment status. Students assigned to control status would receive no compensatory education services but might receive other, unrelated services, such as medical care, nutrition supplements, or the services of a social worker--as described by Ramey in connection with an ongoing early childhood study.

- Compare students who receive differing intensities of compensatory education services.

  Because of the legal and ethical difficulties in withholding compensatory education services from educationally deprived children, Barnett, Bryant, and Ramey suggest that the study compare students who receive either intensive or minimal levels of compensatory education services--in what is termed a "dosage" study in medical research, according to Ramey. This strategy may be preferable to the preceding approach because teachers tend to incorporate successful elements of compensatory education into their regular curriculum, according to Echternacht, thus making it virtually impossible to establish true control groups.

- Use statistical methods to estimate the effect of Chapter 1 participation on educational achievement.

  Although some researchers, including Echternacht, believe it is not feasible to use statistical methods for comparing student achievement levels with and without compensatory education, Frechtling suggests that national testing norms be used to estimate levels of student achievement without compensatory education services. Rogosa proposes use of a regression discontinuity design, in which the growth curves of individual students would be aggregated and analyzed to identify the effects of compensatory education interventions (although Rogosa would also require the designation of control students for comparison purposes). Barnett raises doubts about the appropriateness of that analysis procedure in connection with Chapter 1 services, however.

Echternacht suggests that the problems in establishing comparison groups be avoided by asking a different question from the one that the legislation poses. Rather than asking how Chapter 1 students fare in comparison to nonparticipating students, he suggests that the achievement of

Chapter 1 students be examined in relation to an agreed-upon standard of acceptable student performance, which would be defined in terms of the minimum achievement needed to succeed at a particular grade level. Thus, the achievement of Chapter 1 students would be measured in terms of whether they exceeded a predetermined performance threshold, and different instructional strategies would be compared based on their results in moving students over the minimum threshold.

Cutting across these issues and alternatives are several broad recommendations from the experts regarding the comparison of students in the study, as described below:

- Smith notes that the types of planned comparisons should be based on the study's key questions, which may require comparisons of (1) students within the same school, (2) students or groups of students in similar or different schools within a single school system, and (3) students or groups of students in similar schools in separate school systems--or some other comparative arrangement.

- Smith also recommends that Chapter 1 services to sampled students be strictly additive, rather than a substitute for similar services.

- To control for differences in duration of Chapter 1 services, Ramey recommends that Chapter 1 participants who are included in study samples be required to continue receiving Chapter 1 services for a predetermined length of time, in order to ensure some uniformity of participation patterns.

- Because of increasing numbers of mothers in the work force and national pressures to increase preschool-education opportunities, Coleman recommends that the study include participants in Chapter 1 early childhood programs.

In a warning regarding the findings generated by these comparisons, Smith notes that the preschool studies examined the effects of very intensive treatments (i.e., approximately 300 hours over a year's time) in comparison to no treatment at all. In Chapter 1, the only students with that level of supplementary service are those with very serious educational and social problems, and these students aie the least likely of any students to demonstrate positive behaviors regardless of the services provided to them.

Regional Repr·sentativeness of the Study's Findings

The legislation requires that the study "be conducted throughout the country in urban, rural, and suburban areas and . . . be of sufficient size and scope to assess and evaluate the effect of the program in all regions of the nation." Bryant and Frankel discuss the implications of this requirement for the study sample. Frankel reviews trade-offs between the need for precision in the study's findings and the utility of oversampling populations of particular policy interest. He concludes that it is important to determine the analytic requirements of the study before drafting a sampling plan, in order to ensure that the sample permits all relevant policy issues to be adequately addressed.

Bryant reviews the implications of Chapter 1's unique program features for the design of the study sample. He estimates that the longitudinal study will require a total sample of 10,000 persons, in order to meet the legislative requirements for reporting on subgroups.

Cooley discusses the sample that will be required for his proposed "explanatory observational" study. He anticipates that three school systems (urban, suburban, and rural) would be sampled in each of four geographic regions and that a total of 1,200 students would be sampled from among the 12 school systems.

Time Frame for Implementing the Study

All of the experts noted that the short time lines for the study will require ED to adopt creative mechanisms for collecting longitudinal data for Chapter 1 participants to ages 18 and 25. Barnett, Bryant, Coleman, Cooley, Ramey, and Slavin recommend a design that uses overlapping, "linked" cohorts. By obtaining several years of longitudinal data on children of different ages and using statistical methods to link the different cohorts, they state, the study could draw conclusions about the long-term, cumulative effects of Chapter 1 participation and report on these effects in 1993 and 1997, as required by law.

Coleman's design, for example, uses a series of two-year modules, starting with students in the second, fourth, and sixth grades. In addition, he proposes that the Chapter 1 study "piggy-back" on the analyses of longitudinal data collected by the National Educational Longitudinal Study (NELS:88) on students beginning at the eighth grade and by High School and Beyond (HSB) on older youth. Smith also recommends using NELS:88 and HSB to learn about the effects of compensatory education on older youth. In addition, Smith suggests that ED explore the possibility of locating individuals who participated in the longitudinal Sustaining Effects Study of ESEA Title I; they would be roughly 20 to 25 years old now and could provide useful information on the long-term effects of compensatory education.

Coleman discusses ways of making the "links" between the overlapping cohorts and states that "the success of the modular design depends upon being able to piece together one long causal chain from links in that chain," in order to reveal the "paths" through which the desired outcomes are achieved--or not achieved. He further explains, "This could be conceived as a process of working backwards from the outcomes of interest to those precursor variables that show some effect on these outcomes, from those back to earlier precursor variables, and finally back to examination of the program variables on the early precursor," thus determining what variables should be examined at various ages. Rogosa endorses a similar approach to the design and analysis of data from linked cohorts. Bryant discusses the special technical challenges entailed in the use of this method of assembling a multi-year longitudinal study from short-term longitudinal measurements.

PES asked Frechtling and Slavin to comment on the feasibility of using school records as the basis for generating retrospective data on Chapter 1 participants and nonparticipants. Frechtling states that it may be possible to identify students who received Chapter 1 five to seven years previously, but these data are not likely to include descriptions of the types or amounts of supplementary services, including compensatory education, that these Chapter 1 participants received. Slavin also expresses doubt that retrospective data could provide the amount of instructional detail that will be needed by researchers in a national longitudinal study of Chapter 1.

Chapter 1 Outcomes To Be Assessed in the Study

The legislative mandate identifies the outcomes that the study is intended to measure; they are "academic achievement, delinquency rates, truancy, school dropout rates, employment and earnings, and enrollment in postsecondary education." Within this framework, however, considerable leeway exists for determining what intervening variables will be most important to assess and what specific measures will best capture student performance in these areas.

Before specific measures are selected, however, Ramey, Rogosa, and Smith urge that the study develop (or adopt) a theory of how Chapter 1 affects the growth and development of students. Ramey says that this theory should encompass both cognitive and affective domains and could be drawn from an examination of exemplary Chapter 1 projects. The theory can then serve as the basis for future design decisions, especially regarding variables to be investigated in the study.

The experts varied in their perspectives as to which outcome measures would be most important in the study. While all of them acknowledged the importance of student achievement as a central outcome, they differed in their views as to how it should be measured and whether there are other intervening variables that warrant equal attention for measurement purposes. For example, Slavin would rely on a common standardized test to measure the academic achievement of Chapter 1 participants, including higher-order thinking skills, reading, and writing. Cooley would focus directly on achievement but would measure it using report card grades, which he says are better indicators of school success than are standardized test scores. In addition to achievement, Frechtling would also measure students' success or failure in school by examining patterns of grade retention, participation in extracurricular activities, and placement in special education. Similarly, Ramey would measure achievement, grade retention, and whether students dropped out of school.

Echternacht would measure the performance of Chapter 1 students using a multidimensional construct consisting of standardized test scores, report card grades, and sense of self-efficacy. As indicated earlier, he would set a threshold level of performance and use this construct to measure whether students attained that minimum.

Smith would also measure self-efficacy, which he describes as a good proxy for the positive behaviors that Chapter 1 is intended to stimulate, including high school graduation and postsecondaıy enrollment. Smith also proposes measuring grade retention because of its value as a predictor of whether a student will drop out of school. Similarly, he suggests that early deviant behavior be measured as a prediction of later delinquency. Rogosa summarizes this line of thinking by stating that the study "should assess proximal effects on attributes (e.g., motivation, attendance) that have obvious impacts on longer-term effects."

Barnett urges that the study collect an extremely broad range of outcome data for three reasons. First, ". . . it may be necessary to try to estimate the linkages from grade to grade in order to estimate long-term outcomes using data on overlapping cohorts," which may require many types of data. Second, "the preschool studies revealed a very broad range of effects, not all of which were expected." Third, "the public is interested in real-life outcomes--whether students drop out, get pregnant, stay out of jail, or get a job, not how many points they gain on a test."

In addition to discussing outcomes and measures applicable to students, Ramey also describes several specific types of information that should be obtained from students' parents, in order to determine how program participation may have affected their perceptions of their children's needs and opportunities. In particular, he suggests that parents be asked about their education goals for their children, their understanding of their children's school experiences, the needs they perceive their children to have, their satisfaction with the services their children receive, and their involvement in educational activities at home and at school.

Designing a Chapter I Study: Implications from Research on Preschool Education

Steve Barnett, Ph.D
Early Intervention Research Institute
and Adjunct Assistant Professor of Economics
Utah State University

*Prepared for
U.S. Department of Education
Planning and Evaluation Service*

Designing a Chapter I Study: Implications from Research on Preschool Education

The Congressional goals for a new Chapter I study seem to derive in large part from the perceived success of research on compensatory preschool education, particularly the Perry Preschool study and other studies that have provided very long-term evidence on cost-effectiveness based upon experimental and quasi-experimental designs. Thus, it is worth considering what can be learned from the research on preschool compensatory education that can be used to inform a major research effort on Chapter I programs. This paper addresses three sets of important issues for a Chapter I research plan from this perspective: overall design, the kinds of data to be collected, and methods of data collection. To some extent, these sets of issues are interrelated so that some overlap is unavoidable from section to section.

Overall Design

One source of the influence of the preschool studies on policy makers' and the public's perceptions of compensatory preschool education has been the strength of their designs. This is especially true for the Perry Preschool study, which was prospective, very long-term, and used random assignment to preschool and no-preschool groups (Berrueta-Clement et al., 1984). This design is easy to understand and does not require highly complicated statistical analyses to interpret the data. It provides a strong demonstration of causality--the researcher controls the treatment, and many of the alternative explanatory variables are ruled out as causal. In addition, since the children in the Perry Preschool study entered in waves over 5 years, the potential for unusual events to affect the results (without detection) was greatly reduced.

(A brief description of the Perry Preschool study and its results is included as an appendix.)

Another source of influence is the number of preschool studies with similar findings. Although the preschool studies tend to be small (the Perry Preschool n = 123), there are about 5-10 good studies (depending on how you define this) that produce similar results with a range of different kinds of programs in a variety of settings. This has produced confidence that the results are generalizable even though the studies do not exactly provide a representative sample of preschool programs, locations, and disadvantaged children. Of course, there is a considerable jump from the conclusion that programs like those in the preschool studies are successful to the conclusion that the preschool programs currently run by federal, state, and local governments are successful.

It is recognized that the Chapter I study involves constraints that may not have been present for the preschool studies. Thus, it may well be that no single design can accomplish all of the desired goals. Nevertheless, it would be extremely wasteful to miss an opportunity to include at least some research components that make use of random assignment in a prospective longitudinal study. On a large scale, this could be done by funding perhaps six studies on the scale of the preschool experiments (125-200 students) of different types of programs for various ages at different locations around the country. Separate authorization for program funding could be provided that allow random assignment. On a smaller scale, perhaps only one or two studies would use random assignment to address specific issues. The primary advantage of these multiple studies if they were carried out by different researchers would be independent replication of findings. At the same time, the preschool studies suggest that the limits on generalization would not be fatal. Linking these to a larger study (that uses another design) would address the generalization issue directly.

It should be noted that random assignment is not just an issue at the student level. In choosing a design, one should be aware that school, classroom (teacher), and student factors must all be considered. Classrooms could be randomly assigned or even schools, provided enough schools or classrooms were involved, whether or not students are randomly assigned to treatments within these larger units. It is sometimes overlooked that random assignment of children to two alternative programs does not assure complete disentanglement of the program from other elements when classrooms/teachers and schools (or communities) are not randomized. In the simplest case of a single teacher for each alternative program, the teacher effect cannot be separated from the program effect.

It is sometimes objected that random assignment to compensatory education programs is unethical. However, it seems more reasonable that it is unethical only if it is known that one treatment is better than another or if the researchers have not obtained the fully informed consent of the study participants. In my view, this should not be a problem for the study of Chapter I services. It is possible that some of these services are ineffective or even counterproductive. There are well-known examples in social science research where programs intended to help disadvantaged and handicapped children have been found to make them worse off. Moreover, it is not clear that students are in any sense entitled to the most effective program (which may be only a minimal improvement) regardless of the cost. Of course, it may be that some children are legally entitled to some Chapter I services whether they are effective or not, and this presents problems for random assignment. The legal problem can be overcome by obtaining special funding for programs to be included in the study. If this is not possible, it might be preferable on legal grounds to compare two alternative programs, one of which is considered minimal. The problem here is that there _is_ an ethical concern with offering a minimal program if it is believed to be ineffective. On the other hand, if the possibility is held open that the minimal program has a meaningful effect, then estimation of the difference between the program and no-program is potentially confounded.

There are several alternatives to random assignment. None of them are as strong, but some are better than others. One of the better approaches is the regression-discontinuity design in which students are strictly assigned to alternatives based on a score. Barnow, Cain, and Goldberger (1980) have shown that the regression-discontinuity design uses information over the full range of observations, can deal with nonlinearities (interaction effects), and that no randomization of ties at the borderline is needed. The disadvantages relative to random assignment are that it is necessary to specify the correct functional form of the equation estimated in order to correctly estimate the effect over the full range of observations, and a larger sample size is required (3 times larger, according to one source).

Another alternative with some promise is the use of a before and after design with school-wide (or grade-wide) programs. In this design, one compares the performance of cohorts that pass through after a new program has been instituted with the performance of previous cohorts. This design is strengthened if it is instituted in several successive years at subdivisions of a fairly homogeneous area (ruling out other historical changes as a cause) or at multiple sites in different areas. As with random assignment, the systematic manipulation of the treatment variable increases confidence that observed differences in outcomes are caused by the treatment.

The weakest alternatives are simple before and after designs, matching of participating children with nonparticipants, and matching of schools. In preschool research, such designs have produced results that are sometimes hard to believe. In the Westinghouse study of Head Start, which used a matching design, it was not just that Head Start was "found" not to work, but that it was found to decrease children's school success. There are two common problems with matching. One is that in order to do it, the number of variables is usually limited severely and the groups end up being different on some variable. The other is that we are usually concerned that some unmeasurable or difficult-to-measure variables (for example, potential for school success or predisposition to hard work) are correlated with selection to the treatment groups. Matching schools may be better than matching children, but it is questionable whether any such comparison will ever satisfy the general public, much less skeptics and program critics.

Recently, statistical techniques have been developed to address the problem of selection bias. If, as is the case with Chapter I programs, either self-selection or administrative selection leads program participants to differ from nonparticipants, then ordinary multiple regression or ANCOVA produces biased estimates of the treatment effect. Thus, researchers have attempted to create more complex statistical procedures that provide consistent estimators of the treatment effect (i.e., bias tends toward zero as the sample size tends toward infinity). A major area for the development of statistical approaches to correction of selection bias has been in labor economics research on the effects of emp'·ymment training programs (Heckman & Robb, 1985), but related approaches have been developed in other areas of applied statistics.

In theory, these models for eliminating selection bias can provide consistent and even asymptotically efficient estimates of the effects of alternative educational treatments. In practice, however, it is very difficult to determine if the complex assumptions are met that assure the estimates have desirable properties. Unlike the regression-discontinuity design, the actual selection rule is not usually known and must be guessed at to some extent. In addition to the assumption that the functional form is known (which is also needed to get the most information out of the regression discontinuity design), there are additional assumptions that are not necessarily met and can be difficult to test (Heckman & Robb, 1985).

In my own experience with these models (applying them to the Westinghouse data, for example), I have encountered several practical difficulties--failure of the full information maximum likelihood estimation to converge so that no estimates are obtained, large differences in estimates between alternative estimators, extreme differences in estimates between alternative functional forms (negative and significant v. positive and significant), and identification problems in specifying the selection model and treatment effects equation. Barnow, Cain, and Goldberger (1980) indicate that robustness of the techniques to non-normality of disturbances is a serious problem. It seems to me that matching may introduce additional problems for the elimination of selection bias, because it may attenuate the links between observed variables and the selection rule. Thus, these techniques may be more productive using data from natural variation. Finally, the problem of estimation in the presence of selection becomes even more complex if the selection rule varies from place to place within the sample, which tends to be the case for Chapter I.

No matter what the other characteristics of the design, a prospective study is desirable. A major reason for this is that there are some data that simply cannot be collected (or are collected with more error) in a retrospective study--description of the actual program that was experienced, description of the alternative experiences, family attitudes and behavior at the time of entry to the program, behavioral response to the program, and teacher ratings are all examples. Moreover, matching on, or using in statistical analysis. variables that were not measured at the time of selection increases the amount of error involved. Marital status of parents, parents' employment, and family income measured at age 15 are not good proxies for the values of those variables at age 6 or 7.

The major difficulty in attempting a prospective study is the length of time required to produce the results. Of course, if Head Start had commissioned Westinghouse to produce a prospective study with random assignment in 1968, they would have 20 years of solid data today. Instead, they have little strong evidence about Head Start per se. One way to produce results in a shorter time without sacrificing the prospective design is to have overlapping cohorts. For example, if Congress could be persuaded to provide two 6-year funding periods (two 5-year funding periods are too snort to produce the results that they want without an additional cohort), the cohorts could be as follows:

	Period 1						Period 2						Period 3				
Year	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5
	Ages																
Cohort 1	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Cohort 2	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20		
Cohort 3	11	12	13	14	15	16	17	18	19	20	21	22					
Cohort 4	16	17	18	19	20	21	22	23	24	25	26	27					

Using this kind of design, the study could produce what Congress has asked with only four cohorts over the course of 12 years. The design shown above provides for one cohort that begins at preschool (age 4) and one that begins at first grade (age 6). There are two other overlapping cohorts that link with the first grade cohort. These data could be used to build longitudinal models that would predict effects from preschool and early elementary Chapter I programs to adulthood. As will be explained below, it is necessary to go beyond age 19 to get the information that Congress requested for age 18 and necessary to go beyond age 25 to get the age 25 data. At the end of this funding cycle, it would be extremely productive to secure another 5-year funding cycle in order to obtain true longitudinal data on two cohorts through age 19.

What Kinds of Data Should be Collected?

Given the wide range of data that have been asked for, it may be sensible to conduct multiple studies, or at least sub-studies within a larger framework for research, that make different choices about the types of variables for which data are collected and the frequency with which data are collected. For example, in the illustrative overlapping cohort design given above, some data might be collected every year, but interviews with the subjects and their families might be conducted only at entry to the study, and ages 15, 19, and 25. Similarly, official records data (on schooling, crime and delinquency, or welfare) might be collected only at those intervals. One reason for the varying data collection activities across several different studies is that the dense data collection activities conducted in the Perry Preschool study may not be feasible on a much larger scale.

Data on Child Outcomes

There are at least three reasons to collect an extremely broad range of data on child outcomes that extends far beyond tests of how much children have learned. One is that theory is not a strong guide to the kinds and magnitudes of outcomes to look for in terms of immediate or long-term effects, especially since it may be necessary to try to estimate the linkages from grade to grade in order to estimate long-term outcomes using data on overlapping cohorts. Another is that the preschool studies revealed a very broad range of effects, not all of which were expected. Unexpected relationships among variables were found as well. For example, most preschool studies show only temporary effects on IQ, but permanent effects on educational attainment. Some researchers claim to have found that preschool curriculum differences that do not appear to produce long-term educational differences influence later delinquency (Schweinhart, Weikart, & Larner, 1987). Finally, the public is interested in real-life outcomes--whether students drop out, get pregnant, stay out of jail, or get a job, not how many points they gain on a test.

Among the data on children that should be collected in at least some parts of the study are: IQ; achievement test scores; school records information on grades; special education placement; grade retention; teacher ratings (of motivation, ability, conduct); students' self-reports (of motivation, effort, aspirations, expectations, relationship with parents, perceptions of parents' aspirations and expectations, perceptions of ability, and values); involvement in delinquency and crime including drugs, property crimes, violence, gangs, school-related vandalism and violence (as victims as well as perpetrators); extra-curricular activities associated with school and church; other social activities; work; earnings; welfare; pregnancy; children; and marriage and other relationships. Most of this information can be obtained by self-report. Some can be obtained from official records and from parents.

For many kinds of data, it would be interesting to compare responses from different sources. For some kinds of data, like crime and delinquency, it is indispensable to have multiple sources, because differences in responses are expected and data from a single source would be disputed by advocates of one source over another. Obviously, multiple sources may provide a back-up that decreases the amount of missing data for a particular variable as well. Given the number of variables in which we are interested and the limits of our knowledge about their interactions, it would be foolish to try to select any single outcome as "the" variable for research to focus on.

One of the most persistent questions about preschool compensatory education has been how it has its effects on school success. It is frequently asserted that effects on motivation were the mechanism. This cannot be established clearly from the data, however. The preschool programs increased IQ (at least temporarily), achievement, and teacher ratings of ability. In later years, interviews revealed that preschool was associated with greater aspirations and expectations by parents and children and perhaps with greater motivation and effort as well. It is quite possible that the preschool group's superior ability and performance in school led to those later differences and not vice versa. If there is any hope of unravelling effects on motivation, effort, and ability, then data collection must begin before the Chapter I program has had an opportunity to affect these variables and be repeated at least annually during the program and the first few years following it. Furthermore, it will be necessary to have separate measures of expectations, aspirations, motivation, effort, ability, and achievement. At least in some studies, IQ and standardized achievement tests have been considered to be the same or at least substantially overlapping. The preschool studies indicate that they should be treated very differently. For measures like motivation and effort that we may not be very good at measuring, it may be necessary to have multiple measures.

Data on Chapter I Programs

The data collected on the programs may be considered as important as the data collected on the child outcomes, although this is a traditionally neglected area of research in the preschool studies. The result for preschool studies is that relatively little is known about why and how preschool education produces its effects or how program effectiveness varies with program characteristics. This is not a merely academic problem, because variations in program characteristics have important effects on program costs. Thus, it is recommended that the research obtain program descriptions that are adequate for replication. These descriptions should provide not just information on educational content, but on the resources needed to carry out the program so that administrators can determine what the program would cost to implement elsewhere.

Data collection on programs may be considered to have two purposes for which different data collection strategies are needed. We want to be able to describe programs for those interested in reproducing them, and we want to be able to describe the experiences of individual children. My experience in conducting public school preschool studies is that it is difficult to get teachers and administrators to use universal definitions in description and that school personnel tend to view themselves as doing what they are supposed to do or believe that they should do, even if they are not. (One reason is that they may fail to understand what it is that they are attempting). Thus, direct observation is essential to the collection of accurate descriptions of the program and children's experiences. It can be practical and reliable as well.

No matter what design is used, it will be critical to know what children's educational experiences were, with and without the Chapter I programs that are to be studied. This means that the experiences of the comparison group have to be measured (preferably, observed). This problem of measuring the difference in treatment has been easier to some extent in preschool research, where the difference is greater. It has been widely assumed that children who attend preschool programs have more (and better) educational experiences than those who do not, and measurement of the no-preschool experience has been largely ignored. This assumption is not so widely held for Chapter I. It will also be necessary to have some descriptions of the subsequent academic experiences of the child, as these cannot be assumed to be the same (it is hoped that they are not) for treatment and no-treatment subjects. These descriptions can be less detailed than those of the experiences during the "treatment period."

Measuring the treatment for Chapter I and comparison groups can mean a variety of things. At one level, it means determining attendance, at another measuring the time spent in different environments, at another observing the activities of teacher and child in the classroom. A number of systems are available for systematically coding classroom experiences that produce a quantitative measure of treatment. These vary with the philosophical orientation of the developer: one might assess the type of teacher-child interaction or degree to which the teacher creates learning opportunities in activities as diverse as free-play, reading to a group, and snack time; another might count minutes of direct instruction or number of responses to the instructor. More explicitly qualitative approaches to observation are also available.

Providing information for replication and documentation of treatment are not the only reasons to collect the program data needed to estimate program costs. We have enough experience with compensatory education to begin to address questions of cost-effectiveness. It should be obvious that it is more useful to compare the costs and effects of educational programs than their effects alone. This is one of the characteristics of the Perry Preschool study that captured the public's imagination: it showed the program to be economically sensible not just educationally effective. The estimation of costs (and where possible benefits) and the conduct of a cost-effectiveness analysis entails only a very small increase in research costs over the collection and analysis of the underlying program and effects data that are required in any case. However, the experience with economic analyses of preschool education suggests that the participation of economists specializing in this field is needed to produce sound research (Levin, 1983).

Some thought should be given to the types of programs that should be studied. This will differ depending on whether the objective of the study is to describe the impact of what is going on now or what is possible (or both). It is probably safe to say the study should at least include the programs/program variations that are the most theoretically and practically interesting. A study should take into account the prior views on the programs of both practitioners and researchers. This may not only affect the reception of the study's findings, but it can lead to more useful research. It has been an advantage for preschool research that a wide range of popular alternatives have been studied. It is also a consideration that most of the preschool interventions that show the greatest impact are relatively impressive programs (in terms of duration, intensity, staff, and cost), although there are sometimes surprises (Consortium for Longitudinal Studies, 1983). Clearly, it makes sense to study the separate effects of Chapter I preschool and school programs (and perhaps even to compare them), and to make sure that their effects are not confounded in some studies. The possibility of interactions between some preschool and school programs might be worth investigating, too. It should be noted that the relatively small number of preschool programs for which there is longitudinal research and the nonrandom selection of programs for study has not been a major consideration in drawing policy conclusions from the preschool literature.

Several suggestions have been made regarding criteria for selecting programs to be studied. One is that programs for which some evidence of efficacy is already available be selected. The problem with this suggestion is that it capitalizes on chance. Programs that seem to be more effective because of good luck or because the treatment group just happened to be superior to the comparison group to begin with will be selected for study, and the estimates of program effect will be misleading. Another is that only exemplary programs be selected, where exemplary may be defined in terms of the resources and effort required or expert judgment. Certainly, it would be a mistake not to include exemplary programs. The study would be extremely limited in its usefulness if it included only programs that experts considered unlikely to produce substantive educational benefits. On the other hand, some preschool research suggests that "typical" programs can be effective, perhaps as effective as excellent programs in terms of basic educational outcomes (Barnett, Frede, Mobasher, & Mohr, 1988). Also, the inclusion of some poor programs in the study provides a check on the design and method. It would be reassuring for a study to find that exemplary programs were effective and poor programs were not. If a study found that very poorly executed, minimal interventions produced results equal to those of well-implemented, intensive interventions, questions would be raised about potential bias in the design and methods of analysis.

Data on Family, School, and Community

A key insight from the ecological approach to education and human development (e.g., Bronfenbrenner, 1979) is that the effects of an intervention like preschool education or Chapter I programs can only be understood in the context of the larger environment. The relevant environment is not just the one that is concurrent with the intervention, but prior and subsequent environments can be equally important. To take this into account, it is essential that a study measure characteristics of the most significant systems that make up the social environment: family, school, and community. Many of the preschool studies were conducted before the ecological approach was well-developed in education and psychology, and they tended to neglect these systems. This is an important limitation on the generalizability of the findings from the preschool studies. It is nut known how the effects may vary with family, school, or community characteristics. For example, the failure of two of the preschool studies conducted in large cities (the Perry study was conducted in a small town) to find educational benefits for boys may be an indication that persistent effects on boys are more difficult to produce in poor big-city neighborhoods. Similar interactions might be suspected for Chapter I effectiveness.

As the family is considered to be the most important influence on the child's development, it would be logical to devote substantial effort to obtaining information about the families of children in the study. It is relatively easy to obtain information that describes the family structure. It is somewhat more difficult to obtain information on family function. Instruments that measure the home environment may be considered to be indicators of aspects of family function related to children's development and achievement. Also, it would seem to be important to interview parents (or other adults and perhaps even siblings who may influence the child's educational progress) about their aspirations and expectations for the child, attitudes toward school, activities with the child related to education, and involvement in school and school-related activities. A limitation in the preschool research has been the failure to measure these family variables prior to the intervention that is the focus of study (to the extent that they are measurable before school entry). If a goal of the Chapter I program is to increase parent involvement, a major focus on the family would seem warranted to investigate effects on family behavior and interactions with family characteristics (who agrees to participate initially and who drops out later). Programs might lead to differences in parents' efforts, attitudes, and expectations.

As indicated above, the effects of a Chapter I program may also vary with the characteristics of the school and community. In some cases, it may be possible to describe the situation in some detail through available statistics and direct observation. Even in a retrospective study, one might obtain useful information from oral histories of the school and community. Of course, with more children in out-of-home care than when most of the longitudinal preschool studies were begun, it is much more important now to obtain at least indications of the other care arrangements that children experience, before school-age and during the school years (do they have before or after school care by persons other than parents; are they "latch key" kids?).

Methods of Data Collection

While the choice of prospective and retrospective data collection is to some extent tied-up with design, there is some flexibility within most designs, and it is worth considering the implications of time of data collection for data quality as a separate issue. There is no denying that the quality of data tends to be improved by collection as near to the time of interest as possible and by repeated collection at frequent intervals. However, there is always a tradeoff between quality and cost. For some types of data, there is minimal attrition and negligible loss of accuracy over 5 years, perhaps even 10. For example, in the Perry Preschool study, we were able to locate all of the 123 subjects at age 19 even though the most recent contact with any of them was at age 15. We were able to obtain cumulative school records for 112 of 123 at age 19. Experience in other studies suggests that the Perry study benefited from being focused in a relatively small town where most subjects remained in the area for at least the school years. Cooperation with schools, courts, and police also might be more difficult to obtain in larger cities (we "lost" more school records in big-city schools). In general, the more compact the geographic area (and less dense the population of that area), the easier it would be to collect data from people who have to be tracked down over time. It is also worth noting that some education agencies may retain data for shorter periods of time than the Michigan schools we worked with.

There may be some decay in the availability of official records from nonschool agencies over time, and after a while a person's recall on some variables will not be very accurate. Juvenile court records may be destroyed after a certain amount of time. Welfare records may be purged after a certain number of years of inactivity or may only record a certain number of years (perhaps the last 5, for example). Information related to employment may be forgotten quite rapidly for individuals who change jobs frequently, which is likely to be the case for teenagers. Even if they remember the job and their hours, they may be uncertain about the pay and any benefits they may have received.

The potential loss of accurate recall is a reason to collect data on such things as employment, earnings, savings, welfare, delinquency and crime, and attitudes toward education at more than one time during adolescence and the young adult years. In the Perry Preschool study, data were collected at age 15 and 19. The age of data collection suggested in the authorization for the study is highly problematic. At age 18, many of the variables that are to be the subject of research have not yet been affected by the school to post-school transition. Especially with grade retention, many may still be in school. There will be little about post-secondary education, welfare (they may be part of their parent's case until they are adults), and employment. The contractor should have at least a full year beyond high school graduation--two or more would be better--to collect social and economic data on the subjects.

If attrition is to be kept at a reasonable level, it is critical to have a wide time window for data collection. For example, if the intent is to collect data through age 19, the contractor should be allowed to collect data for the subjects up to age 21. It simply takes time to track some subjects down, and 1 year may not be enough to avoid significant attrition. If it is possible, one strategy for extending the time available for most subjects and spreading out the data collection burden (when it is clustered around specific ages) is to have subjects enter the study in waves (25% a year for 4 years). This is the strategy used in the Perry Preschool and Abecedarian studies (Ramey & Campbell, 1987), and I think that it is an important reason for their lack of attrition. When data are collected by wave, all of the data at a given measurement point are not collected in a single year, and much of the time the research team can be looking for difficult and easy-to-find cases at the same time.

The use of existing infrastructure to collect data will help keep costs down and the quality of data up. Schools, police, courts, social service agencies, and employers are all sources of information that can be obtained with informed consent. In order to use these sources, it will be necessary to have a great deal of identifying information about the research subjects. The use of social security numbers for parents and children could make the process easier and more accurate. However, it s my recollection that the Social Security Agency has a policy of not cooperating in research projects. To obtain their cooperation, Congress might have to intervene. Computer matching with school records is a method that is often suggested, but does not always work well--children change the names they use, their parents divorce and remarry, and at older ages they marry and divorce. Also, these school data are kept by grade, not by cohort. When children enter school late or are held back, they fall into different grades than the rest of their cohort. Thus, a cohort has to be built by matching across several grade levels. Similarly, use of existing student testing programs that are administered by grade as the source of achievement data means that children take the same tests at different ages because of grade retention. In using the existing infrastructure to collect data, it is important not to turn responsibility for data collection over to an agency such as the school. There are too many pressures for them to cut corners. Moreover, it is highly likely that if schools collect the data, the rate of attrition and quality of data will be correlated with the quality of the Chapter I program the school runs.

If attrition is to be held down, it is reasonable to provide financial incentives for participants. They are much more apt to respond if they are being paid for the time and effort it takes to show up on time for an interview and to complete the interview. Participants may also be more willing to help the researchers find them at follow-up times. Their friends and family may be more willing to help the researchers find the participants if they know that there is something in it for the participant. They are doing a favor r a friend in that case, not for a research project in which they have no interest. For adolescents, the payment probably does not have to be very high given the kinds of (legal) opportunities they have to make money and the small effort required to cooperate.

References

Barnett, W. S., Frede, E. C., Mobasher, H., & Mohr, P. (1988). The efficacy of public preschool programs and the relationship of program quality to efficacy. Educational Evaluation and Policy Analysis, 10(1), 37-49.

Barnow, B., Cain, G., & Goldberger, A. (1980). Issues in the analysis of selectivity bias In E. Stromsdorfer & G. Farkas (Eds.), Evaluation studies review annual (Vol. 5). Beverly Hills: Sage.

Berrueta-Clement, J. R., Schweinhart, L. J., Barnett, W. S., Epstein, A. S., & Weikart, D. P. (1984). Changed lives: The effects of the Perry Preschool program on youths through age 19. Ypsilanti, MI: High/Scope.

Bronfenbrenner, U. (1979). The ecology of human development: Experiments by nature and design. Cambridge, MA: Harvard University Press.

Consortium for Longitudinal Studies. (1983). As the twig is bent...lasting effects of preschool programs. Hillsdale, NJ: Erlbaum.

Heckman, J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions: An overview. Journal of Econometrics, 30, 239-267.

Levin, H. (1983). Cost-effectiveness: A primer. Beverly Hills, CA; Sage.

Ramey, C. T., & Campbell, F. (1987). The Carolina Abecedarian project. In J. Gallagher & C. T. Ramey (Eds.), The malleability of children. Baltimore, MD: Paul H. Brookes.

Schweinhart, L. J., Weikart, D. P., & Larner, M. B. (1986). Consequences of three preschool curriculum models through age 15. Early Childhood Research Quarterly, 1, 15-46.

TITLE:	Adjoining Randomized Experiments to Long' udinal Surveys in Chapter I Evaluation: Satellite Policy
AUTHOR:	Robert F. Boruch
PREPARED FOR:	U.S. Department of Education, Briefing Seminar on Chapter I Evaluation, November 10, 1988.
REPORT #:	A-488 (Related A-490)
DATE:	12-06-88
STATUS:	Final Draft
TO BE SUBMITTED TO:	AERJ or EEPA, maybe and with addition and co-authorship of Sally Freels

Department of Statistics and Psychology
Northwestern University
Evanston, Illinois 60208

SUMMARY

1. Longitudinal surveys based on well designed probability samples are the best possible approach available to describing growth of individuals and change at the national level. Such surveys often do not yield defensible estimates of the effect of intervention, e.g. of Chapter I programs.

2. Controlled randomized experiments are the best possible approach to estimating relative effects of interventions, program variations, etc. They are often not feasible at the national level however.

3. Coupling controlled randomized tests to longitudinal study can provide both understandings of growth or change and unbiased estimates of what works better in local contexts.

4. A formal policy for coupling experiments to longitudinal study then seems sensible. Such a policy is analogous to research policy in satellite use. The major vehicle for generating information, the satellite, is periodically reoriented and partly dedicated to special experiments and is analogous to a longitudinal study system.

5. The main justification for the proposed satellite policy for Chapter I is scientific and policy relevant: better data to inform policy about how to improve programs. The secondary justifications include: economic ones, e.g. local experiments capitalize well on longitudinal infrastructure; methodological reasons, e.g. learning about how to improve data quality generally; political reasons, notably permitting answers to several questions.

6. Selection of interventions for experimentation should be guided by several criteria: theoretical import of the intervention, empirical support for its promise, propriety of a test, feasibility of implementing both the interventions and the randomized experiment.

7. In Chapter I, replication of exemplary projects may meet all these criteria. The experiments may for example test new ways of sustaining parental involvement, reducing drop-out, decreasing low grades and failures, tutoring, and so on.

8. Executing controlled experiments in Chapter I projects requires resources: well trained researchers and practioners and support for both. Failure of some projects is likely simply because learning how to improve and generating evidence on it is difficult. Assuming a failure rate of 20% for executing the experiment (regardless of program success) is reasonable.

9. Statistical characterization of the target groups (who is eligible, who gets service) etc. is essential for design of the experiments. So is careful literal and statistical description of the processes engendered by the program, e.g. time in Chapter I variation, nature of variation. Both can be generated at least crudely by longitudinal study.

10. Theory will be important in the longitudinal study to estimate, at the macro-level, effects/ The experimental programs will, if based on similar theory, help to adjust statistical vulnerability of the longitudinal work.

11. A major legislative implication of this perspective is that mandates for longitudinal study must also authorize demonstrations, i.e. implementations of new programs, variations, and components.

CONTENTS

1. Introduction

   1.1 Organization
   1.2 Context

2. Definitions

3. A Proposal for Satellite Policy

4. Justifications for a Satellite Policy

   4.1 Scientific and Statistical Rationale
   4.2 Economic Rationale
   4.3 Prophylactic Rationale
   4.4 Calibration Rationale
   4.5 Methodological Rationale
   4.6 Policy and Political Rationale

5. Related Research Policies and Policy Origins

   5.1 Related Policy and Practices
   5.2 Recent Origins
   5.3 Earlier Origins

6. Examples of the Contexts to Which The Proposal is Relevant

   6.1 Multicohort, Multicity Studies of Criminal Behavior
   6.2 Education: High School and Beyond
   6.3 Employment and Training

7. Probable Issues and Options

   7.1 Standards for Adjoining Experiments
   7.2 Adjoining Experiments to Existing Surveys
   7.3 Adjoining Experiments Regardless of Longitudinal Study Type
   7.4 Feasibility and Appropriateness

8. Summary

9. References

Adjoining Randomized Experiments to Longitudinal Surveys In Chapter I Evaluation; Satellite Policy.

Robert F. Boruch
Northwestern University

1. Introduction

There are a variety of ways to enhance the usefulness of longitudinal surveys. In this paper, one such strategy is considered: attaching controlled randomized field experiments periodically to ongoing surveys. Social research policies that encourage coupling the two approaches will make both longitudinal data and experimental data more useful for social science and public policy, decrease the artificial separation of the sample survey and experimentation traditions, and reduce unnecessary debates over policy-relevant data analyses. In short, such a policy would combine the strengths of each design while compensating for their respective analytical and administrative weaknesses.

1.1 Organization

What follows is based heavily on Boruch and Pearson (1988) and Boruch (1975). The material exploits recent work by Blumstein, et al (1986), Farrington (1988), Fienberg and Tanur (1986) and others, and presents some new ideas. The organization is as follows:

2. Definition
3. Proposal for Satellite Policy
4. Justification for a Satellite Policy
5. Related Research Policies and Policy Origins
6. Examples
7. Probable Issues and Options

1.2 Contexts

The stress here on research that is now being planned. The two contexts of special interests are the Chapter I Education Evaluation and the Program in Human Development and Criminal Behavior.

Sections 1461 and 1462 of Public law 100-297 (1988) requires that the Secretary of the U.S. Department of Education employ a longitudinal study to "assess the impact" of children's participation in Chapter I programs relative to "comparable children who did not receive Chapter I services". The outcomes of interest include "academic achievement, delinquency rates, truancy, school dropout rates, employment and earnings...". The U.S. Department of Education, responsible for implementing the statute convened experts to discuss research designs during 1988.

The Program in Human Development and Criminal Behavior, begun in 1988, is supported jointly by a private foundation, the MacArthur Foundation, and a public mission oriented research agency, the National Institute of Justice. The Program's main objective is to learn about how delinquency and criminal careers grow and cease to grow or cease entirely. A secondary object is to learn about how to decrease or delay onset of criminal activity and to increase desistance. The Program initiated discussion of research designs during 1988.

The two efforts are independent. But some similarity in national scope, design, aims, and difficulty invite their joint consiueration at least for illustration's sake. Local contexts, especially randomized tests of specific programs are also considered below, notably the Broward County School Board's t. ; of the A.I.M. program for children at high risk of academic failure.

2. Definitions

Longitudinal surveys are defined here as repeated observations of the same persons or organizations or other entities in the interest of documenting growth and change. A major purpose of such studies is to understand how individuals (or organizations etc.) change over time. Interest may, for example, lie in the growth of children's intellective achievement and how that growth accelerates rapidly during some periods (e.g. early childhood) and accelerates less rapidly in other periods. Or, the interest may lie in variations in level of delinquent activity over some period. When based on well designed national probability samples, such surveys are the best possible approach to statistical characterization of individuals' growth, development, and engagement in various educational and social systems. Excellent compendia of national longitudinal surveys are given in Taeuber and Rockwell (1982) and in Verdonik and Sherrod (1984).

Randomized experiments are defined as settings in which individuals (or organizations, or other units of study) are randomly assigned to one of two alternative regimens. The object of the experiment is to estimate the relative differences among regimens in a way that is unbiased in a logical sense and that permits formal probabilistic statements to be made about one's confidence about the estimates. Interest in long-term differences between what are frequently referred to as "treatment" and "control" groups are often of interest and may engender the repeated observations that characterize longitudinal or panel research designs. A compendium of field experiments is given in Boruch et al. (1978). See also Riecken et al (1974).

The statistical models used to analyze each kind of data usually differ. Heckman and Singer's (1985) edited monograph, for instance, reviews methods of analysis but not the design of such studies. But one can develop analyses that simultaneously exploit contemporary experimental design models and models designed for common panel or longitudinal data (e.g., Boruch, 1975; Fienberg and Tanur (1986; 1987B)).

3. A Proposal for Satellite Policy

The proposal for joining experimental studies to ongoing longitudinal surveys may be stated as follows (amended from Boruch and Pearson, 1988).

Any longitudinal study should be designed so that independently designed experimental studies can be adjoined to the longitudinal surveys so long as (1) the experiment is compatible with the mission of the longitudinal survey, (2) the risks of disruption to the survey can be managed, especially in regard to the time frame, respondent's burden, and institutional cooperation, (3) designated contractors are responsible for oversight of the process, and (4) the experiment involves no appreciable cost to the agency supporting the longitudinal study.

This proposal is analogous to policy on satellite use that have been used by astrophysicists. That is, the satellite, like a longitudinal survey, has a primary monitoring mission and requires considerable resources to place and maintain. Further, scientists can obtain access to part of the satellite periodically for limited, temporary investigation of important scientific questions.

The policy recommended here for longitudinal survey allows the social scientist or evaluator the option of using the infrastructure of the ongoing survey as a vehicle for conducting prospective experimental studies. The proposal also extends a scientific tradition of "data sharing" in the social and behavioral sciences and education research (Fienberg et al., 1985). In particular, it requires that resources be shared: population listings and sampling frames, the organizational vehicles for longitudinal surveys, and so on, not just data.

Adjoining experiments to ongoing longitudinal surveys is likely to be feasible, for example, for only for a few projects, perhaps only one every year or two, because of the difficulty of coupling studies to an already complex longitudinal enterprise. In the case of the Program on Human Development and Criminal Behavior, for example, there might be a half dozen city based longitudinal surveys. Each may involve multiple cohorts of individuals. Any given city might, for a given cohort, also be the site for an experiment joined to the longitudinal study in the interest of testing new regimens that are thought to control delinquency or crime.

In the case of the Chapter I longitudinal surveys, experimental tests of variatiens on Chapter I programs that are thought to be important might be adjoined periodically to the surveys in the interests of understanding what program variations work better. But Chapter i sites will vary in their capacity and willingness to test innovations under controlled conditions. See the examples below.

4. Justifications for a Satellite Policy

Longitudinal surveys have clearly been useful for science and public policy in revealing how individuals (or institutions) change over time. For example, they avoid the logical traps that cross-sectional studies invite in overlooking cohort effects, or in failing to measure and control a wide range of unobserved individual traits that may explain the relationships that are the focus f inquiry. Most important, the surveys can be based on national probability samples that permit one to make generalizations at the national level.

For instance, those with an interest in the kinds of crimes that households encounter could, during the 1970's, rely on the National Crime Survey to estimate monthly encounters of individual families over time in major cities. The High School and Beyond Surveys, begun in 1980 and conducted periodically since then, help to understand the transition from high school to adulthood, , employment, and so on, and the social influences on individual progress.

There are several kinds of justification for adding controlled tests to such a study design.

4.1 Scientific and Statistical Rationale

The mathematical conditions under which longitudinal (nonrandomized) study will fail to yield an unbiased estimate of relative program effects are well understood. Rubin (1978) provides a basic description in the context provides a basic description in the context of education; Campbell and Boruch's (1975) treatment is more rudimentary. Heckman and Robb (1985) provide elaborate description for analysis of both longitudinal and cross-sectional data in an economic context.

Despite remarkable advances in the mathematical aspects of topic, the problem of assuring that mathematical assumptions are tenable remains. Indeed even determining whether assumptions are met can be difficult often impossible, especially where theory is not adequate. All approaches to estimating the effects of intervention based on longitudinal nonrandomized data depend heavily on the assumption that performance of individuals in the absence of the intervention can be estimated accurately.

The assumption is patently suspect to judge from recent empirical comparisons of evaluations based on longitudinal against evaluations based on randomized evaluations. LaLonde (1986), Fraker and Maynard (1987), and Maynard (1987), show how estimates of program effect based on the former have been demonstrably wide of the mark in evaluation of manpower programs.

The economist's work is, in some sense, belated. Research on nonrandomized clinical tests in medicine and on randomized clinical trials has routinely shown differences in results between the two. Boruch and Riecken (1975) gave relevant illustrations.

More recent work by Gray-Donald and Kramer (1988) reiterates the point for research in pediatrics. Observational studies have typically shown a definite association between infant formula supplementation in hospital settings and lower subsequent breast feeding by mothers. The inference has been that supplementation then has an important potentially negative effect. Controlled randomized tests show no such difference, eliminating pediatricians' concerns about supplemental feeding in hospitals.

The point of this and other illustrations is this. Though longitudinal studies may be enormously useful for rational description of growth and change, they cannot be relied on for accurate estimates of the effects of new intervention programs, at least not in the absence of strong theory.

The implications for Chapter I evaluations based solely on longitudinal study are direct and have identified by Cooke (1988) and Smith (1988). the law's demand that Chapter I effects be estimated using only longitudinal study cannot be met without heroic assumptions about children's behavior in the absence of such programs. Such assumptions may be tolerable politically but they are often indefensible scientifically. The implications for longitudinal study of the Program on Human Development and Criminal Behavior are related if indeed the Program seeks to determine how onset of delinquent behavior and desistance can be affected by intervention. They are reiterated by Farrington (1988), Farrington, Ohlin, and Wilson (1988) among others.

A second justification for adjoining experiments to longitudinal study is that the science and technology of randomized field tests of projects has developed more or less independent of the technology of longitudinal surveys. The intellective separation is often sufficient to prevent researchers from thinking about both in designing tests of new programs or in designing longitudinal studies of important topics. There are good scientific reasons to avoid intellective parochialism here and to understand the union of approaches when the opportunity arises.

A third scientific justification stems from Fienberg et al.'s (1985) observation that although major experiments involve collecting longitudinal data, their analysis is often based on dynamic models that were not incorporated into the design of the experiment. The failure to involve these models in design of the survey, they suggest, ultimately leads to less defensible analyses of experimental results. The argument seems sensible. But little formal research on the relative gains and costs of basing designs on analytical models appears to have been undertaken.

The scientific justification for coupling experiments and longitudinal surveys is then to capitalize on the strongest merits of each. That is, one obtains both the information produced by national probability samples - often conducted over a considerable length of time - and the information produced by smaller comparative experiments in which causal inferences are more appropriately deduced. Insofar as the experiments can be adjoined systematically, their generalizability will be enhanced.

4.2 Economic Rationale: Less Costly Policy Experiments

It takes considerable effort to mount high quality longitudinal surveys. It also takes considerable effort to mount randomized tests of policy relevant programs, more effort if we recognize the difficulty of maintaining control over selection of individuals into programs and over program operations. To the extent that an experiment can capitalize on the resources and data of a longitudinal study, the experiment becomes a less costly enterprise.

Recent experiments undertaken by the Broward County School Board's Department of Research (1987) are a related case in point. Their experimental tests of the A.I.M. project for youth at high academic risk capitalizes heavily on a regular system of standardized testing using Iowa Achievement Tests and the infrastructure to which regular testing was based to execute the experiments. The infrastructure was especially useful in tracking the large number of children who migrated from the original six school to 18 schools (Carey, Sutton, Personal Communication, November 11, 1988). In a longitu instance, we might reasonably expect the adjoined experiment to exploit one or more ot the following elements of the basic study:

- interviewers cadre, the investments in their training, supervision and quality control.

- questionnaire and interview design,

- information generated in the longitudinal study about local institutional, political, and managerial constraints and stakeholders,

- knowledge emanating from the study about structure and quality of administrative records, e.g. police records education records.

Two kinds of local statistical data generated in surveys are often crucial to a well executed experiment: estimates of the number of individuals relevant to a particular experimental project and estimates of the temporal flow of such individuals through various systems.

So for instance, a longitudinal study that included attention to youthful co-offenders might generate good information on their number, their geographic stability and their general geographic location or locatability. Such pipeline studies based on longitudinal data could arguably help to avoid the problem of some experimental tests in police handling of domestic violence and others (Project Review Team, 1988). Such information is basic to a pipeline study that would inform the design of a experiment dedicated to preventing illegal activity by co-offenders.

It would of course, be a mistake to depend on a longitudinal study to inform all aspect of the design of experiments. It usually cannot help much, if at all, in understanding the ethical or legal propriety of experimental tests, for instance. Nor would a longitudinal study help to understand the obstacles to implementing a new regimen.

The implication of all this is that field experiments can and should exploit longitudinal surveys done in the areas in which the experiment will be emplaced, simply to decrease the cost of experiments. The reduction in cost stems from capitalization on human and statistical resources and savings in time.

4.3 Prophylactic Rationale

Cross-sectional and longitudinal surveys, are often pressed to produce evidence that they cannot always support as, for example, in an important class of questions in the social sciences and public policy concerning the impact of social programs. The Continuous Longitudinal Manpower Survey for instance has been justified and supported primarily on grounds that it is important for understanding the changing nature of the pool of human resources available to society. Its second justification was that it could help understand the effect of special programs, in youth employment and job training.

The second justification may be useful for rhetorical purposes, e.g. to gain political and fiscal support for the survey. But it is not always appropriate and will be counterproductive in the long run insofar as the claim is exaggerated. That is, longitudinal surveys alone are usually not sufficient to estimate the effects of programs designed, say, to affect the earnings of individuals, some of whom happen to participate in the survey. Nor are these designs sufficient for making causal statements about the effects of programs in health, criminal justice, and other areas. See the earlier remarks on scientific justifications nd the reference to Fraker and Maynard's (1985; 1987) and LaLonde (1987) comparisons of program effects based on randomized experiments against effects based on data, notably the CLMS and the Current Population Survey (CPS).

In the case of evaluating of Chapter I programs, relying on longitudinal study will merely continue a practice that is known to be invidious. The estimates of effect, if one follows the instruction of law, will be ambiguous at best and misleading at worst. To the extent that randomized experiments are a prophylactic to such results, and have been recognized as such in medicine and education since the early 1970's (Campbell and Boruch, 1975), then such experiments ought to be considered seriously.

The Program in Human Development and criminal Behavior has grappled with this issue (Farrington, Ohlin, Wilson, 1986) and continues to do so.

4.4 Calibration Rationale

An engineering justification for joining experiments to ongoing longitudinal surveys is that one may use the experiments to calibrate estimates of program effects that are derived entirely from the longitudinal survey (Boruch, 1976). That is, the biases in estimates of program intervention that are based on longitudinal data can be assessed, and periodically corrected, through controlled experiments. Longitudinal studies are then likely to be more policy-relevant and less ambiguous with respect to biases in estimating program effects. Experiments are likely to benefit from their greater generalizability, lower costs, and more manageable administration. As a practical matter, systematic calibration is a couple of decades in the future. Also as a practical matter, one can develop a subjective sense that informs theory and decision, based on rude comparisons of results from both kinds of study. In the work on comparing estirates in supported-work manpower training programs, for instance, the biases engendered by relying on longitudinal study differ depending on whether one considers youth or recipients of Aid to Families with Dependent Children. To be specific, the estimates for the impact on youth in 1979 was near zero for the experiments and minus $1200 for the nonrandomized study. Estimates for AFDC women do not differ appreciably.

It is especially appealing to consider calibration in the case of Chapter I programs because the better parts of the Chapter I Reporting and Information System and infrastructure might be exploited (See Reisner et al (1982) for work up to 1981). The comparison of estimates of program effect based on grade equivalents against estimates based on randomized tests may reveal that the former does well under certain conditions, e.g. for second graders. The accumulation of experience about when each type of estimate is in accord can help us to understand when experiments are not needed.

4.5 Methodological Ration^le: Better Methods and Data

The methodological rationale for joining experiments to longitudinal study can be narrowly construed, and often is, to understanding how to reduce measurement error in tests and interviews.

Some of the methodological reasons for joining experiments to longitudinal studies are implicit in the earlier remarks. The economic rationale for instance, carries the implication that experiments can be better designed; a methodological concern. The statistical and calibration justifications also accord with methodological interests.

Understanding how to elicit accurate information from people in the face of poor memory, difficulcy in understanding questions, and reluctance to provide responses seems important. The problem has at times prompted the design of experiments in the general context of longitudinal studies.

Malvin and Moskowitz (1983) undertook randomized experiments to understand how to better elicit information from junior high school students on their drug use and attitudes. The work involved comparing completely anonymous responses to ones in which identification was elicited but privacy assured by the substitute teachers responsible for administration of questionnaires. The biases reported in identified questionnaires appear to the authors to be very small except for current use of drugs.

The Weis (1987) review of research on reliability of reports on delinquent and criminal behavior suggests that new methods of eliciting information do often not work better than high quality conventional ones. The paper is persuasive on this account. Still, need to improve quality invites attention to better controlled tests. Some of the tests can be adjoined to longitudinal study.

Mathiowetz (1987) for instance has mounted studies to understand how to better ask questions about the unemployment spells of employees of a large company partly to improve quality of data in the Panel Study of Income Dynamics (Mathiowetz and Duncan, 1984). Her object was to ask questions in two different ways to determine which yielded more valid results; validity standard, company employee records, was available. Although in this case the same sample was asked both kinds of questions, an experiment could have been designed to achieve related ends.

4.6 Policy and Political Rationale

A longitudinal study's usefulness to policy lies partly in its capacity to show change. A national shift in truancy level may, for instance, direct attention to the problem.

Consider then that the scholarly and policy use of longitudinal data is high soon after a wave of measurement. The use tapers off rapidly until the next wave. Consider further, several waves of measurement may be characterized by little change in the phenomenon of interest.

The implication is that "surprises" in the sense of new understanding will be infrequent and will decay rapidly. If they occur at all, they will be tied to frequency of measurement and frequent change. To the extent that this is true, one might choose to measure frequently. This may make possible results that show, for instance, that only 10% the individuals involved in high crime commission rates in one year are involved in low or zero rate in a subsequent year. This finding has implications for policy: the high rate individuals are not durable in their enterprise and so perhaps one ought to invest in prevention rather than punishment.

Still, such surprises will be infrequent. And the longitudinal study may have to be refreshed, in the interest of generating understanding that is not obvious.

To refresh and invigorate the study, it seems intellectually justified and politic to join policy experiments to the enterprise. That is, on guarantees surprises - new understanding of a policy relevant kind - by doing controlled experiments that are designed to inform policy. The regimens tested are of course unknown with respect to their effectiveness. On this account they also guarantee new understanding.

Consider for example the proposed Chapter I program evaluations. The expectation of some observers, to judge by P.L. 100-297, is that such programs will indeed affect truancy. A national longitudinal study may detect no effect of program on truancy simply because a national study cannot measure as specifically, frequently, and reliably as is desirable; nor is it reasonable to expect that despite the enormous variation in such programs all will be directed toward truancy. Controlled tests of programs that replicate what appear to be the best of the *existing* programs might then be undertaken in sites that do not have such programs.

In the case of the Program in Human Development and Criminal Behavior, one might also refresh the longitudinal study periodically by undertaking experiments. For instance, handling of students at risk of further truancy varies a great deal. Ethnographic studies of the sort implied by Cooley (1988) may help to identify how most schools handle the matter and how the most conscientious do so. Designing formal programs based on what *appears* to be the best and testing these in a variety of settings is likely to be at least as important, more important perhaps, and as newsworthy as a longitudinal finding that "truancy is associated with delinquency and subsequent crime."

5. Related Research Policies and Origins

5.1 Related Polices and Practices

Precedents exist for coupling prospective methodological experiments to ongoing surveys. The Bureau of the Census, the Social Security Administration's Office of Research and Statistics, and other agencies have undertaken experiments to assess the validity of information reported to them. Measurement error and validity studies have for example been adjoined to the National Longitudinal Study of the Class of 1972. In the social scientific community, the General Social Survey, which regularly employs split half designs to study such phenomenon as the effects of question ordering.

The proposal adjoin experiments to longitudinal surveys is related, of course, to piggybacking in observational surveys, i.e., adding questions to a questionnaire to meet the special needs of sponsors or the public. It is related also to the common practice of augmenting samples to investigate special groups that cannot be explored in a conventional national probability sample. The sample augmentation procedure of the National Assessment of Educational Progress, for example, permits states to add respondents within their states so that confident statements can be made about the state's students' achievement test scores, statements that would have not been possible with the survey's national sample design (Messick, 1984).

Joining experiments to ongoing longitudinal surveys can also be regarded as a special case of matching and linkage of records. The topic of linked files has been of interest to researchers in Sweden, Norway, the United Kingdom, the United States, and other countries for at least 10 years (Schueren, 1985).

The satellite policy proposed here differs from earlier policies and precedents in that it suggests that the studies adjoined to the survey be *prospective randomized tests* of programs, substantive program variations, or their components. Such studies are not designed primarily to inform the methodologist; that aim is important but secondary here. Rather, they are designed to help understand what works better. The distinction is an important one insofar as social experiments engender problems that are not encountered (or are encountered in less extreme forms) in methodological experiments.

5.2 Recent Origins

The proposal for joining experiments to ongoing longitudinal surveys has origins in the debate among scholars and bureaucrat-scholars about how much one can depend on longitudinal data. It shares an interest with those who have discussed the more issue of combining experimental and sampling structures (Fienberg and Tanur, 1986; 1987b). There is no doubt about the need for such data for understanding change. The debate lies in whether these data can be used sensibly to understand the causes of change.

For example, Richard Berk and others at a recent MacArthur Foundation conference vigorously discussed whether longitudinal surveys of criminal careers can effectively be exploited to understand the impact of programs designed to affect these careers. The discussion led to, among other things, a MacArthur-funded policy paper on social experimentation (Berk et al., 1985) that stressed the importance of controlled randomized experiments relative to other approaches to estimating program effects.

The National Research Council's Panel on Criminal Careers makes longitudinal study paramount in its proposed research agenda (Blumstein, Cohen, Roth, Visher, 1986). Randomized field experiments are considered generally in the context of longitudinal study as a device to test hypotheses emerging from such study and to test projects in prevention, criminal career modification and selective incapacitation. Specific linkages between each approach to understanding are implied but not discussed in detail.

Similarly, the National Academy of Sciences' Committee on Youth Employment Programs examined major studies to understand whether one could draw firm conclusions about program effects from earlier research (Betsey et al, 1985). The committee concluded, among other things, that longitudinal surveys are no substitute for randomized experiments when the object is to estimate the effectiveness of new youth employment programs. Moreover, the committee urged the use of randomized experiments for this purpose; a satellite policy is discussed in an appendix to its report.

The proposed guideline for coupling randomized design to longitudinal surveys can also be traced to a technical advisory committee for employment program evaluation appointed by the Department of Labor. The DOL scught to learn whether analyses of manpower programs based on conventional longitudinal surveys against estimates based on randomized trials. The conclusion of this exercise was that the two estimates are not always in accord. Indeed, they differ remarkably.

5.3 Earlier Origins

The justification for the coupling of longitudinal, cross-sectional and other surveys with randomized experiments appeared in the early 1970s. In particular, the Social Science Research Council's Committee on Experimentation as a Method for Planning and Evaluating Social Interventions devoted considerable attention to the problem of generalizing from experiments.

The Committee produced two state-of-the-art monographs: Riecken et al. (1974), Boruch and Riecken (1975), and a variety of papers. One of these papers concerned the coupling of randomized experiments to "approximations to experiments" such as longitudinal surveys and the models used to underpin their analyses (Boruch, 1975).

Proposals for adjoining experiments to longitudinal and some cross-sectional studies have since this early work been presented formally to policy boards responsible for enhancing data bases and their utility. The groups include the Policy Advisory Board of the National Center for Educational Statistics (1982), the Policy Advisor Board of the National Assessment of Educational Progress (Boruch and Sebring, 1983), the National Science

Foundation's Human Resources Division (1982), and others.

6. Examples of the Contexts to Which the Satellite Policy is Relevant

To illustrate the kinds of setting to which the proposal is pertinent consider some examples. In what follows, different longitudinal studies and different experiments are considered. The settings bear on out-of-school youth and young adults, high school students, and children in early grades who are at risk.

6.1 Chapter I Evaluation

Consider Broward County's AIM project as a possible model. The project was directed at second graders at risk of academic failure. Risk was determined by the students' performance below the 26th percentile on the Iowa Test of Basic Skills. The A.I.M. program involved random selection and assignment of these students to all day programs in small classrooms, emphasis on basic skills, classes being taught by specially selected teachers.

The project was undertaken in a District that has considerable standardized testing and a Research Department that is active. The experimental field test of the A.I.M. project exploited the testing and research infrastructure in several ways that can be emulated in evaluating Chapter I programs.

- Candidates for the program were identified on the basis of regular testing, i.e. low ITBS scores.
- Impact of the program was based on the ITBS administered to project participants and comparison students.
- Routinely collected administrative records on absences and behavior problems were used to understand implementation and outcome,
- specialty tests were developed to capture localized differences between the randomized A.I.M. and non-A.I.M. students.

- the administrative system for tracking students was used too.

Not all school districts are interested in improving programs in ways that are testable of course. Not all school have sufficient numbers of students at risk to justify the investment in either program innovation or formal test. Broward County School District is, for instance, the largest in the country.

The implication is that not all districts with Chapter I programs are capable, much less willing, to emulate such tests. Nonetheless, the Broward experience can help to inform the work of others, and to inform the way we think about coupling experiments to surveys and to routine administrative and academic information systems.

6.2 Multicohort-Multicity Longitudinal Studies of Delinquent Behavior

Consider surveys currently being designed by the Program on Human Development and Criminal Behavior. These surveys are relevant to proposals for Chapter I evaluation in the sense that both studies are longitudinal in character, are likely to focus on at least some common outcome variables such as truancy, and will be national in scope.

It is not hard to identify potentially interesting experiments that might effectively exploit a longitudinal study infrastructure and be worth doing. In fact, the number of options is sufficiently great to make choice difficult. The feasibility of any option may then be the determining factor, e.g. willingness of the site's public service agencies, such a police department or court, or community based organizations to cooperate.

For example, relatively innocuous and small but useful side experiments might be adjoined in all longitudinal studies to determine which methods are most effective locally in eliciting cooperation in the main longitudinal study or in improving the accuracy of reporting on delinquent or criminal activity. A strategy that comports with this aim might simply replicate and improve earlier experimental tests of such methods, such as:

- Malvin and Moskowitz (1983) on drug attitudes and use among junior high school students.
- Goodstadt and Grusen and others on the use of randomized response and other methods for eliciting sensitive information (Boruch and Cecil, 1979).
- Bradburn and Sudman (1981) and others on alternative methods of interviewing and questionnaire design to improve data quality.

Potentially useful experimental tests are implicit in Weis (1987).

For adolescent or in-school cohorts, it may be desirable and feasible to design and test programs based on a variety of theoretical perspectives. Differential association theory (Ohlin, 1988), for instance, suggests that association of target adolescents with others who are more or less delinquent will affect the targets' delinquent behavior. To the extent that school based programs (e.g. that focus on unacceptable social behavior) or programs that attract individuals who are out of school into employment or other programs are worth testing, the longitudinal infrastructure will facilitate such testing. The extent to which shifts in association can be controlled at all seems worth testing in a controlled education, sociological and training contexts.

Taking this idea further, Reiss (1987) reviewed available research on co-offenders generally. He endorses the idea put forward by Klein and Crawford that external sources of cohesiveness of gangs, if eliminated, would lead to gang dissolution or degraded cohesion. He recognizes that conventional approaches, e.g. incapacitation and social work attention, do not reduce internal cohesion and, on the contrary, may increase it. The options that are explicit in the Reiss paper and that lend themselves to experimentation include:

- court oriented efforts to sanction co-offenders in ways that are different from sanctioning individuals (to increase sense of risk), e.g. early sanctions to all co-offenders.

- interventions designed to reduce external sources of cohesiveness (e.g. threats from gangs, revenues from drug sales)

- intervention designed to disrupt recruitment of co-offenders.

Consider now a different kind of coupling, one that involves a randomized test, a time series analysis, and longitudinal study. The idea of combining these has precedent in at least one major economic effort: the Experimental Housing Allowance Program. In EHAP, poor families within certain cities were randomly assigned to various kinds and levels of housing allowance (e.g. for home repairs). In other cities, involved so called saturation experiments, the providers of housing were given federally subsidized support to understand how to enlarge the supply of quality housing for the poor; the effect was in these projects based on times series analyses.

Current related kinds of couplings are underway in Wisconsin. Irv Garfinkle and his colleagues have begun randomized experiments on better ways to extract child support from delinquent fathers. And to understand how community wide interventions affect such payment, saturation tests have been designed for county level implementation. It is conceivable that similar randomized tests and nonrandomized time series or panel analyses can be executed in other areas, in the interest of understanding how to assure that young, out of home fathers provide financial support to their children.

Alex Weiss (1988) has considered the merits and shortcomings of randomized experiments on police handling of crime His stress on the use of time series approaches suggests a coupling of the approaches. So, for instance, if the general effects of delinquency deterrence are plausible at all they ought to emerge from community wide programs that focus on norms, associations, handlers, sanctions, and so on. And in some geographic areas, pertinent saturations experiments that exploit time series or longitudinal data may be feasible. Elsewhere, deterrent effort that focus on offenders and co-offenders might be designed and tested in randomized experiments that also include long term (longitudinal) follow-up.

6.3 Education: High School and Beyond

Consider the National Longitudinal Study of the High School Class of 1972 (NLS) and High School and Beyond (HSB), a national longitudinal study of the high school class of 1980. These surveys are costly and widely used by the educational research and policy community. They are sponsored by the National Center for Education Statistics (NCES) and have led to a variety of provocative reports, e.g. Coleman et al. (1982).

There are a variety of reasons why HSB is relevant to proposals for a Program on Human Development and Criminal Behavior is relevant. To the extent that the Program or Chapter I evaluation will involve study of the onset and desistance of delinquency among in-school children, the HSB sample might be augmented to focus on the high risk geographic areas and people that are of primary interest. Questions might be added to ordinary HSB questionnaires to add to the fund of knowledge.

More to the point, consider that the Program in Human Development and Criminal Behavior may be in a position to augment not is own longitudinal survey, but to augment HSB or a Chapter I evaluation that is coupled to HSB. That is, if the program invents, extends, or facilitates the invention of programs that reduce delinquency among high school students, then the Program's interest in testing them could drive the tests beyond its own borders. The drive may stem from inadequacy or irrelevance of its own target samples, or from simple interest in better use of institutional resources.

For instance, differential association theory explored by Ohlin suggests that an individual's desistance from crime results in part from a change in associations, notably a change from criminal associations to noncriminal. Inducing and maintaining such a change may involve jobs, military service, or other special handling methods. Programs designed to do the job should take account of history in locations, number of those at risk, level of risk and so on. Information about these are available or can be collected at marginal cost from target areas in HSB. Further, the relations between HSB and local sites are sufficiently good to consider providing opportunities to do side experiments on effectiveness of such programs.

The example implies a link between delinquency research and educational research. Why would a federal office of educational research and statistics benefit from an explicit satellite policy more generally? There are several reasons. First, issues of data and resource sharing have emerged often during meetings of advisory committees for the HSB and the NLS, and it seems reasonable to expect their reoccurrence. It then seems sensible to develop a program of joining experimental studies to these surveys that would help such committees and their staff understand how to respond to these issues equitably and efficiently.

6.4 Employment and Training

Let us suppose that randomized trials of employment and training programs are not always appropriate or feasible. Suppose further that there is some interest in learning from such trials, especially through using longitudinal surveys as a vehicle for their implementation. How might such experiments be carried out?

Several strategies may be appropriate, and are reflected, for example, in current plans to evaluate programs of the Job Training and Partnership Act (Bloom et al., 1987). All of the following discussion assumes that experiments can be conducted in a way that permits one to take advantage of the longitudinal data and the organization structure used for its collection without disrupting that process.

Specific components of full programs may warrant testing. For example, we know very little about when, why, and how different varieties of job counseling "work." Mounting experiments in a selection of sites to assess the effects of the components of an employment and training program will often be more feasible and perhaps more appropriate than national trials on full-blown programs. See, for example, Bickman (1985) on assessing preschool programs for children in Tennessee.

Augmenting the existing employment and training regimens may be feasible in some sites. For example, how "residential' does residential training have to be? We know that some residential programs work (e.g., the Job Crops). We do not know how brief the residential experience can be while continuing to be effective (see, for example, Betsey et al., 1985, on such programs).

There is little good evidence to help answer the question "Does it 'pay' to treat the most needy, rather than the least needy?" The most "tractable" people (i.e., those most likely to benefit from training) often lie at the margin of need. And this margin often defines a population for which randomized trials are likely to be most feasible. randomization at the margin can be coupled with other designs as well, e.g., regression-discontinuity (Riecken et al., 1974).

Selecting only the best of an array of research sites that are capable and willing to conduct experiments will not give fair estimates of the impact of programs. But such sites will demonstrate the best that can be done, thus providing evidence that may be sufficient for purposes of making policy and producing research that is heuristically rich for the social sciences.

7. Probable Issues and Options

The idea of adjoining field experiments periodically to longitudinal surveys is not new. But it has not emerged often and this accounts perhaps for the scarceness of thoughtful papers on the topic. Another reason for the scarceness of papers may be the difficulties of executing the idea.

Some of the difficulties are resolvable given the current ability of research-managers and manager-researchers. Others require more thinking and perhaps pilot tests.

The following considers issues and options that are general, in the sense of not depending on whether the experiments are adjoined to an existing longitudinal study or to a proposed study. Respondent burden is important regardless of design for example. It also treats issues that depend on whether the experiment is adjoined to an existing study, e.g. proprietary interests, or to a proposed one.

7.1 Standards for Joining Field Experiments to Ongoing Surveys

The proposal put forward earlier suggested that adjoining experiments to a longitudinal study be regarded as a legitimate research policy options so long as:

(1) the experiment is compatible with the mission of the longitudinal survey;

(2) the risks of disruption to the survey can be managed;

(3) designated contractors are responsible for oversight of of the process; and

(4) the experiment engenders no appreciable cost to the agency supporting the longitudinal research.

Adhering to these standards is likely to reduce or eliminate obvious problems.

Still, one must decide which of a variety of potential experiments should and can be adjoined to the longitudinal study. Greenwood's (1987) draft paper lays out five criteria that help in making a choice. Paraphrased, the criteria include:

(1) theoretical importance of the program(s) proposed for experimentation

(2) empirical evidence for the worth of the program(s),

(3) "amount of difference" between proposed regimens and current practice,

(4) compatibility with the longitudinal design, and

(5) political feasibility.

The fourth item of course is part of the Boruch-Pearson (1988) proposals. Discussions and criteria for understanding political and managerial feasibility are important and have been given in, among others, Chelimsky's (1985, edited volume on evaluation at local, regional, and federal levels of government, and in Riecken et al (1974) on managerial, ethical, institutional and political issues, engendered by social experiments.

Greenwood's second criterion implies that evidence ought to be available from quasi-experimental or other randomized experiments. It seems sensible, given the likely cost of mounting new experiments, the need to anticipate outcomes, and the need in most field experiments to rely on earlier pilot testing of randomization procedures, measures, and negotiation strategies (Boruch and Wothke, 1985).

Criterion number three is interesting in part because one can easily argue two sides. To the extent a difference between proposed regimens and existing control regimen is small, then detecting a difference in outcome will probably be difficult and perhaps not worth the effort. On the other hand, a small change is likely to be politically and managerially more feasible than a large one.

Similarly, to the extent that the difference between proposed regimen an existing control regimen is large, differences in outcome are likely to be more detectable and the product may be useful on policy and theory ground. But the managerial problems may be difficult. Riecken et al's (1978) handling of this matter is to encourage some testing of extreme program levels, the reasoning being that most interventions are weaker than they are predicted to be and that effects are, if the variation is effective, more detectable (p. 33-34).

7.2 Adjoining Experiments to Existing Surveys

Proprietary interests of researchers are important of course. The principal investigators in a longitudinal study such as a Chapter I evaluation may be disinclined to permit another research group, such as the Program on Human Development and Criminal Behavior, to augment Chapter I samples or questionnaires because this would capitalize on the Chapter I infrastructure, expertise or ideas. It would yield no obvious benefit to the Chapter I researchers. Similarly, the major sponsor for a Chapter I evaluation , the U.S. Department of Education, may see no benefit in sharing credit for an important survey by cooperating with another federal agency, e.g. the National Institute of Justice.

Some ways, quid pro quos, to meet proprietary interests then must be developed to make satellite policy possible. As important, proprietary interests are the managerial problems that the policy can engender. The National Opinion Research Center, for instance, operates HSB and is under no obligation to cooperate with organizations responsible for surveys or experiments in another area. Moreover, developing such an obligation through contract and negotiated agreements may be difficult. There are few precedents for interorganizational cooperative research in policy and social science research. There are none for the satellite research of the kind proposed here.

7.3 Adjoining Experiments Regardless of Longitudinal Study Type

Respondent burden is and will continue to be important. For example, if an experimentation effects of Chapter I program variations asks a substantial fraction of children in early grades in a set of school districts to respond to a questionnaire and a separate study of delinquent behavior directs other questions to the same individuals, the burden on the respondents and their guardians (who must provide consent) may be increased and be notable.

Monetary payments may offset the burden. Indeed, the experience in at least some studies of adolescents suggests that payment leads to not only good cooperation of the target sample members but to requests to cooperate from those outside the sample (Howard and others, 1988).

Monetary respondents are irrelevant if there is competition for respondents in any real sense. That is, if local rule or custom dictates that the respondent can participate in only one study, then payment by a second aspiring researcher will not be relevant.

Further, monetary payments to respondents ought not be relevant if the experiment adjoined to the ongoing survey can disrupt the survey. In this case, augmenting the basic sample targeted for survey may be the only way to obtain additional information for the experiment.

Similarly, and more important, an experiment adjoined to a survey will disrupt the results of a survey in a special sense. For example, the survey researcher that members of the sample encounter "ordinary" conditions. The experiment will perforce introduce an extraordinary condition, albeit for a small fraction of the sample. The experimental regimen will, if effective, then affect the estimates of prevalence for incidence that are important to the longitudinal study. Again, the only resolution to this problem appears to be augmenting the sample targeted in the longitudinal study.

Augmentation of a targeted sample to reduce individual respondents' burden then may help to resolve one problem but it generates another. If a central federal, state or local agency dictates the permissible tocal number of respondents, then the tactic does not help. Paying additional respondents may do so, as might other tactics.

7.4 Feasibility and Appropriateness of Experiments

Conducting controlled experiments to plan and evaluate new programs, program variation, or components is no easy matter. This is regardless of whether the experiment is coupled to a longitudinal study.

The standards for judging their appropriateness and feasibility have been laid out elsewhere, e.g. Boruch (1985). put briefly, appropriateness hinges on answers to questions such as:

- Does current practice need improvement?
- Is there important uncertainty about the proposed innovation?
- Will methods <u>other</u> than randomized experiments yield good estimates of relative effectiveness?
- Will results of the experiment be used?

These are closely linked to standards for ethical propriety of experiments.

The standards for feasibility hinge on answers to the following questions:

- Have standards for appropriateness and propriety been met?
- Are technical and financial and human resources sufficient?
- Is the process of the new program or variation understood, described, capable of replication?
- is the target group and context well understood?

Methods for addressing these questions and enhancing feasibility are discussed in Bloom et al 91987), Betsey, et al (198:), Boruch and Wothke (1985), Riecken et al 91974), Boruch and Riecken (175), among others.

The Human resources are perhaps most important in assuring quality and feasibility of controlled experiments. For Chapter I evaluations, it seems clear from precedent that some school districts have relevant capacity, e.g. Broward County, Florida and Austin, Texas, Some, not all, of the Chapter Technical Assis ..nce Centers are likely to have the expertise necessary to provide counsel to school districts on the use of randomized tests for program improvements (Reisner, Turnbull, and David, 1988). Indeed, directors of TACs, such as Echternacht, constitute a resource that can be capitalized nicely in this arena.

8. Summary

Longitudinal surveys based on well designed probability samples are the best possible approach available to describing growth of individuals and change at the national level. Such surveys often do not yield defensible estimates of the effect of intervention, e.g. of Chapter I programs.

Controlled randomized experiments are the best possible approach to estimating relative effects of interventions, program variations, etc. They are often not feasible at the national level however.

Coupling controlled randomized tests to longitudinal study can provide both understandings of growth or change and unbiased estimates of what works better in more local contexts.

A formal policy for coupling experiments to longitudinal study then seems sensible. Such a policy is analogous to research policy in satellite use. The major vehicle for generating information, the satellite, is periodically reorientel and partly dedicated to special experiments and is analogous to the longitudinal study system.

The main justification for the proposed satellite policy for Chapter I is scientific and policy relevant: better data to inform policy about how to improve programs. The secondary reasons include: economic ones, e.g. local experiments capitalize well on longitudinal infrastructure; methodological reasons, e.g. learning about how to improve data quality generally; political reasons, notably permitting answers to several questions.

Selection of interventions for experimentation should be guided by several criteria: theoretical import of the intervention, empirical support for its promise, propriety of a test, feasibility of implementing both the interventions and the randomized experiment.

In Chapter I, replication of exemplary projects may meet all these criteria. The experiments may for example test new ways of sustaining parental involvement, reducing drop-out, decreasing low grades and failures, tutoring, and so on.

Executing controlled experiments in Chapter I projects requires resources: well trained researchers and practioners and support for both. Failure of some projects is likely simply because learning how to improve and generating evidence on it is difficult. Assuming a failure rate of 20% for executing the experiment (regardless of program success) is reasonable.

Statistical characterization of the target groups (who is eligible, who gets service) etc. is essential for design of the experiments. So is careful literal and statistical description of the processes engendered by the program, e.g. time in Chapter I variation, nature of variation. Both can be generated at least crudely by longitudinal study.

Theory will be important in the longitudinal study to estimate, at the macro-level, effects. The experimental programs will, if based on similar theory, help to adjust statistical vulnerability of the longitudinal work.

A major legislative implication of this perspective is that mandates for longitudinal study must also authorize demonstrations, i.e. implementations of new programs, variations, and components.

8. References

Bailar, B.A. and C.M. Lanphier (1978) Development of Survey Methods to Assist Survey Practices. Washington, D.C.: American Statistical Association.

Barnes, R.E. and A.L. Ginsberg (1979) "Relevance of the RMC Models for Title I. Policy Concerns." Educational Evaluation and Policy, 1(2), p. 7-14.

Berk, R.A. et al. (1985) "Social Policy Experimentation." Evaluation Review 9:387-429.

Betsey, C., R. Hollister, and M. Papgiorgiou (1985) The YEDPA Years:Report of the Committee on Youth Employment Programs. Washington, D.C.: National Research Council.

Bickman, L. (1985) Improving Established Statewide Programs. Evaluation Review, 9: 189-208.

Bloom, H.S., M.E. Borus, and L.L. Orr (1987) "Using Random Assignment to Evaluate an Ongoing Program: The National JTPA Evaluation". Presented at the Annual Meeting of the Statistical Association, San Francisco, August 17-20.

Blumstein, A., J. Cohen, J. Roth and C.A. Visher, eds. (1986). Criminal Careers and "Career Criminals". Washington, D.C.: National Academy Press.

Boruch, R.F. (1975) "Coupling randomized experiments and approximations to Experiments in Social Program Evaluation." Social Methods and Research 4: 31-53.

Boruch, R.F. and Otehrs 91985). "Randomized Experiments for Evaluating and Planning Local Programs: A Summary on Appropriateness and Feasibility." In Chelimsky E. (Ed.) Program Evaluation: Patterns and Directions. Washington, D.C.: American Society for public Adminsitration, PAR Classics Series, pp. 165-175.

Boruch, R.F. and H.W. Riecken (eds) (1975) Experimental Testing of Public Policy, Boulder, CO: Westview.

Boruch, R.F. and J.S. Cecil (1979) Assuring confidentiality of Social Research Data. Philadelphia: University of Pennsylvania Press.

Boruch, R.F., A.J. McSweeny, and J. Soderstrom (1978) "Bibliography: Illustrative Randomized Experiments". Evaluation Quarterly, 655-695.

Boruch, R.F, and W. Wothke (1985) "Seven Kinds of Randomization plans for Designing Field Experiments". New Directions for Program Evaluation (Jossey-Bass), 28, 95-118.

Boruch, R.F. and Pearson, R.W. (1988) Assessing the Quality of Longitudinal Surveys. Evaluation Review, 12: 3-58.

Bradburn, N. and S. Sudman (1981) Improving Interview Method and Questionnaire Design. San Francisco: Jossey-Bass.

Broward County School Board. Department of Research (1987). Achievement Through Instruction and Motivatin: Program Evaluation Report for 1986-87. Fort Lauderdale. Florida: School Board of Broward County, Research Department.

Campbell, D.T. and R.F. Boruch (1975) "Making the Case for Randomized Assignment to Treatments by Considering the Alternatives," pp. 195-297 in C.A. Bennett and A.A. Lumsdaine (eds.) Central Issues in Social Program Evaluation. New York: Academic Press.

Chelimsky, E. (1985) Program Evaluation: Patterns and Directions. '/ashington, D.C.: American Society for Public Administration (PAR Classics Series).

Coleman, J.S. et al (1982) High School Achievement: Public. Catholic and Private Schools Compared. New York: Basic Books.

Cooley, W.W. (1988) "Design for a Longitudinal Study of Chapter I" Briefing to the U.S. Department of Education, Washington, D.C.

Cottingham, P. and A. Rodriguez (1987) "T e Experimental Testing of the Minority Female Single Parents Program." Presented at the Annual Meeting of the American Statistical Association, San Francisco, August 17-20.

Cronbach, L.J. et al. (1980) Toward Reform of Program Evaluation. San Francisco: Jossey-Bass.

Duncan, G.J., d G. Kalton (1985) "Issues of Design and Analysis of Surveys Across Time." Presented at the centenary session of the International Statistical Institute, Amsterdam.

Duncan G.J., F.T. Juster, and J.N. M an (1984) "The Role of Panel Studies in a World of Scarce Research Resources," in S. Sudman and M.A. Spaeth (eds) The Collection and Analysis of Economic and Consumer Behavior Data: In Memory of Robert Ferber. Champaign, IL: Bureau of Economics and Business Research.

Farrington, D P. (1988) "Advancing Knowledge About Delinquency and Crime: The Need for a Coordinated Program of Longitudinal Research." Behavioral Sciences and Law, 6(3), pp. 307-331.

Farrington, D.P., L.E. Ohlin and J.Q. Wilson (1986) Understanding and Controlling Crime: Toward a New Research Strategy. New York: Springer-Verlag.

Fienberg, S.B. and J. Tanur (1986) "From the inside out and the outside in: Combining Experimental and Sampling Structures." Technical Report 373, Carnegie-Mellon University (December).

Fienberg, S.B. and J. Tanur (1987a) "The Design and Analysis of Longitudinal Surveys: Controversies and Issues of Cost and Continuity:, in R.F. Boruch and R.W. pearson (eds) Designing Research with Scarce Resources. New York: Springer-Verlag.

Fienberg, S.B. and J. Tanur (1987b) Experimental and Sampling Structures: Parallels Diverging and Meeting. International Statistics Review, 55:75-96.

Fienberg, S.E., B. Singer, and J. Tanur (1985) "Large scale social Experimentation in the United States," pp. 287-326 in A.C. Atkinson and S.E. Fienberg (eds) A Celebration of Statistics: The ISI Centenary Volume. New York: Springer-Verlag.

Fienberg, S.E., M.E. Martin, and M.L. Straf (1985) Sharing Research Data. Washington, D.C.: National Academy of Sciences.

Fraker, T. and R. Maynard (1985) The use of Comparison Group designs in Evaluation of Employment Related Programs. Princeton, NJ: Mathematica Policy Research.

Fraker, T. and R. Maynard (1987) "The Study of Comparison Group Designs for Evaluations of Employment-Related Programs." The Journal of Human Resources. 22: 194-227.

Frederikson, C.H. and J.A. Rotondo (1979) Time Series Models and the Study of Longitudinal Change. pp. 111-154 in J.R. Nesselroade and P.B. Baltes (eds) Longitudinal Research in the Study of Behavior and Development. New York: Academic Press.

Gray-Donald, K. and M.S. Kramer (1988) "Causality Inference in Observations Versus Experimental Studies". American Journal of Epidemiology, 127. pp. 885-892.

Greenwood, P. (1988) The Role of planned Interventions in Studying the Assistance of Criminal Behavior in a Longitudinal Study: Concept Paper Developed for the Desistance Group of the Program on Human Development. Santa Monica: RAND Corporation. (July).

Gueron, J.M. (1985) "The Demonstration of State Work/ Welfare Initiatives." New Directions for Program Evaluation, 28, 5-13.

Heckman, J. and B. Singer, (eds) (1985) Longitudinal Analysis of Labor Market Data. Chicago: University of Chicago Press.

Heckman, J.J. and R. Robb, Jr. (1985) "Alternative Methods for Evaluating the Impact of Interventions," pp. 156-246 in J.J. Heckman and B. Singer (eds.) Longitudinal Analysis of Labor Market Data. New York: Cambridge University Press.

Howard, K. et al. (1988) A Survey of Adolescents and Their Access to Mental Health Services. Psychology Department, Northwestern University, Evanston, IL.

LaLonde, R. (1986) "Evaluating the Econometric Evaluations of Training Programs with Experiments". American Economic Review, 76(4): p. 604-620.

Linn, R.L. (1979) "Validity of Inferences Based on the Porposed Title I Evaluation Models". Educational Evaluation and Policy Analysis: 1(2). pp. 23-32.

Malvin, J.H. and J.M. Moskowitz (1983) "Anonymous Versus Identifiable Self Reports of Adolescent Drug Attituds, Intention and Use." Public Opinion Quarterly, 47, pp. 557-566.

Mathiowetz, N.A. and G.J. Duncan (1984) "Temporal Patterns of Response Errors on Retrospective Reports of Unemployment and Occupation." Proceedings of the American Statistical Association: Section on Survey Research Methods, 652-657. Washington, D.C.: American Statistical Association.

Mathiowetz, N. (1987) Response Error: Correlation Between Estimation and Episodic Recall Tasks. Proceeding of the American Statistical Associaton: Survey Research Methods Section. Washington, D.C.: American Statistical Society, pp. 430-435.

Maynard, R.A. (1987) "The role of Randomized Experiments in Employment Training Evaluations". proceedings of the Section on Survey Research Method: American Statistical Association, Washington, D.C.: American Statistical Association, pp. 109-113.

Mazur, A. and E. Boyko 91981) "Large-Scale Ocean Research Projects: What Makes them Succeed or Fail?" Social Studies of Science, 11: 425-449.

Messick, S. (1984) "A New Design for the National Assessment of Education Progress." Proceedings of the Section on Survey Research Methods, American Statistical Association. Washington, D.C.: American Statistical Association, 83-87.

Mundel, D. (1979) Memo to Franklin Zweig (November 15) Congressional Budget Office.

Ohlin, L.E. (1988) Memo to Working Group on Desistance Regarding Policy Statement as Program Objective. Program on Human Development and Criminal Behavior. Castine, Maine (May 12).

Pearson, R.W. (1987) Researchers' Access to U.S. Federal Statistics. Items 41: 6-11.

Project Review Team. (1988) Report on the Spouse Assault Replication Project to the National Institute of Justice. Department of Statistics and Psychology, Northwestern University, Evanston, IL.

Reisner, E.R., M.C. Alkin, R.F. Boruch, R.L. Linn, and J. Millman. Assessment of the Title I Evaluation and Reporting System.. Washington, DC: U.S. Department of Education, 1982.

Reisner, E.R., B.J. Turnbull, and J.L. David (1988) Evaluation of the ECIA Chapter I Technical Assistance Centers. Washington, D.C.: Policy Studies Associates, Inc.

Reiss, A.J. (1986) Co-offending Influences on Criminal Careers. In A. Blumstein, J. Cohen, R. Roth, and C. Visher (Eds.) Criminal Careers and 'Criminal Careers'. Volume I. Washington, D.C.: National Academy of Sciences, pp. 121-160.

Reiss, A. and others(1988) Pipeline Studies in the Spouse Assault Replication Project. In: Report of the Program Review Team, Spouse Assault Replication Project, to the National Institute of Justice. Departments of Statistics and Psychology, Nortwestern University, Evanston, IL.

Riecken, H.W. et al. (1974) Social Experimentation. New York: Academic Press.

Rubin, D.B. (1974) "Estimating Causal Effect of Treatment in Ranomized and Nonrandomized Studies". Journal of Educational Psychology. 66, pp. 688-701.

Rubin, D.B. (1987) Multiple Imputation for Nonresponse in Surveys. New York: John Wiley.

Schueren, F. (1985) "Methodologic Issues in Linkage of Multiple Data Bases." Prepared for the Panel on Statistics for the Aged Population, national Academy of Sciences, Washington, D.C.: National Academy of Sciences.

Smith, M. (1988) "Thoughts on the Chapter I Longitudinal Evaluation Design" Briefing to the U.S. Department of Education, Washington, D.C.

Stafford, F. (1985) "Forestalling the Demise of Empirical Economics; The Role of Microdata in Labor Economics Research," in O. Ahsenfelter and R. Layard (eds.) Handbook of Labor Economics. New York: North-Holland.

Stone, E.F., D.G. Gardner, H.G. Gueutal, and S. McClure (1983) A Field Experiment Comparing Information Privacy Values, Beliefs, and Attitudes Across Several types of Organizations. Journal of Applied Psychology,68: pp. 459-468.

Taeuber, R. and R.C. Rockwell (1982) "National Social Data Series: A Compendium of Brief Descriptions." Review of Public Data Use 10:23-111.

Verdonik, F. and L.R. Sherrod (1984) An Inventory of Longitudinal Research on Childhood and Adolescence. New York: Social Science Research Council.

Weis, J.G. (1987) Issues in the Measurement of Criminal Careers. In A. Blumstein, J. Cohen, J.A. Roth, and C.A. Visher (eds) Criminal Careers and "Career Criminals". Washington, D.C.: National Academy Press, 1986.

Weiss, A. (1988) Randomized Experiments and Time Series Analysis in Police Research. Department of Political Science, Northwestern University, Weiss, A. (1988) Randomized Experiments and Time Series Analysis in Police Research. Department of Political Science, Northwestern University, Evanston, IL.

USE OF COMPARISON GROUPS IN THE EVALUATION OF CHAPTER 1

by

Edward C. Bryant
Westat, Inc.
September 27, 1988

Basic Factors in the Evaluation Design

Evaluation of the impact of Chapter 1 assistance implies quantifying the amount by which students have benefitted by participation in the program. Quantifying the benefit implies the ability to estimate what the participants' measures of attainment and achievement would have been if there had been no Chapter 1 assistance.

If experimental evaluation were possible, one could randomly withhold Chapter 1 services from a sample of students who would otherwise be eligible for participation in Chapter 1. The achievement of this group (the control group) would constitute the baseline against which the achievement of the treated group could be compared. But no such experimental evaluation is possible for Chapter 1 if the interim and final report deadlines are to be met. Note, however, that an experimental evaluation could be used to supplement the overall evaluation. Many details need to be worked out to make this approach feasible.

The term "comparison group" is sometimes applied to a group of persons who are similar in characteristics to the treated group but who, for a variety of reasons, have not received the treatment. "Natural experiment" and "quasi experiment " are terms often used when a program is evaluated using a nonrandom comparison group. Such evaluations are not experiments, but they can be reasonably successful if the characteristics that distinguish between participation and nonparticipation are independent of outcomes in the absence of the treatment. Another way of saying the same thing is that one would expect the treatment group, if they hadn't received the treatment, to have the same outcome as the comparison group. Because Chapter 1 participation depends on the economic disadvatagement of the school population, and that, by assumption, affects outcomes, it is difficult to see how the concept of a natural experiment can apply if, in fact, the assumption that outcome is related to disadvantagement is correct.

In some kinds of evaluations, a "before and after" study can be conducted, in which the effectiveness of the program is judged on the basis of the difference in relative standing of participants before participation and after participation. The approach will not work in the Chapter 1 evaluation, however, because of a phenomenon known as "regression toward the mean." Under Chapter 1, the students chosen for participation in eligible schools are selected primarily on the basis of their low achievement. The measurement of such achievement is subject to substantial error and, if the students were tested at another time, their rankings might be substantially different. Therefore, those chosen to participate, being at the low end of the scale of achievement at time of initial testing, could be expected to move upward, on average, at a later test date, whether they received any services or not. Reducing the effect of this phenomenon is one of the principal reasons for creating comparison groups.

The Basic Evaluation Model

Various alternatives in the evaluation of Chapter 1 include (1) a series of cross-sectional studies, (2) a longitudinal study, (3) a retrospective study, (4) a series of short term longitudinal studies, and (5) combinations of the above. Each of these approaches has advantages and disadvantages. Regardless of the approach taken, the selection of a comparison group is critical to the success of the evaluation.

In a perfectly designed and executed experimental evaluation there is no need for a complex model linking outcomes with participant characteristics. Random variation in participant characteristics not controlled in the design can be relied on to average out, permitting the evaluation to be completed by simply comparing the average outcome of the treatment group with the average outcome of the control group, with suitable estimation of the standard errors. In the Chapter 1 evaluation, however, one knows in advance that the characteristics of the nonparticipants will not match exactly the characteristics of the participants. If such a match were possible it would constitute persuasive evidence that Chapter 1 assistance was not being given to those segments of the population deemed by the law to be in most need of it.

The best one can hope for in a nonexperimental evaluation is that a model can be found that relates characteristics associated with participants and nonparticipants in such a way that a reasonable estimate can be made of the outcome that participants would have achieved if they had not participated in the program. Such characteristics must be considered both in the design of the study and in the analysis of results. The design provides the rules for the selection of the participant and nonparticipant groups. Rules, built into the design, that equate the groups, to the extent feasible, require fewer subsequent (and less valid) statistical adjustments than would be required if the adjustment were left entirely to the analysis.

Figure 1 displays a conceptual model for the evaluation. It is assumed that the outcome of a student is related to a number of characteristics that can be grouped into community, school and family factors, student factors, measurement factors, and whether the student participated in the Chapter 1 program.

For simplicity in the presentation, it is assumed that participation in Chapter 1 is known and is dichotomous. In reality, students participate for varying periods and at various grade levels, and their former participant status, in the case of transfers, may not be known. It is clear that some definitions need to be developed and, possibly, participation needs to be quantified in terms of the duration of participation. For purposes of presenting the model, however, these problems have been submerged.

The small overlap between Chapter 1 and nonChapter 1 students with respect to school factors, as shown in Figure 1, is deliberate. This is the most difficult part of the match between participant and nonparticipant characteristics. There are many economically disadvantaged students who do not participate in Chapter 1 services. While there are some technical problems with identifying them for use as a comparison group, there is no shortage of them. But there is a real shortage of schools serving disadvantaged neighborhoods that do not participate in Chapter 1. By the rules that determine participation, a student cannot participate unless he or she is in a school that participates, and it is the characteristics of the school that determine whether the school participates. While it is true that some schools in disadvantaged neighborhoods do not participate if they are in economically poor districts, the extremely poor schools will all participate in that district.

The problem is portrayed by Figure 2 which shows Chapter 1 participation by schools having various percentages of students eligible for free lunches (a suitable surrogate for economic disadvantagement of the community). Clearly, there are nonparticipating schools in every economic category. But it can be assumed that, in the categories of schools having higher proportions of free lunches, the nonparticipating schools represent less disadvantagement than that of the participating schools. Thus, in a comparison of achievement, in the absence of Chapter 1, one would expect the less disadvantaged schools to score better. Whether this difference would apply equally to achievement *gains* is more problematic. In any case, the creation and careful use of a model of achievement appears to be a necessity.

Some Potential Sources of Comparison Groups

Since the characteristics of a school are presumed to be so important to student achievement, the construction of comparison groups within the sample of Chapter 1 schools used in the evaluation might be considered. This strategy would automatically eliminate school differences which, as pointed out above, are a potentially major source of noncomparability. The difficulty is that the students selected for participation in Chapter 1 in participating schools are the low achievers, while those who are not low achievers do not participate. Therefore, noncomparability between participants and nonparticipants is practically guaranteed by the method of selection.

It is true, however, that, except for kindergarten and possibly first grade, selection for participation is based upon testing which is subject to substantial error. Thus, there is a possibility that, by administering an independent test, one would find some substantial overlap in the independent test scores between the participant and nonparticipant groups. If so, the independent test score could be used as a regressor in an achievement model to adjust for differences in ability (as measured by the independent test). However, it is unlikely that many nonparticipants would have independent test scores comparable to those of the lowest achievers in the participant group. Thus, strong reliance would have to be placed on the validity of extrapolations beyond the range of actual observation. It seems unlikely that such reliance can be justified.

A second approach is to select a sample of schools as a comparison group. This approach was discussed above. It might be possible, through testing and a parental survey, to find nonparticipating students who are comparable to participating students on the bases of achievement and economic circumstances. But the opportunity for a nonparticipating student to learn may be substantially different from that of a participating student, with like similar achievement and economic circumstances if their school environment is fundamentally different. This concept underlies the process of allocating participation. Thus, again, noncomparability would be built into the system by this means of selecting the comparison group. It is possible that the noncomparability could be reduced through use of a model in which the school factors drive the adjustment. But, as in the case of adjusting for differences in student characteristics, if the comparison group were drawn from non-chapter 1 schools, strong reliance must be place in the validity of the model which includes school factors as variables.

A third approach would limit the comparison group to the schools that are Chapter 1 participants. It is presumed that, in some of the large center city districts, almost every school is "eligible" to participate, in the sense that its students are as economically disadvantaged, or more economically disadvantaged, than the students of some other schools in other districts that do participate. Thus, there is a tendency on the part of the administration of the large and economically poor districts to spread the limited amount of funds as widely as possible, even though there is Federal pressure to concentrate the funds sufficiently to produce a positive effect. In order to accomplish this objective, some of the treatments are very "light" treatments. On the other hand, there are some districts and schools in which the treatments are much more intensive, often providing a considerable amount of one-on-one teaching. There is, then, a spectrum of intensity of treatment among the participating schools. If this presumption is correct, one could develop a model of achievement in which a measure of intensity would be the principal regressor variable. Some considerable thought would need to be paid to the construction of the measure of intensity, perhaps based on the number of hours of teacher or teacher aid per pupil.

As in the first two approaches discussed above, the approach would require extrapolation to the situation of "no treatment", but the hazards would seem to be less, since observation could be made over a wider part of the total achievement spectrum.

This approach might be used in conjunction with a sample of students from nonparticipating schools in order to validate the extrapolation to no treatment.

The Need for Synthetic Cohorts

Public Law 100-297 requires an interim report to Congress not later than January 1, 1993 and a final report not later than January 1, 1997. These dates will require the contractor to complete work on the reports sometime in 1992 and 1996. These are the dates, then, that must be considered in the schedule. The Law also requires that "The study shall assess the impact of participation by such children in chapter 1 programs until they are 18 years of age." How can one measure the impact at age 18 of participation in (say) the lower grades when the first report is due in 1992? The problem is clear, but the solutions aren't.

One alternative is to do a retrospective study of 18 year olds, some of whom will have participated in Chapter 1 and some of whom have not. The difficulties seem insurmountable. School records likely will not be available for that length of time. Students will not remember (and, indeed, may not know) whether and when they participated. School dropouts (an important outcome) will be missed. And so on. Such an approach is simply not possible.

Another alternative is to do a series of short term longitudinal studies. In the extreme, two samples of each grade, K through 12, would be drawn in a given year. One sample would represent participants and the other would represent nonparticipants. Achievement would be measured as the gain in achievement during the year. Change in status (such as school dropout, or becoming a disciplinary case) would also be measured as the one-year change. The impact of participation in Chapter 1, for any combination of years of participation, would be found by aggregating the achievement indicator for years for which participation is assumed. For example, if a student participates in Chapter 1 during the third, fourth and fifth grades, and the net impact of such participation has been estimated to be 1.02, 1.04 and 1.01, respectively, the net gain from participation in grades three through six might be estimated as the product of 1.02, 1.04, and 1.10, or 1.07. (Whether the model should be multiplicative, as suggested here, or additive needs to be the subject of some study. In particular, whether the estimates of impact should consider participation in the previous year or two should be considered) Since no student is followed for the whole period, the hypothetical student group is referred to as a synthetic cohort.

The design sketched above is an extreme case of a synthetic cohort, since single years' changes are used to build up the estimation of impact for various lengths of participation in various grade ranges. It is clear that two-year links could be used, or three or four-year links in the case of the final report. If the links are too long, problems will occur with availability of school records and other interpretive matters. Some of these matters are discussed below.

Some Problems with Synthetic Cohorts

Stability of the system. If short term measures of achievement are to be linked, it is essential that the services provided under Chapter 1 be consistent over time. Is it reasonable to assume that services received by a first grader in (say) 1980 are the same as the services that will be rendered to first graders in 1990? And is it reasonable to assume that the impacts in 1980 are the same as they will be in 1990? These are questions that cannot be answered statistically, at least not in the time available before the final report must be submitted. There is also the problem of whether cumulative effects are more than or less than the product (or sum) of individual-year effects.. With links of two or more years one could accumulate some data on whether the impact is changing, but one certainly could not protect against the possibility of substantial change in very long term impact due to cumulative effects.

Comparability of test scores. Evidently, common practice is to use spring test scores as the baseline from which the next year's achievement will be measured. Although it may not correspond with practice everywhere, suppose the Normal Curve Equivalent (NCE) is used, both in the baseline measurement and in the posttest. Comparability of the pre- and posttest scores is compromised because one or the other of them may be obtained through the district's testing program and the other one through the special testing of Chapter 1 participants. More emphasis and motivation to try harder on the test are likely to be placed on the district's testing program. Students will be more mentally ready for it, it will be conducted more carefully, and provisions for testing absentees may be different. These factors will tend to reduce comparability and could create bias in the comparisons.

Measurement of achievement in nonparticipating schools. Presumably, test scores will be available in all participating schools, but only scores obtained from the district's testing program will be available for nonparticipating schools, unless at least a sample of the nonparticipating comparison schools can be persuaded to test annually. Even then, the cycle of testing may differ between the district containing the participant school and the district containing the comparison school. It is assumed that one cannot always find comparison schools in the same district as the participant schools. Otherwise, the comparison schools would always be less disadvantaged than the Chapter 1 schools.

One way out of the testing dilemma (which would work for the final report, but not the interim report) is to use only the district tests which are customarily given on something approaching a three year cycle. That is, gain in scores of Chapter 1 participants (after adjustment) would be divided by gain in nonparticipant scores (again, after adjustment) where the testing interval was n years, to arrive at a ratio, r. Then, by assumption, the annual gain would be a number, g, which, when raised to the nth power, would equal r. Note that not all district testing cycles would need to be identical. However, it would be important that comparisons be made among schools on the same cycle. The rules for aggregating across different cycles seem straightforward.

A problem would arise when a student in a Chapter 1 school had transferred from a nonChapter 1 school during the testing cycle or when, for other reasons, a participant at time of testing had not been a participant during the entire cycle interval. It seems likely, however, that rules could be worked out which would classify students as having participated "substantially."

The problem of school transfers. Transfers within the district pose a smaller problem than transfers between districts. The problem posed by transfers within the district is that students may be in a participating school one year and in a nonparticipating school the next. Thus, some definition of participation is needed. (See above.) Transfers between districts may cause loss of data because both test scores and participation may be unknown. Resolving such cases can absorb a lot of resources and, if the number of such cases is relatively small, it may be wise to simply consider them as missing data. The number of such cases will be a function of the length of the testing interval.

The problem of dropouts. It will be virtually impossible in a school-based sample to obtain outcomes for students who have not remained in school until age 18, the age chosen in the Law for which conclusions are to be drawn. But dropout is an important characteristic and may be considered a terminal outcome which is to be analyzed as a variable in its own right. This would mean that academic achievement would be measured only for those students who remained in school, thus becoming a conditional outcome.

Relatively few high schools participate in Chapter 1 programs, so the finding of suitable comparison group schools may be simplified for the high school grades. Finding 18 year olds who have participated significantly in Chapter 1 in elementary school, but not in high school, will be substantially more difficult, however. This puts additional emphasis on the need for creating synthetic cohorts.

Validation of the year-to-year linkage. The measured impact based on aggregation of relatively short term longitudinal comparisons may yield different results than would be obtained by comparing long term longitudinal gains directly. It seems important to compare the aggregate of short term impacts with the estimate of gains for longer longitudinal periods. This would not be hard to do if annual testing results (for Chapter 1 purposes) were entirely comparable to results of district testing. But, as pointed out above, the outcomes may be substantially different. One possibility is to conduct a substudy to calibrate an adjustment to the Chapter 1 test scores. Another is to cover two cycles of district testing in at least some of the sampled districts to check on the validity of the aggregation concept. It also seems advisable to include in the mathematical model a variable for previous participation.

The Special Problem of Evaluating at Age 18

The requirement to measure impact of Chapter 1 services on 18 year olds poses special problems. One of the problems is that most services to students occur in the elementary grades, and an 18 year old person may have participated in Chapter 1 during (say) grades 3, 4, and 5, but not since. Relatively few high schools participate, so if one were to limit the sample of Chapter 1 participants to participating high schools many, if not most, participants during some portion of their schooling might be excluded from the evaluation. The concept of the synthetic cohort, discussed above, may prove to be a satisfactory way to link together academic achievement, but the approach seems to be totally inadequate as a means for evaluating broader outcomes.

One possible approach to the problem is to draw a sample of high school seniors, determine their participation, and follow them into their post high school jobs. However, many, if not most, seniors would not know whether they had participated in Chapter 1. Also, records of the extent of their participation and academic achievement along the way would not be readily available, particularly if they had transferred across districts. Finally, dropouts would be missed. As suggested above, dropout might be considered an important outcome. But the suggested plan would not make it possible to identify dropouts.

Another possibility is to draw a sample of students in (say) grades 7 and 8 and follow them to age 18. This would require obtaining information on their previous Chapter 1 participation retrospectively and their future participation prospectively. Both might be feasible, particularly for students who didn't transfer across districts. If the sample were drawn early during the evaluation period it would be possible to make the evaluation for 18 year olds within the time frame specified in the law. It is evident that statistical adjustments to equate participant and nonparticipant groups would have to be made, and some attention would have to be paid to the extent of participation.

It is clear that such a plan would be costly, both in terms of acquiring the baseline data and in following the students to age 18. Equally difficult longitudinal studies have been conducted successfully in the past, however. It may well be the only feasible approach.

Some Comments on Possible Experimental Evaluation

If political considerations can be overcome, there would be a fine opportunity to conduct a randomized experiment to test the effectiveness of Chapter 1 on within-school achievement. Schools within participating districts could be paired in terms of a number of characteristics. One member of the pair could receive Chapter 1 support and Chapter 1 participation would be withheld from the other. The difference in achievement score gains would be a raw estimate of the impact of Chapter 1. It could be refined by some regression adjustments.

The approach would only work within large districts, but those may be of the greatest policy concern in any case. Also, districts have substantial leeway in selecting the grades to which the program applies. Thus, by staggering the withholding of services across the various grades, the plan might be applied for one or two years without serious disruption of the administration of the Chapter 1 program. That is, no student would have to be denied participation permanently. Estimates could still be made using the concept of a synthetic cohort. The decision to withhold treatment during a given year would have to be made as the result of a randomization process, however. I believe the potential gains from such an approach are great enough that it should receive serious consideration.

Some Notes on Sample Sizes

Much more attention needs to be paid to matters of design and estimates of variances than can be given here before any estimates of needed sample sizes can be made. All that can be done here is to point out some of the factors that affect sample sizes so that they can be taken into account during the planning stages of the project.

For these purposes, I will assume that the variable of interest is reduction in dropout rate prior to graduation from high school. I will also assume that a reasonable estimate of the gross dropout rate for economically disadvantages students is in the neighborhood of 40 percent The figure may have no validity--it is simply an assumption. An estimate with some validity may be made at a later date.

It is likely that the universe of students receiving Chapter 1 assistance will need to be subdivided into subsets (for example, males receiving Chapter 1 services below the fourth grade). Assume that some such subsets for which generalizations are to be made contain no more than one-eighth of the universe. Assume also that one wants to be able to detect a drop of five percent in the dropout rate with probability 0.95 in a five percent significance test. Some manipulation not presented here will show that, if random comparison and treatment groups were possible, one would need about 1,000 students (500 comparison and 500 treatments) to achieve the required precision.

It is not possible to construct simple random samples. Students must be grouped by school, which cuts the precision of the estimates. Offsetting this grouping effect is the fact that some matching of groups is possible. However, the effectiveness of such matching is questionable. This makes it virtually certain that at least 2,000 students will be required in each subgroup of interest. But, some saving is possible because not every subset needs to have its own comparison group. For example, a comparison group of students who have never received Chapter 1 services can be used as a comparison group for those who received services in the lower grades, in the upper grades, etc. In any case, it seems likely that a sample of at least 10,000 will be required.

The amount of speculation in these figures must be recognized. They are only intended to give a rough idea of the size of the project. Also, one must be aware of the level of precision that can be expected from such an evaluative study. In addition to the normal kinds of sampling error in such a study, there is the problem of noncomparability of the comparison group and uncertainty with respect to the models that adjust for such noncomparability. Thus, even with large samples, detection of small differences cannot be assured. However, if there are big differences between the gains of students given Chapter 1 services and those who aren't, most social scientists would be likely to accept the results.

Use of Data from Demonstration or Exemplary Schools

A question arose at the September 27 meeting concerning the use of data obtained from schools that were chosen in special ways, or that, having been selected by probability methods, received special attention in some manner. Such data can always be used in a national evaluation, but the weight it receives may be substantially different from that of the data collected in the usual manner.

Suppose, for example, that a school is known in advance to have an effective program and therefore it is to be included in the evaluation. Since its selection is not subjected to a randomization process, it enters the sample with certainty and receives a weight of 1.0 in the national estimates. Other schools receive a weight equal to the reciprocal of their probability of selection. This is true whether a single child has been included in the study or a thousand children. All that increasing the numbers of children accomplishes is to reduce the variability of the average for the school which is to be weighted by the sampling scheme.

If a school is selected according to the specified sampling plan and then is found to have a special program in which there is some interest, how can it be handled in the national estimates? The answer is the same as above. The school results must receive the weight specified by the sampling plan. Note, however, that one is not prevented from making conditional estimates for schools having particular charzcteristics. After such schools have been identified, the sampling fractions within them can be increased to provide additional precision for students having received the specified treatment. But in the national estimates, the average for the school must receive the weight specified by the sampling plan. The same rules apply, of course, for schools in which demonstration projects are conducted.

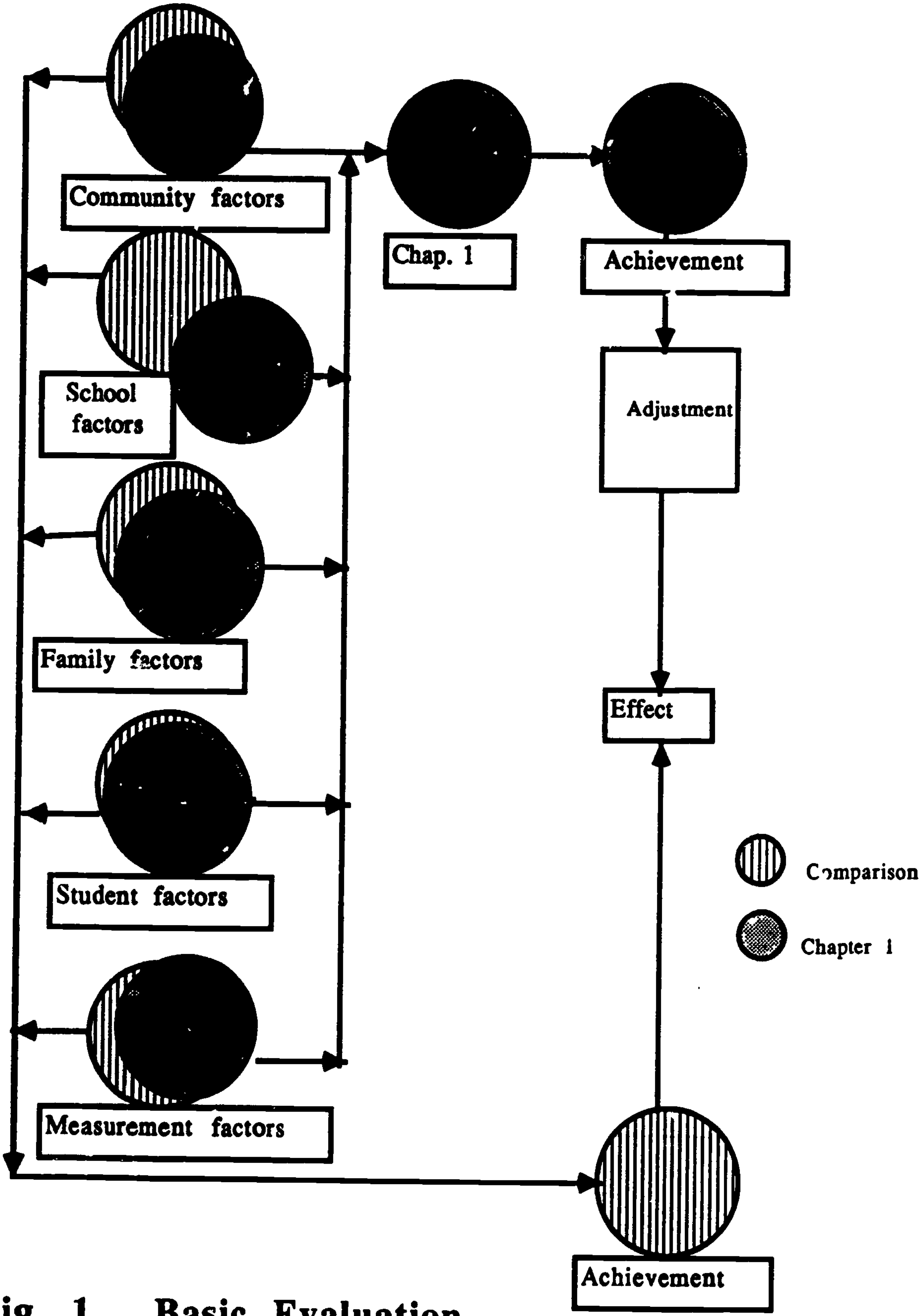
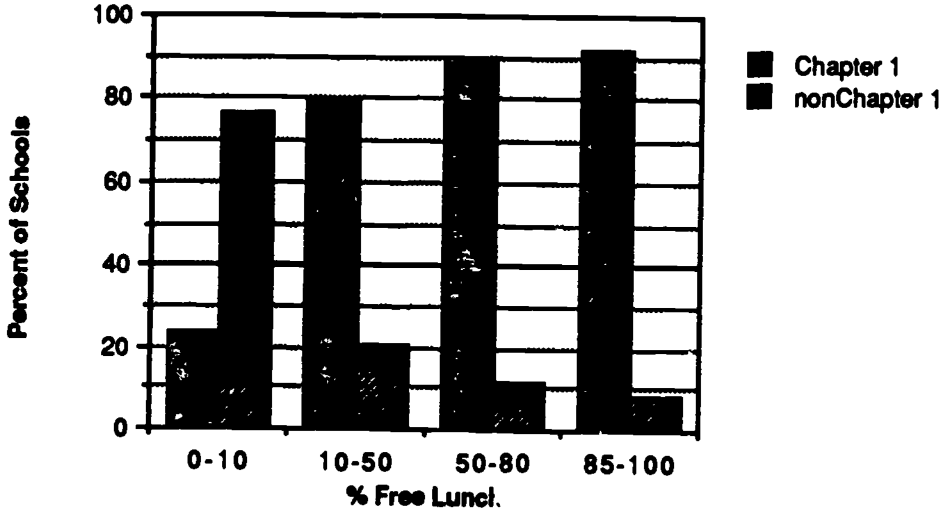


Fig. 1. Basic Evaluation Model

Fig. 2. Percent Chapter 1 of Elementary Schools by Category



Source: National Evaluation of ECIA Chapter 1 Schools, Westat, 1987

DESIGN PROPOSALS FOR STUDY OF CHAPTER 1 PROGRAMS AND THEIR EFFECTS

James S. Coleman
The University of Chicago

Initial considerations

The legislation provides an extraordinarily broad and demanding directive. An initial reading of the legislation seems to indicate an unachievable goal: to follow children now participating in Chapter 1 and report by January 1, 1997 on the effects of that participation up to 25 years of age. However, if this legislation is taken as an opportunity, it can prove extremely valuable: It provides a mandate, and with that mandate, the power, to initiate activities that can be very important for the future of educational evaluation, but would otherwise not be possible. For example, some standardization of the kinds of school records could be made a part of Chapter 1 participation, to make them more usable for subsequent evaluation than they currently are (as Jay Frechtling's briefing indicates). In addition, such a mandate imposes a demand on the researcher that can lead to more ambitious research designs, leading to answers to questions that were previously regarded as too difficult.

My initial reaction to the legislation includes the following points:

1. A major goal of the research should be to provide evidence about effects that is politically defensible; and politically defensible means scientifically defensible, for research results that enter the political arena are subject to closer scientific scrutiny than are purely academic research results. This implies that the research must be quantitative; qualitative research simply does not yield results that are defensible when scrutinized by unsympathetic investigators.

2. A major goal of the research should be to provide evidence not about "Chapter 1" as a whole: that it works or doesn't work, that it's good or bad. "Chapter 1" is not a well-defined educational intervention. Rather, the goal should be to provide evidence on which kinds of Chapter 1 interventions are most effective, and for what outcomes. The aim should not be to provide information to legislators that will help them decide whether to increase or decrease Chapter 1 allocations. It should be instead to provide information that will help in the decision of what kinds of intervention programs to put in place. This may have a secondary impact on the kind of Chapter 1 funding (for example, providing a force to increase expenditures through finding some extremely effective programs, or providing a force to decrease expenditure through finding that the way Chapter 1 programs are funded results in programs that are more often ineffective than effective). The research should aim to give information on how best to use Chapter 1 funds, not whether to have more or less funds.

3. The research should proceed on the assumption that whatever Chapter 1 programs are potentially feasible already exist. It should also proceed on the assumption that the evaluation of existing programs is going to provide more valid information than will the evaluation of programs newly designed for this study (because of Hawthorne-effect problems of the latter). Both of these considerations militate against demonstration programs. This should not be research testing out new ideas; it should examine how well existing ideas, having been put into practice, work.

Although control groups which have not had Chapter 1 programs can and should be included in the research design, principal emphasis should not be on comparison between "Chapter 1 and non-Chapter 1," because as indicated above Chapter 1 is not a well-defined educational input. Information from such comparisons will be the least valuable part of the research results.

The research should be carried out with the recognition that the kind of information it can provide is going to be increasingly useful, apart from the specific Chapter 1 question, for two reasons: a) As employment of mothers of young children an increasingly established institution, the institutionalization of pre-school children will become an increasingly established fact. This will generate increasing need for programs that are effective for pre-school and first grade.

b) The fraction of the birth cohort that is from disadvantaged families will increase, as the correlation between having children and socio-economic status becomes increasingly negative. This will mean that an increasingly large fraction of the next gereration will require some extra educational inputs if they are to be brought up to a level of productivity they will need as adults.

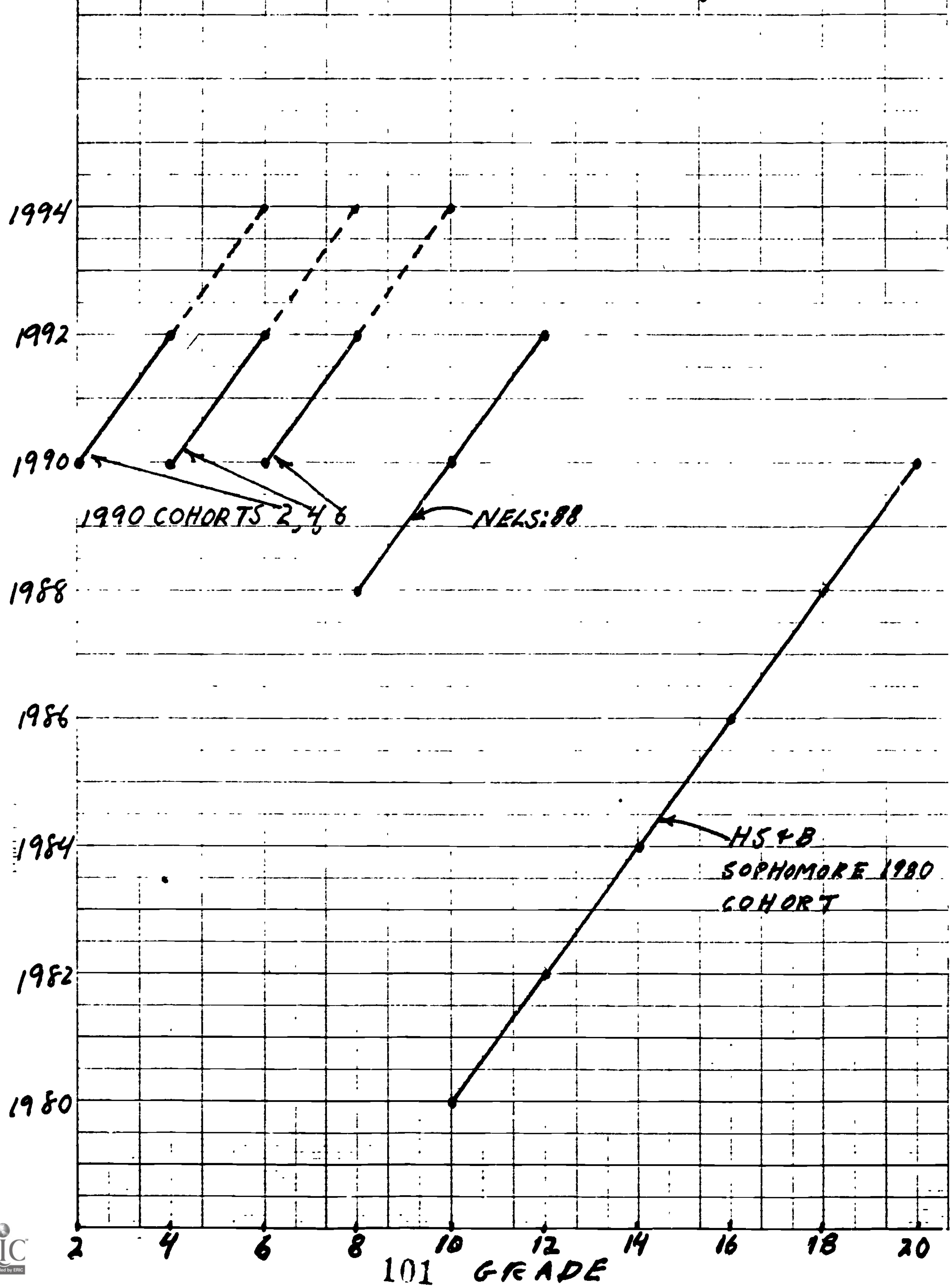
Specific research design recommendations

The principal target date should be taken as the date of the 1993 interim report: January 1, 1993. This imposes a very tight schedule. My proposal for an optimal research design is to use a series of two year modules, according to the design shown in Figure 1.

FIGURE 1 HERE

The design involves a set of linked cohorts, with each cohort having a data collection point in Spring of 1990, and a second point in Spring of 1992. The research would be designed to piece together the data from the set of cohorts to provide information on the long-term effects of specific Chapter 1 programs. As Figure 1 shows, the research would involve obtaining information on three cohorts (the 1990 cohort of 2nd, 4th, and 6th quarters), and using data from the NELS:88 cohort (8th grade in 1988) and the High School and Beyond sophomore cohort (10th grade in 1980) to link together with the data from these younger cohorts to provide information on long-term

Figure 1: Modalar Design for Study of Long-term Effects of Chapter 1 Programs



effects. In Figure 1, I have drawn broken lines extending each of the three younger cohorts to 1994. This is intended to make possible information that would be used for the 1997 final report. They are not drawn as solid lines to emphasize my point that the principal focus should be on providing an outstanding intirim report in 1993. It is feasible, within a data-collection framework of Spring 1990 - Spring 1992, to provide a strong ntirim report by January 1, 1993, and that should be the goal. (An experience from HS & B is relevant here: Data were collected in Spring 1980, and reports were available on Public and Private Schools, work during High School and Discipline in High Schools by September 15, 1980. This experience shows the feasibility of this timing.)

How can a modular design work? The success of a design like that shown in Figure 1 depends upon special data-collection procedures. Obviously, it cannot depend on 6th graders, 8th graders, or 10th graders remembering whether they participated in a Chapter 1 program, or on school records showing whether they participated or not. It cannot, for three reasons: First, neither students' memories not school records (given the move from one school to another as the student progresses in grades) can be counted on to provide such information. Second, in the case of NELS:88 and HS & B, the data have already been obtained, with no Chapter 1 participation information. Third, even if such retrospective data were available from students or from school records, they would not be what is needed, for they would not give information that would allow characterizing the program. As emphasized earlier, "Chapter 1" is not a well-specified policy input for children's education, and actual data on the programs is necessary in order to specify the properties of the inputs as experienced by children.

The success of the modular design depends upon being able to piece together one long causal chain from links in that chain. One's concern is with long-term consequences, say at age 21 or 25 (equivalent to "grade 16" and "grade 20" respectively in Figure 1). If researchers had more time than sense, they could attempt to discover the long-term effects of Chapter 1 programs by an extended input-output model: the inputs are Chapter 1 program variables at an early grade, say grade 1, and the outputs are things like school attainment, occupation, economic independence, and psychological well-being, say at age 25. If such an analysis found effects, it would be definitive, but not very helpful. It would not tell what the paths were through which there were effects, it would probably not provide specific information on the aspects of programs, and types of programs, that were effective, and it would not give information by which persons engaged in a Chapter 1 program could gauge the effectiveness of what they were doing.

The key to the modular design is the recognition that if a program that occurs at time t has some effect, n units of time later, at time t + n, this effect must take place through changes ir. some characteristics of the student that can be observed at time t + n-1; and these effects in turn must take place through changes observable at time t + n-2, and so n back to the starting point. To take a well-publicized case: If Head Start does have long-term effects, as the results of one study seem to indicate, these long-term effects did not suddenly blossom after the end of high school. Either some kinds of intermediate changes could have been observed throughout the period from Head Start to the point at which effects were observed, or there are no effects. Effects don't suddenly blossom after remaining submerged for ten or more years.

Implementation of a modular design requires recognition that the paths through which ultimate effects may occur are multiple, and that the changes that take place between time t and t+1 may involve characteristics of the child that are very different from those that are observed as ultimate effects of the program at time t+n. Thus to use the Head Start example again, it may be that the early research which looked at immediate or proximate effects did not cast its net widely enough, but looked instead too narrowly at the achievement measures that were of the same type as the ultimate outcome measures desired. As an illustration, suppose that there was an effect of a certain Head Start program on a child's sense of control, apart from any direct effect on verbal skills. Even if the direct effect on verbal skills washed out in the first year, suppose the sense of control did not. This might then have a long-term effect on later verbal and mathematical skills.

If this were in fact the case, and the investigators measured only achievement at t, t+1, and t+2, they might find that there was a differential gain in achievement from time t to t+1 due to the program, but that the difference washed out by time t+2. What they would miss is the second path through sense of control, a path through which the program had a long-term effect.

The general strategy, then, for a modular design in the study of Chapter 1 effects, must be to take a wide variety of outputs as potential changes from grade 2 to grade 4. These variables have been measured (in 1990) in grade 4, and are taken as potential input resources for changes from grade 4 to grade 6; and soon. Thus at grade 4, the dependent variables, in which potential consequences of a Chapter 1 program should be sought, must include things like absences, being late to school, attitudes toward school and toward self, parental involvement, discipline problems in school, grades in school, along with scores on standardized tests. All of these variables are not only dependent variables in 1992 for the 1990 grade 2 cohort. They are variables measured for the 1990 grade 4 cohort, where they serve as independent variables affecting changes at grade 6. The dependent variables at grade 6 include not only these same variables, but also initial measures of delinquency and drug involvement, as well as any other attitudinal or behavioral measures that could not have been manifested at grade 4. Many of these variables will be identifiable from HS & B analyses and NELS:88 analyses as precursors of dropout, and early pregnancy, and other variables of direct interest as outcomes. This could be conceived as a process of working backwards from the outcomes of interest to those precursor variables that show some effect on these outcomes, from those back to earlier precursor variables, and finally back to examination of the program variables on the early precursor.

Strategic variations and representative samples

The design of the sample for the grade 2 cohort should involve two components. One component should consist of a representative sample of programs to enable the question, "What is the effectiveness of Chapter 1 as currently implemented?" to be answered. This component of the sample will also be of value in determining what kinds of Chapter 1 programs, and what aspects of Chapter 1 programs, are most effective for particular (intermediate) outcomes, but it is necessary for the overall question. This component of the sample should be supplemented by a second component which might be called "strategic variations." These are Chapter 1 programs that are selected because they represent a wide range of variation in program goals and content, and because there is some prospect of their being effective programs. The principal value of the study of these programs should be the knowledge of what components of programs, and what kinds of programs, are most effective for particular outcomes. Although the representative-sample component of the total sample will aid in this, it is unwise to expect that the full range of program variation, with sufficient representation of each, will be found in a self-weighting representative sample of programs.

Obviously, these two components of the sample could be combined into a single sample design by using program types as strata, and sampling sufficiently within each stratum to insure that reliable statements can be made about programs in each of the strata. This would involve, of course, a pre-sampling characterization of the types of program variations.

Transactional analysis

All that I have written so far implies the kind of causal analysis that has become standard in quantitative studies of effects of educational variations. It is important to note, however, that something is captured in qualitative studies based on classroom observation that attempt to examine just what takes place in the classroom. Some of us at Chicago have been working on methods for bringing into quantitative analysis the study of transactions that take place in the classroom. These methods are in their infancy, but they could be especially valuable in aiding the characterization of a program. The methods involve the treatment of the classroom as a system of action, with extensive social exchanges going on between teacher and students, and among students. The principal use of these methods for the research on Chapter 1 would be to characterize the actual functioning of a particular Chapter 1 program, based on observation of what takes place in the classroom, and on analysis of these observational data.

It is not useful to go into the detail of these methods here. I will attach a paper which gives some description of their use with questionnaire data from High School and Beyond - although the methods themselves are more appropriately used with observational data.

10/21/88

Design for a National Longitudinal Study of Chapter 1

William W. Cooley

Professor of Education

University of Pittsburgh

Background

In section 1461 of public law 100-297 of 1988, Congress mandated that the Education Department sponsor a national longitudinal study of the impact of Chapter 1 participation on a broad list of outcomes: "academic achievement, delinquency rates, truancy, school dropout rates, employment and earnings, and enrollment in postseconday education." This mandate was inspired by the Perry Preschool study, which showed the impact of a well designed preschool program upon these broader outcomes of interest to society.

This encouragement to move beyond achievement test scores in thinking about the value of educational programs is certainly laudable. That aspect of the Perry Preschool study is clearly applicable to a national study of Chapter 1. However, the randomized design, which provided the logical basis for causally linking the Perry Preschool treatment with those subsequent outcomes, is not applicable in the case of Chapter 1. Therefore a different kind of study design is required in order to establish the causal links between Chapter 1 participation and these broader outcomes. This paper describes a design for such a study and the rationale for it. It utilizes an ethnographic approach as the primary method of data collection.

General Design Consider^tions

In designing this longitudinal study, the first requirement is to shift ones thinking away from experimental design, either randomized or non-randomized. There are two reasons for this. One is that Chapter 1 cannot be thought of as a treatment. Chapter 1 participation indicates possible access to a wide array of services which varies dramatically among participants at any given time, and varies dramatically for any given student over time. For plausible causal attribution in experimental studies (i.e., to be able to say that this program produced these effects) it is necessary to have a well defined treatment that is well controlled. Chapter 1 is not such a treatment.

The other reason to shift from experimental design thinking is that no comparable control group is possible. The best method of establishi ıg a control group, the way the Perry Prescnool study did it, is to randomly withhold Chapter 1 services frcm Chapter 1 eligible students. Random assignment is not an acceptable optior for such a well established program. Also, it is not a feasible option under the congressional constraint to conduct a 20 year longitudinal study in seven fiscal years, which requires some retrospective looks at what happened prior to initiating this study.

An alternative to randomization in establishing a comparison (control) group is to match on factors known to affect the outcome measures. Because of Chapter 1's targeting mechanisms, this is not possible. Certainly it is possible to find non-participants with the same test scores as participants, but closer examination inevitably reveals other significant differences which make them non-comparable, the most important of which is the probability that the "matched" non-participant is attending a Chapter 1 ineligible school. Such a school would tend to serve families with higher socio-economic status, which we know would give the control an advantaged educational environment (Birman, et al, 1987).

A alternative to experimental design is the explanatory observational study (e.g. Cooley, 1978), which makes it possible to estimate causal impact if one has reliable measures of all of the factors known to affect a reliably measured dependent variable, such as student achievement. This is the approach that guided the Instructional Dimensions Study (Cooley and Leinhardt, 1980), as well as many of the analyses of the Sustaining Effects Study (Carter, 1984). The problem with applying this approach once again is that we will learn nothing new. If the analyses are guided by an adequately specified model of tested student achievement, only very small effect sizes will be found for Chapter 1 services. The reasons why this is true and yet it is still possible for Chapter 1 services to have an educationally significant impact on the lives of disadvantaged youth is a long and complex story, but the main reason has to do with the fact that achievement test scores are very dependent upon the overlap between what was tested and what was in the curriculum. In the presence of measures of curriculum overlap, the effects of other treatment variables tend to be insignificant. For all of these many reasons, the recommendation here is to turn from quantitative efforts to find the Chapter 1 effect in student test score variance, to a study that is primarily qualitative in nature (see, for example, Patton, 1980 or Schofield and Anderson, 1987). The purpose of the study would be to show how Chapter 1 supported services is making a difference in the lives of Chapter 1 participants as well as help us understand the factors that lead to student failure.

A Focus Upon Student Failure

Previous efforts to establish the effect of Chapter 1 services have focused upon achievement test scores as the dependent variable. One reason for the enthusiasm surrounding the Perry Preschool study is how it showed the power of shifting to other outcome measures. It is hard to get excited about marginal increases in test score performance, not only because they tend to be so educationally insignificant, but because we know that achievement test scores are so weakly related to outcomes that people really care about. Achievement test scores are not as good an indicator of delinquency, truancy, dropout, employment, earnings and post-secondary education as are report card grades. The almost exclusive focus upon test scores in Chapter 1 evaluations has been unfortunate. Most test score differences are a function of who happened to be taking the test, what happened to be in the curriculum, and how "standard" the test administration happened to be.

I have been unable to find any national compensatory education study that has systematically looked at grades. But the "at-risk" literature has (for example, Wehlage and Rutter, 1986, Ekstrom et al, 1986, Bickel et al, 1986, Miller et al). The students who are at-risk of becoming a burden to society are the ones who fail the basic courses in school. I very highly recommend that this longitudinal study contribute to our understanding of the factors that lead to student failure as it seeks to document the ways in which Chapter 1 services are reducing the likelihood of student failure. Understanding failure includes understanding truancy, disruptiveness anu motivation as well as tested performance in academic skills. Teacher grades reflect those broader factors.

It is important to recognize that reducing early school failure is not just a matter of "fixing" (remediating) the students reading and mathematical abilities. In fact, it is often not just a matter of "fixing" the student. It is also important to consider ways in which the classroom or school could be "fixed", or the ways in which school-home relationships could be improved.

Another unfortunate aspect of the evaluations of federally supported compensatory education programs for the past twenty years has been the almost exclusive emphasis upon summative evaluation. In that search for proving Chapter 1's value, we have tended not to find ways to improve the program. The study suggested here can reveal ways in the which Chapter 1 services could be improved so as to reduce the likelihood of student failure.

Recommended Design

The best available method for establishing the impact of Chapter 1 services upon the lives of Chapter 1 participants is to directly observe the causal mechanisms that are operating in their day to day lives. This can be done by observing and noting what is happening during school and out of school. The observers need to notice what problems students are having in school and how Chapter 1 Interventions are helping. What is the student's school day like? How much direct instruction is occurring? How much of that is with a Chapter 1 supported teacher? Do the mainstream and Chapter 1 teachers plan together? What is the home like? What home factors are increasing the likelihood of school failure? How could Chapter 1 services be structured to reduce that likelihood? What happens during the summer? The way to answer such questions is to observe a student for two consecutive days and repeat that about six times each year.

Table 1 outlines the general structure for such a study. The overlapping longitudinal design makes it possible to study a 20 year developmental process in seven years. As outlined there, the study would begin with five cohorts, A to E. Cohort A, for example, would begin with first graders and follow them through grade 7. Cohort E, which begins with 19 year olds, is necessary if indeed you have to examine this process until age 25 by 1997.

One reason for starting with the grade levels suggested in Table 1 is that the first three years of the study would then cover the 12 years of schooling, and it seemed important to have that coverage for the interim report due in 1993. Another reason for the grade levels suggested is to cover major transitions within cohort, a transition being the movement from one type of school organization to another, such as elementary to middle school.

Before getting into sample size, let's examine the Congressional request to conduct the study "throughout the country in urban, rural, and suburban areas." To make things manageable for the type of study envisioned here, I recommend that the country be divided into as few regions as possible (i.e. as politically feasible). Table 2 suggests four regions. Any fewer would probably not be credible, and more would be less manageable. Four regional centers would be established for data collection. Within each region, three school districts would be enlisted, a rural, an urban and a suburban, all with average Chapter 1 partic'pation levels and range of services. One essential criterıon for district selection would be good student record keeping systems. This is important since cohorts B to E require knowing prior educational history, particularly the nature and extent of previous Chapter 1 participation, grades, attendance, and disciplinary actions. Within districts, students from grades 1, 4, 7 and 10 would be randomly sampled from among current Chapter 1 participants, or prior participants if there is currently no Chapter 1 service at that grade level.

Observers must have had prior experience in teaching in the schools. Recently retired or substitute teachers could be easily trained for this observational task. They would be trained in the production of field notes keyed to clock times, and done in a manner which would allow generalizations across students (Allington et al, is an example of this type of study in the Chapter 1 context). Each observer should be able to observe each student for about ten days over the course of a school year and for two days during the summer. That means about 15 students per observer, given the 180 day school year. (Some time must be allowed for training and planning sessions, illness, etc.) Total sample size will be a function of how much you want to spend on data collection, of course. An estimate of $2000 per student per year would be a rough guide for such estimates.

Let's assume that a total sample of 1200 is feasible. As Table 2 suggests, that would mean about 300 students per regional center, with 20 observers per center. Sampling across urbanicity could vary, as indicated. In es+ablishing the extensiveness of the sample it is important to recognize the fact that you are not trying to estimate population parameters (e.g., what percent of Chapter 1 students attend rural schools), but rather design a study that provides the demographic diversity that Congress requested, so that you can reassure Congress that the educational/oevelopmental processes that you are observing are not significantly different in these various demographic settings (or how they do differ).

Cohort E represents a special problem in sample selection. It would be better to start with 12th graders, but then they would tend not to be 25 by 1997. If it was not the intention of Congress (in section 1462 part [b]) to follow to age 25 by 1997, then I would drop Cohort E altogether. The ^ritical years in the transition from high school to post secondary are covered in cohort D.

One design consideration must be what to do about the high mobility of this low SES population. Because of the high mobility among schools within an urban district, I recommend that you start with a random sample of participants across the district, rather than select particular schools. You will soon be in all schools anyway as the selected sample transfers about, so you might as well begin with establishing contacts in all schools where Chapter 1 participants are found. That would also result in a richer variety of schools and classrooms. I also recommend that you follow students as they move from district to district, so that that aspect of the problem can be studied. This could be done by passing the responsibility for tracking a student from one regional center to another when a student makes a cross country move. Just following students who stay put would result in a very biased sample of this target population of low achieving, low SES students.

The linking variable across cohorts would be performance in school as measured by report card grades. It does not matter that grades may not be quantitatively comparable from one school context to another. What is important is that within a particular school context, failing grades is the best single indicator of an at-risk student. Having a linkin‿ variable is important in an overlapping longitudinal study. Although cohort A, for example, may not be followed long enough to establish ultimate outcomes such as dropping out, a pattern of failing grades is an excellent proxy for those subsequent negative outcomes that Congress hopes Chapter 1 is reducing.

What Could be Learned?

This longitudinal study of Chapter 1 participants could reveal how schools respond to early school failure. It could document effective practices and show that they often require a broader array of services than providing a remedial math &/or reading teacher for students who happened to have scored below an arbitrary achievement test score cut-off last spring.

While documenting the ways in which Chapter 1 funds are making a difference in the lives of disadvantaged students, it can also reveal why the studies of Chapter 1 impact upon student achievement have been so disappointing in the past. Such findings could have important implications for the improvement of Chapter 1 services.

This study could help Congress see that targeting students has prevented Chapter 1 from being as effective as it could be if the focus was on schools heavily impacted with children from families living below the poverty level. The study could help Congress to see that the big issues surrounding compensatory education--setting (pull-out or in-class), discontinuity of services from year to year, the narrowness of services, the lack of intensity of services, and the stigma of labeling students as disadvantaged--would be reduced or eliminated by targeting schools, not students. With schools as the target, it would then be possible to use Chapter 1 funds to apply what has been learned about improving schools and not worry about individual student eligibility. All students in such schools are operating at a disadvantage unless extra resources are available to make those schools effective.

TABLE 1

A Seven Year Overlapping Longitudinal Design

Cohort	90-91	91-92	92-93	93-94	94-95	95-96	96-97
A	1	2	3	4	5	6	7
B	4	5	6	7	8	9	10
C	7	8	9	10	11	12	19*
D	10	11	12	19	20	21	22
E	19	20	21	22	23	24	25

----

* 19 and up refer to approximate age group. Other cell entries are grade level. I would make cohort A largest, getting smaller as you go from B to cohort E. Total sample size of about 1200 seems about right.

TABLE 2

Distribution by Region and Urbanicity

	Region				
	(NE)	(SE)	(MW)	(W)	
Rural	100	100	100	100	400
Urban	150	150	150	150	600
Suburban	50	50	50	50	200
	300	300	300	300	1200

References

Allington, R., Stuetzel, H., Shake, M. and Lamarche, S. What is remedial reading? A descriptive study. ERIC ED254822

Bickel, W., Bond, L. and LeMahieu, P. Students at risk of not completing high school. Pittsburgh, PA Learning Research and Development Center, 1986.

Birman, B. et al. The Current Operation of the Chapter 1 Program. OERI: 1987.

Carter, L. "The sustaining effects study of compensatory and elementary education." Educational Researcher. 1984, 13. 4-13.

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

Cooley, W. and Leinhardt, G. "The instructional dimensions study" Educational Evaluation and Policy Analysis. 1980, 2(1), 7-24.

Ekstrom, R. et al. "Who drops out of high school and why?" Teachers College Record. i986, 87, 356-373.

Erickson, F. "Qualitative methods in research on teaching." In Wittrock (Ed.) Handbook of Research on Teaching. New York: MacMillan, 1986.

Miller, S., Leinhardt, G. and Zigmond, N. "Influencing engagement through accommodation: An ethnographic study of at-risk students" American Educational Research Journal. In press.

Patton, M. Qualitative Evaluation Methods Beverly Hills: Sage, 1980.

Schofield, J. and Anderson, K. Combining quantitative and qualitative compontents of research on ethnic identity and intergroup relations. In Phinney and Rotheram (Eds.) Children's Ethnic Socialization. Newbury Park, CA: Sage, 1987.

Wehlage, G. and Rutter, R. "Dropping out: How mulch do schools contribute to the problem?" Teachers College Record. 1986, 87(3), 374-392.

Issues in Designing a National Study of Compensatory Education

Gary Echternacht
Educational Testing Service

Background

Since beginning in 1964, compensatory education provided through federal Chapter 1, formerly Title I, funds has undergone continual evaluation. Evaluation of the program occurs not only at the local school district level, but also at the national level through studies funded by the U.S. Department of Education. Evaluation of the program will continue, as a major national evaluation was authorized in 1988 in the new Chapter 1 law.

The U. S. Department of Education conducted several national studies early in the program (e.g., Wargo, et al 1972, Hendrickson, 1978, Carter, 1980, OERI, 1986 and 1987). All attempted to estimate the effect of the program on raising student achievement as indicated on standardized test scores. The findings ranged from small to no effects. All these studies had significant design problems and the results often were criticized by program advocates.

The most thorough study of compensatory education began in 1974 and was known as the sustaining effects study. It was a three-year longitudinal study looking at achievement in reading and mathematics for students in grades one to six. The study used a control group obtained from small schools without compensatory education programs and from schools that had no compensatory programs. Standardized achievement test results administered specially for the study were the primary outcome variables. Only small positive effects were found, stronger in mathematics than in reading. Nevertheless, the effects were not carried over to the next grade level.

The most recent study of compensatory education was conducted by the Office for Educational Research and Information. Completed in 1987, the study findings were presented in three reports:

- The first report established the link between being in poverty for a long time and remaining in poverty.
- The second report rehashed the sustaining effects study and came up with the same findings.
- The third report synthesized a potpourri of specially funded studies regarding operation of the program.

T'ꜛ first study provided a lesson in the politics of evaluation and compensatory education for all involved with the program. The study presented strong evidence that the more time one spent in poverty, the more likely one was to remain in poverty. If the goal of compensatory education were to help people out of poverty, the study findings suggested that program funds be more heavily concentrated in areas with histories of long-term poverty. Ninety percent of all the school districts receive Chapter 1 funds. Given the current emphasis on limiting government spending, the study findings suggested that program funding be redistributed to put more funding into areas with significant long-term poverty. This suggested policy was criticized by much of the Chapter 1 status quo who wanted no change in the methods for distributing funds. The proposal went nowhere in congress.

With the exception of the last study, studies have attempted to estimate the effect of the program on student performance on standardized tests. The results have been consistent and not terribly useful in the sense that the programs have changed little, if at all, as a result of the national evaluations. The programs have changed, but the changes stem primarily from movements within the content areas or to a lesser extent, through local program evaluations. For example, there is a movement within compensatory education to emphasize the teaching of reading comprehension. This has come about because of the research in reading rather than to evaluation research.

Because there is a history of evaluation in compensatory education, the design of the current study needs to look beyond the immediate question of whether or not the program has an effect on student achievement as measured by standardized test results. Tough design issues need to be faced directly, so we can reach a better understanding of why programs work or do not work. In this paper, I address five issues that are basic to the national evaluation. They are:

- what are the appropriate outcome variables?
- what is the treatment?
- what do we want to study?
- are there any control groups?
- what concept of evaluation shall we employ?

I will go on to argue that the answer to these questions will be to develop a multidimensional index of achievement for the outcome variables, that the treatment must be considered both as a funding source and as a set of instructional conditions, that our goal should be to better understand why programs work rather than to estimate their overall effects, that there are no control groups, and that we must consider a threshold attainment as well as a gains design.

What are the appropriate outcome variables?

Historically, standardized reading and mathematics tests have been used as outcome variables. These are the same tests commonly used by schools throughout the country. Although they are not the only outcome measures used, they are by far the most commonly referenced.

The advantage of using standardized tests as outcomes is that they are relatively independent of the school curriculum. Although there is certainly a great deal of overlap between school curricula and standardized test content, the tests are not directly tied to a specific curriculum as are end of unit or year tests supplied by the textbook producer. In that sense, they represent an independent application of the knowledge and skills taught in the classroom. This provides test scores with generally perceived credibility. Their results are also comparable over different school districts, and they are relatively easy to collect and analyze.

Standardized achievement tests do have some major disadvantages, however. In particular:

- they do not represent a national standard of performance
- they are limited in content
- they provide only an indirect measure of the real purpose of compensatory education

Standardized achievement test publishers go to great expense to make sure that the content of their tests is a representative sample of the content taught throughout the country. Nevertheless, even though the content of standardized tests may adequately represent the grade level content taught, there is no performance standard set for the tests. Publishers have left that task up to their users and users have done a poor job in setting performance standards. For example, if a compensatory program student obtains a score at the 45th percentile rank at the end of the school year, schools seldom attach an adjective such as good or bad to the score. When performance standards are set, they generally are set arbitrarily and are unrelated to classroom performance. There are few cases where school people have attempted to relate levels of test score to either current or future classroom performance.

Even if schools set standards for performance, those standards would surely differ. Standards at suburban schools with little poverty, many resources, desirable teaching conditions, and high socio-economic status students are likely to be higher than at inner-city schools with high poverty, few resources, undesirable teaching conditions, and low-socioeconomic status students.

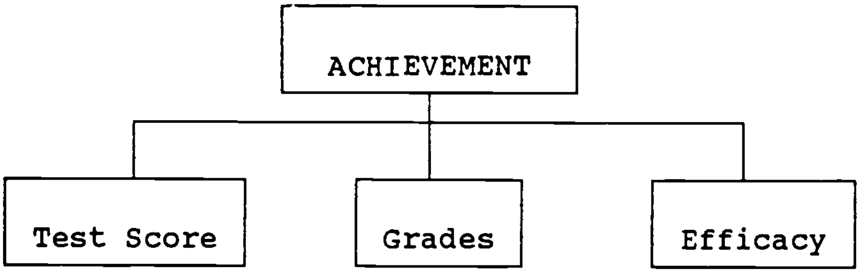
Tests are also limited in content. This is a necessary feature of standardized testing with young people as they begin to fatigue during testing after about 45 minutes. This means that a year's worth of content must be tested in that short time. Limitations are affected by the nature of the subject being tested. For example, reading comprehension is affected by knowledge in the content areas. For that reason, it is most desirable to sample reading passages from many content areas. But to measure understanding of a reading passage rather than simple literal recall, it is necessary to have a sufficiently long passage so that elements of the passage can be related and interpreted. This trade-off between passage representation and length faces every reading test developer and necessarily results in limiting the reading content in a reading comprehension test.

Perhaps most importantly, however, standardized tests provide only an indirect measure of the purpose of compensatory education. One can argue that the real purpose of compensatory education is to provide extra help to students who are having trouble in class so they can "get along" in the regular class setting. How would you know if the program was accomplishing this generally accepted goal? A test score only helps indirectly. It gives a general picture of achievement, not directly related to classroom performance. In-class performance is most directly measured by classroom grades in areas directly related to achievement in the subject. If a compensatory education student in reading receives satisfactory marks in areas directly related to reading achievement, then the program is accomplishing its goal.

Grades or marks have their own set of disadvantages. Systems of grading differ from school district to school district. Grades are not comparable over different teachers. Sometimes grades are influenced by judgements based on factors other than achievement. Nevertheless, they represent an indicator of achievement that is related to future achievement and commonly used in educational studies.

Compensatory education programs aim to do more than simply raise achievement. Most programs aim to instill a value for achievement and an efficacy for learning. Teachers try to help students believe that they can achieve and that achievement is a valued product of education. This goal is sometimes stated in the applications schools make for compensatory funds and the manner in which teachers work with students. Rarely is this aim measured in an evaluation.

The appropriate method for dealing with the outcome issue, in my judgement, is treating the outcome as a multidimensional construct consisting of standardized test performance, classroom grades, and an assessment of student efficacy in achievement. The outcome would be represented by an appropriate index made up of a composite of test score, grades, and a self efficacy scale. The figure below illustrates the index.



Using grades and another scale requires developing systems for making different grading systems comparable and selecting or developing the self efficacy scale. Certainly, grades and self efficacy would be less reliably measured than test score. Nevertheless, interpretations from the study would be more valid than they would be by using test scores alone.

<u>What is the treatment?</u>

There are two ways to define what a Chapter 1 treatment is. On the one hand, you can think of compensatory education as a funding source. It is the collection of all instructional and support activities that are funded under compensatory education. On the other hand, you can think of compensatory education as a group of specific instructional practices.

Past evaluations of compensatory education have used the first definition. Compensatory education has been considered as all those services paid for with compensatory education funds. A student was considered in the program if the student was receiving any of these services. This way of defining the treatment gave researchers a clear method of deciding who was in the program and provided a complete coverage of all the types of services that were provided.

But there are disadvantages in taking this approach. In studies thus far conducted, results have been general and applied to the program as a funding source rather than a specific instructional application. Findings at this level may have been useful to national policy makers, but were of little use to people who designed and carried out programs in schools.

As people began to conduct longitudinal studies, other problems were identified. One might be called the longitudinal problem. If one identifies a cohort of people who are in compensatory education in the first grade and follows that cohort through sixth grade, a number of different participation patterns emerge. In each of the succeeding years past first grade, an individual student may be either in or out of compensatory education. Over five grades it means there are 32 different participation patterns ranging from only participation in the first grade to participation in all six grades. For a student participating only in the first grade, compensatory education is only providing a short "dose" that presumably prevents the need for further help. For a student participating in all six years, compensatory education is an integral part of that student's schooling. At the individual student level, the program is very different for these extremes.

Although researchers have considered compensatory education as a funding source, there is no reason why it cannot be considered as a set of specific program types or models. One could develop a classification system for compensatory education programs and sample from that classification system when evaluating the program at a national level. Programs using the same general methodology and materials could be considered equivalent for the study.

Using such an approach would provide a clearer picture of the nature of the treatment being studied. In that sense, it would provide information to those who design and conduct programs and lead to improvements in the program rather than summative judgements of the value of the funding. If the study sampled in such a way that a wide range of instructional applications were studied, we could assess their relative effectiveness.

There are disadvantages to this approach. There are many different program types. It is not feasible to include all in any national study. The longitudinal problem is still there. And, it does not allow for differing effects by grade level, or the comparability of programs at different grade levels.

This last area is most significant. Most compensatory education people believe that programs should be preventative. They should be placed in the lower grades so that difficulties with achievement later in school can be avoided. It is based on the notion of individual differences in achievement among students when they enter school and that achievement is highly correlated from year to year. In other words, they believe students who start low in achievement remain low in achievement throughout school.

There is little evidence supporting this notion and much suggesting that it simply may not be true. For example, although there are individual differences among students when they enter school, those differences increase through the years even with compensatory education. This is a common pattern in any area of developed characteristics. There is evidence that fragmenting reading instruction by having a regular classroom and a compensatory teacher both work with a student is not effective in the elementary school grades (Allington, 1986). There is evidence that although girls achieve better than boys in mathematics in the early primary grades, this pattern of achievement reverses in the secondary grades (Marshall and Smith, 1987). In the elementary school grades compensatory education is rarely supplementary in the sense that extra school time is provided for instruction in the basic skills (OERI, 1987). In fact, compensatory education students receive little more instruction in reading and mathematics than do non compensatory students. Finally, we need to realize that the achievement that we should be most concerned about is the achievement students leave school with. In that sense, achievement in the higher levels of basic skills, for example reading comprehension and mathematics problem solving, are most important. These are developed later in schooling.

In the national study, I argue, we should approach the problem of defining the treatment by using both methods. We need to look at the program as a funding source so that we can retain continuity of findings with previous research. Nevertheless, given the political climate, it is best to devote most resources to studies that consider the treatment as a collection of specific instructional applications. Compensatory education has wide support in both the congress and in education. People believe the program is effective whether a study finds it so or does not. What is needed is an examination of specific instructional applications and their effectiveness which can lead to improvements in program design and implementation.

What do we want to study?

The first compensatory education evaluations emphasized the effects of the program on achievement. Effects were defined as statistical effects, that is, achievement above that which would occur without the program. Estimating the statistical effects of the program has come to dominate all evaluations since.

I would argue that the emphasis on overall effects is no longer appropriate as a study objective for both statistical and political reasons. The statistical reason is based on our inability to adequately estimate how people would achieve without the program. The political reason is that congress supports compensatory education regardless of what evaluation studies find. The important questions a national study should address concern program design and implementation. Those questions are:

- In what grade spans (e.g., pre-K-K, 1-3, 4-6, 7 and above) are compensatory programs most effective in terms of current and future classroom performance?
- What is the relative importance of various design factors that school districts can control on student performance?
- What is the relative effectiveness of various program components (e.g., parental involvement, extracurricular activities, outside class learning) on performance?
- What models for integrating compensatory education with the whole school experience are most effective in improving performance?

We should realize that these are difficult questions. The concepts are vague. Measurement is poor. Statistical models are only marginally helpful. But the questions are right. It is better to answer the right questions with weak models than it is to apply strong models to the wrong questions.

Are there any control groups?

When we estimate the effects of a program in a statistical sense, we mean the difference between the observed outcome and what the outcome would have been without the program (Rubin, 1977). This later quantity is unobservable, but can be estimated if we have a quantitative model of how people are selected for the program and knowledge of the functional relationship between the outcome and selection variables. The most powerful results are obtained when students are selected for the program at random.

In school settings, assignment to compensatory education is never random. Schooling is not an experimental enterprise. Nor is it possible to construct an adequate quantitative model of selection. Students are selected by teacher judgements and other factors in a nonsystematic manner. The functional relationship between the selection and outcome variables is also problematic in that it is usually nonlinear when the variables are test scores. In snort, we cannot get a statistical effect estimate from our evaluation studies because randon selection is not possible.

Indeed, we cannot even get a good quasi-statistical effect estimate. Within any school at any given grade level having compensatory education, all of the low achieving students will be in some type of compensatory education. If we look for a school that is not eligible for compensatory education and apply the same selection procedure, we not only confound a large school effect with the program effect, we also find there are too few comparison students identified. Also, some elements of compensatory education will have made their way into the instructional programs. When school people find effective practices in compensatory education, they try to institutionalize those in other schools.

Because there are neither control groups nor good comparison groups, effectiveness must be determined relative to a standard treatment that the evaluation study team must define. I would suggest that the study should first find the most common type of program. Outcomes for those programs should be determined. This would be a standard against which other programs would be comparred.

What concept of evaluation should we employ?

Past evaluations have always approached their tasks by applying quasi-experimental designs to gain scores. The existing methods for evaluating local projects consists of obtaining gains for participating. The general idea is that if participants are gaining in achievement in relation to some standard of gain, it is good. This is the way that most educational programs are evaluated.

There is another way, however. It involves setting a performance standard and seeing how many people meet this performance standard. This is the philosophy behind many state testing programs and all licensing and certification test programs.

Under this approach, a threshold is set for minimum performance. In a national study that threshold would be the minimum achievement needed to succeed in a grade. It is a pass-fail system. The proportion of students scoring above this threshold is the key evaluation statistic. The idea behind this type of evaluation is that students must achieve at certain levels to progress through the educational system. Gain is irrelevant. It matters little how much one is gaining if those gains result in achievement that is inadequate for grade-level performance. The purpose of compensatory education is to help people succeed in the regular classroom. An evaluation approach along this line is more directly related to the purpose of compensatory education that is an approach emphasizing gains.

References

Allington, R. Policy constraints and the effective delivery of remedial instruction: a review, in Hoffman, J.(ed.) The Effective Teaching of Reading: Research and Practice. Newark, DE: International Reading Association.

Carter, L. The sustaining effects study of compensatory and elementary education. Educational Researcher, 1984, 13(7), p. 4-13.

Hendrickson, G. Review of Title I evaluation studies. DHEW Office of the Assistant Secretary for Planning and Evaluation, 1978.

Marshall, S. and Smith, J. Sex differences in learning mathematics: a longitudinal study with item and error analysis. Journal of Educational Psychology, 1987, 79(4). p. 372-383.

Office of Educational Research and Information. Poverty, achievement and the distribution of compensatory education services, U. S. Department of Education, 1986.

Office of Educational Research and Information. Current operation of the Chapter 1 program, U. S. Department of Education, 1987.

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

Wargo, M., Tallmadge, G., Michaels, P., Fipe, D., and Morris, S. ESEA Title I: a reanalysis and synthesis of evaluation data from fiscal year 1965 through 1970. Palo Alto, CA, American Institutes for Research, 1972.

A DISCUSSION OF SOME STATISTICAL SAMPLING

ISSUES RELATED TO THE

PROPOSED CHAPTER 1 LONGITUDINAL STUDY

BY MARTIN R. FRANKEL, PH.D.

The purpose of this document is briefly articulate several statistical sampling issues that will arise in conjunction with the planning, design and implementation of the proposed longitudinal study of Chapter 1. This proposed study of the impact of Chapter 1 on participants until ages 18 and 25 is mandated by the Hawkins-Stafford Amendments. The legislation indicates that educational achievement of children with significant participation in Chapter 1 programs should be compared with that of comparable children who did not receive Chapter 1 services. This clearly requires some form of comparison or "quasi control" or even "randomized treatment - control" group analysis. Further, the legislation indicates that "(t)he study should be conducted through the country in urban, rural and suburban areas, and should be of sufficient size and scope to assess and evaluate the effect of the program in all regions of the Nation."

ISSUE: BASIC FORM OF THE STUDY

The time requirements for reporting results as well as the basic nature of the Chapter 1 program would appear to restrict the options that are available in the basic design of the study. Ideally, a fully defensible and sound study of impacts should be based on a fully randomized design. Such a design would be based on a selection of a random sample of potential participants followed by a random assignment of these potential participants to treatment and control groups.

Given the time and program constraints, a more realistic design will probably involve the use of random samples of "naturally generated" participants and non-participants with data collection that involves a longitudinal component as well as some degree of retrospective data collection or record retrieval.

Because of the widespread utilization of Chapter 1 funding it is not clear that it will be possible to find a non-exposed comparison sample that may be demographically or statistically matched to the participant sample. In this case the impact analyses will be forced to more generally rely on comparison of outcomes among individuals with different levels of Chapter 1 program exposure.

ISSUE: Analysis Plan - Sample Size

In addition to the general form of the study itself, one of the crucial issues that must be faced is that of sample size. In order to approach the question of sample size it is first necessary to describe the nature of the analysis plan. It is only in this context that the adequacy of any sampling strategy and sample size may be assessed.

From the standpoint of sample size assessment, the nature of the analysis plan may be viewed in terms of the following question:

What are the basic assessment and evaluation measures?

a. Means, Proportions

b. Differences between Means or Proportions

c. Regression coefficients

d. More complex statistics

The ability to answer this question involves a general agreement about the nature of the overall analysis plan. While the expectation that the overall analysis plan be known may seem somewhat premature, it is only after the nature of the inference problem is known that a reasonable assessment of sample size is possible.

ISSUE: Representation of Regions

The legislation authorizing the Chapter 1 study is somewhat unusual since it contains rather specific language regarding the scope of sample coverage.

> "The study shall be conducted throughout the country in urban, rural and suburban areas, and shall be of sufficient size and scope to assess and evaluate the effect of the program in all regions of the Nation"

This language is open to several interpretations, but it appears to indicate that the sample should be distributed throughout the entire nation and should support <u>separate</u> analyses for urban, suburban and rural areas as well as <u>separate</u> analyses among various geographic regions.

Within this general requirement there a number of technical questions that must be answered :

* What are the appropriate definitions of Urban, Rural, Suburban?

* What are the definitions of Regions?

  (Census regions (4) or divisions (9) or other)

* Are the precision requirements the same for Total US and the various sub-domains (urban, rural, suburban, regions)?

DEFINITION OF URBAN, RURAL AND SUBURBAN

There is general agreement with resect to the definition of the classification of rural versus metropolitan (non-rural) areas. Most researchers are comfortable with the use of the Metropolitan Statistical Area (MSA) in this context. More specifically counties that are not contained within an MSA are considered rural and those falling within an MSA are considered metropolitan or non-rural. Table 1 below show the distribution of the total US population among Metropolitan and Rural counties:

TABLE 1

METROPOLITAN VERSUS RURAL COUNTIES

AREA	POPULATION PERCENT
METROPOLITAN: COUNTIES INSIDE CMSA's (PMSA's) AND MSA's	(76.5%[1])
RURAL: COUNTIES NOT IN CMSA's OR MSA's	(23.5%)

There is generally less of a consensus regarding the appropriate subdivision of metropolitan areas into areas that are considered Urban and those that are considered Rural. The most commonly used definition involves the US Census designation of Central Cities. Each MSA contains one or more cities that are given the designation Central City. Some definitions of Urban versus Suburban areas define urban areas as those areas within Central Cities of MSA's and define suburban areas as those areas within MSA's that are not within Central Cities. Under these definitions, approximately 30.0% of the US population is classified as Urban (Central Cities of CMSA's and MSA's) and 46.5% of the US population is classified as Suburban (Balance of MSA's and CMSA's).

For the proposed study of Chapter 1, it might be more appropriate to examine a more refined definition of Urban area which captures the concept of inner or core city. In this case it might be appropriate to subdivide Central City areas into inner central city areas and remaining central city areas.

DEFINITION OF REGIONS

The two most commonly accepted definitions of geographic subareas follow definitions used by the US Census. The Census has subdivided the 50 United States on a state basis into 4 geographic REGIONS: Northeast, Midwest, South and West. These geographic regions are further subdivided (using complete States) into 9

[1] Percent of total population 1985, Statistical Abstract of the US 1987, Table No. 33.

geographic DIVISIONS. New England, Mid Atlantic, East North Central, West North Central, South Atlantic, East South Central, West South Central, Mountain and Pacific. The exact State definitions of these areas as well as the percentage population distribution is shown in Table 2

TABLE 2 GEOGRAPHIC REGIONS[2]

AREA	PERCENT
NORTHEAST	(19.4%)
NEW ENGLAND (ME,NH,VT,MA,RI,CT,	(5.1%)
MID ATLANTIC (NY,NJ,PA)	(14.3%)
MIDWEST	(24.1%)
E. N. CENTRAL (OH,IN,IL,MI,WI)	(17.0%)
W. N. CENTRAL (MN,IA,MO,ND, SD,NE,KS)	(7.1%)
SOUTH	(35.1%)
S. ATLANTIC (DE,MD,DC,VA,WV, NC,SC,GA,FL)	(17.3%)
E. S. CENTRAL (KY,TN,AL,MS)	(6.5%)
W. S. CENTRAL (AR,LA,OK,TX)	(11.3%)
WEST	(21.2%)
MOUNTAIN (MT,ID,WY,CO,NM,AZ,UT,NV)	(6.2%)
PACIFIC (WA,OR,CA,AK,HI)	(15.0%)

[2] Percent of projected 1990 population, Statistical Abstract of the US 1987, Table No 28.

PRECISION REQUIREMENTS FOR TOTAL US AND SUB-DOMAINS

In addition to evaluating the sample size in terms of the overall precision requirements for the total US, a similar evaluation must be undertaken for the various sub-domains (urban, rural, suburban, regions, divisions) that are to be the subject of separate analysis. Further, to the extent that it is desirable to equalize the sample size among the various sub-domains (rather than accept the proportionate allocation) it should be recognized that the requirements for data weights will impact the overall statistical efficiency.

More specifically, it is possible to alter the allocation among the various sub-domains so that it is not proportional to the population. Typically this modification it toward equal sample size among the various sub-domains.

This modification necessitates the use of "weights" when the sample is used to produce national estimates. While this process provides a great deal of design flexibility it extracts a price in terms of increased standard errors for overall estimates. This increase in standard error may be expressed as a quantity called "Statistical Efficiency." The statistical efficiency expresses the statistical reliability of estimates produced by the weighted sample to that of a proportionate simple random sample. For example, if the statistical efficiency of a non-proportionate sample of 1,000 cases is 90%, then for overall estimates, a proportionate simple random sample of 900 cases (0.90 x 1,000) would have the same statistical reliability.

Table 3 shows the statistical efficiency of three possible non-proportionate allocations based on equal sample sizes among the various sub-domains classifications that have been considered in this document.

TABLE 3

STATISTICAL EFFICIENCY OF VARIOUS SAMPLE ALLOCATIONS

ALLOCATION	EFFICIENCY
FOUR (4) REGIONS EQUAL	94.4%
NINE (9) DIVISIONS EQUAL	85.0%
THREE (3) URBAN/SUB/RURAL EQUAL	92.2%

ISSUE: Sample Mobility

One of the critical issues that must be addressed in any longitudinal survey is that of sample mobility. When sample selection and interviewing in the base line survey is carried out in schools, then successive years of data collection must recognize that in subsequent years students may not necessarily be found in their base year schools. This movement may be the result of several factors including: natural progression out of the grade range for the school, movement of a student's home to a different school district, change in school district boundaries, change from public to private school, dropping out, etc.

In most situations, it is absolutely critical that a longitudinal study follow movers. This following need not necessarily be on a 100% basis, but rather it may involve a probability sampling process. It is imperative that movers not be excluded from subsequent data collection since this will leave the study burdened with facing the strong possibility that impacts may be confounded by mobility.

The generally increased costs associated with following movers must be included in the planning of the study. Estimates of moving rates should be obtained so that the increased cost does not come as a "surprise" to the study sponsors.

A good source for basic data to inform assumptions about mobility may be found in other longitudinal studies of students within schools.

Tables and 5 contain information about moving obtained from NELS88 and High School and Beyond respectively

TABLE 4

NUMBER OF TIMES
CHANGED SCHOOL BETWEEN FIRST AND EIGHT GRADES
AS THE RESULT OF CHANGE OF RESIDENCE

PERCENT OF SAMPLE	# TIMES
44.0%	0
22.8	1
10.7	2
9.7	3
5.7	4
5.7	5 or more
1.4	DK/NA

TABLE 5

NUMBER OF TIMES
CHANGED SCHOOL BETWEEN FIFTH AND TENTH GRADES
AS THE RESULT OF CHANGE OF RESIDENCE

PERCENT OF SAMPLE	# TIMES
64.0%	0
16.0	1
7.0	2
10.0	3 or more
3.0	DK/NA

ISSUES IN LONGITUDINAL ANALYSES OF CHAPTER 1 DATA

Paper developed for the U.S. Department of Education for use in planning the national longitudinal study of Chapter 1

Joy A. Frechtling

September, 1988

The U. S. Department of Education has been asked by Congress to plan and conduct a national longitudinal study of Chapter 1. The basic purpose of the study is to examine the extent to which Chapter 1 participation improves the academic and social outcomes of disadvantaged children. Because an important part of this study is assessing program impact on students who may be in their late teens or even older, a retrospective look at efficacy is needed. A critical issue in conducting such a study is determining the extent to which needed data on participation, services, academic and social outcomes can be obtained from existing records.

The purpose of this paper is to discuss issues related to the collection of retrospective data from student records. In doing so, three concerns have been kept in mind: Are the data retrievable? Are the data likely to be accurate? What ar: the probable costs/logistical efforts required to obtain them? The critical data elements considered are:

- names of former Chapter 1 participants
- outcome measures such as grades, test scores, and attendance
- Chapter 1 program descriptors
- descriptors of other "compensatory education services"
- information on comparison students

In addition, a brief discussion is also presented of questions regarding the methodology to be used. The values of a representative survey vs. a purposive study of "successful sites" are considered.

Identifying former Chapter 1 students

At the heart of doing a study of the impact of Chapter 1 services is identifying students who participated in the program. This is not a trivial task; keeping track of students who once participated in the program has not been a priority in many places. And, the boxes of paper containing lists and information on Chapter 1 participants have been a ready target for disposal as program staff have fought the battle of paper overload.

Discussions with school districts indicates that, at best, the Department will be able to identify , with some degree of confidence, students participating in the program only over the last five to seven years. Data on students served later than that time would be spotty and extremely suspect. Assuming that services begin in the first grade in most school systems, this means that the oldest students to be studied would be in the fifth to seventh grades when data collection begins.

Even limiting the retrospective look to the previous five to seven years is no guarantee of easy access to the needed data. Districts will vary in the extent to which an indicator of Chapter 1 performance is available on some readily accessible, computerized data base and the extent to which names exist only on a log or roster. Clearly, the former is preferable and affords a greater opportunity for linking participation information with other relevant data . In some cases, where the information is not available through either means, inspection of individual student records would be the only way of determining who had received services. In all but the smallest of districts, such a record search would probably be prohibitively expensive.

It should be noted that for this study to succeed, it is essential that there be accurate information on who the students are that have participated. When lists are maintained separately from other student information, problems can arise. The best situation is one in which both student names and identification numbers are presented for each participant. In some cases, however, only names will be provided. In small school systems, this may not be any problem. In larger systems, however, care needs to be taken to assure that names correct. This is is of special concern when trying to link separate sources of data on participants with outcome indicators or other student information. Misspellings in one place or another, use of nicknames in one place and full names in another, and actual duplications of names can all lead to incorrect matches and incorrect data. It will be important in designing and carrying out the study to build careful checks whose intent is to catch and fix as many of these problems as possible.

It is difficult to say without actually talking to district personnel who might have the required database and who might not. District size and/or sophistication do not necessarily predict accessibility. One district contacted, for example, does not have a marker for Chapter 1 participation on its central database, despite the fact that a wide range of data on students is maintained. The failure to record Chapter 1 participation is a function of philosophy--a decision was made several years ago that it might be deleterious to have the information attached to the student's record--rather than capability.

Linking data on participation to information on academic and social outcomes

To assess the impact of Chapter 1 participation, outcome indicators of success in school will be collected. Currently, indicators which will be included are academic achievement, delinquency rates, and truancy. Based on experience in conducting similar impact studies, I would recommend some additional measures such as retention in grade, placement in special education, and participation in extracurricular activities. The former two are are useful indicators not only of academic success, but also of the special services (and costs) needed to support students. Presumably Chapter 1 is successful if it reduces that extra costs needed to support a student, even if the student does not achieve as well as one might wish. The latter is a good proxy for adjustment to school life and full participation in the school program.

A critical task of the study will be linking the participation data with the outcome measures discussed above. The magnitude of this task will vary across districts, depending largely on the sophistication and breadth of computerized data base which is available. Spelled out below are the likely alternatives that the study will have to face. Their logistical and practical implications are considerable.

> First, in some systems there will exist a centralized computer base which will contain many of the items linked to student i.d.'s. In such cases, the needed data can be obtained through a simple extract. Districts with such capabilities are clearly prime candidates for the study from a practical standpoint.
>
> Second, other districts may have data on two separate databases--one maintained by the Chapter I office, the other by the regular data management operation. Extracting the needed data may require merging the two data sources, a more complicated effort than a simple extract, but clearly one that can be performed without too much trouble. assuming that the student identifiers match. Districts with information stored in this way also are appealing.
>
> Third, still other districts may have a database which is maintained by the Chapter 1 program on students while they are participating in the program, but lack post participation data in any computerized form. In such cases, the longer term outcome data will have to be gathered through record reviews.
>
> Finally, some districts may have no computerized system and all data extraction will have to take place by manual inspection of records. Further, in some cases two sets of records need to be accessed--a regular file and a confidential file. From a cost and logistical standpoint situations where this level of manual effort are needed should be minimized and, if possible, avoided.

Gaining permission to collect or access what might be considered "confidential data" (delinquency, truancy, special education participation, etc.) may pose some problems regardless of how the information is stored. And, in some cases student or parental permission will be required. This could be quite time consuming and could also lead to a biasing of the sample where permission is not received either because it is actively denied or the family cannot be reached for some reason. One way around this would be to have the districts themselves provide the information in such a way that the anonymity of indivi1uals is preserved. A critical task in the study will be determining the policies of each site regarding this matter and tailoring solutions to each particular set of circumstances.

Obtaining Chapter I Program descriptors

Studies of program effectiveness, including ones examining Chapter 1 services, have learned by bitter experience that what goes on in the name of Chapter 1 cannot be treated as a black box. Although the vast majority of schools provide support in the basic skills areas, not all do so. Some districts still provide supports that are only vaguely academic in focus.

Further, schools and school systems differ in both the amount of time devoted to Chapter 1 and the way in which services are delivered. Variations go beyond whether and inclass or pull-out approach is used. Differences also occur in whether or not the additional services are provided during the school day, before or after school, only during the regular school year or also during the summer.

Chapter 1 reports will provide a good source of information on the global approach taken to Chapter 1. Assuming that reports can be located for the years under consideration, and it is likely that they can be, such data should be readily available. However, this global picture is not sufficient. If the study is really going to try to make some statements about best practice, more detailed data at the school and even the child level are needed. For example, it is essential to know how many years of services each participant received and whetner these years were consecutive. Reasons for leaving the program also need to be documented. It is important to distinguish between the student who left because he "graduated out" and the student who left because he moved to a school where Chapter 1 was not provided.

Some districts will have detailed paper or computerized records on the services received by each participant. In many cases, however, the information will be very spotty. Existing records would have to be supplemented by interviews with program staff. The usability of this information will depend on the longevity and memory of staff, as well as the resources available for sleuthing.

This area of program description may well be one of the most recalcitrant for a retrospective, longitudinal approach. If information on "best practice" is really desired, something other than a large scale survey is likely to provide the best vehicle. I will return to this point later under additional discussion of methodological issues.

Obtaining descriptors of other "compensatory education services"

In order to understand the effects of participation in Chapter 1 programs, it is important to have detailed and accurate information on the other compensatory services that students may have received. While even with such data it may not be possible to separate out the effects of Chapter 1 participation from those of participation in other programs, the extent of confounding can at least be understood or described.

For these other services, in addition to the descriptive data on types and kinds of services described above under Chapter 1 services, some other critical data will have to be obtained. It is important to know, for example, when the "other services" were received. Were the provided before, after, or concurrent with Chapter 1 support? In addition, the criteria for receiving the additional services need to be understood. Are the same problems addressed by Chapter 1 also being addressed by these programs?

The problems in obtaining good, descriptive data on these other services are likely to be very similar to those enumerated above in discussing Chapter 1I. Detailed and accurate descriptions of services are likely to be sparse. However, it is also probable that the other services will resemble greatly those services funded through Chapter 1. In the original study of compensatory education conducted by the National Institute of Education in the middle and late 70's, it was found that by and large "other compensatory services" mirrored the program offered through Chapter 1. The only distinction was the source of funding.

Obtaining information on comparison students

In doing an impact study of this kind it is always desirable to be able to compare the effects on the treated students with those on an untreated, comparison group. This desire is not very often fulfilled, however, as finding a comparison group wh:ch is not different in very significant ways is difficult. For example, while it may be feasible to identify low achieving students in schools not eligible for Chapter 1, the fact that the school is not eligible becomes a confounding factor. Even though two groups of students in the Chapter 1 and the nonChapter 1 schools may appear to be similar at outset, their learning environments clearly differ in some potentially important ways.

When the study is a retrospective one, the problem is exacerbated. First, it may not be possible to identify the comparison students without complicated record searches. Second, some of the data needed such as supports provided to the comparison low achievers are likely to be missing and irretrievable.

It is because of these difficulties in constructing good comparison groups that most studies have measured performance against some other standard--typical growth on a norm referenced test or some expected month by month growth in the absence of treatment. In addition, participants' performance in other areas such as attendance, grades etc. has also be compared to that of the overall population in a given setting. While these strategies clearly represent a compromise, they are probably better than what could be obtained through most "comparison" groups; and, as cost is usually a consideration, such strategies definitely are to be preferred.

Conclusions and additional comments on methodology

Considering the issues discussed above, it can be concluded that while problems exist in doing a longitudinal, retrospective study, they are not insurmountable, if certain compromises are accepted.

> First, the retrospective period cannot exceed five to seven years. Going back further, will probably not be cost effective and will jeopardize the credibility of the conclusions. This means that it will be very difficult to look on any large scale at the success of students who are in high school or beyond.
>
> Second, a mixture of data collection techniques will need to be employed. Because records will be in different shape in different school districts, a variety approaches, from the most to the least technicological, will need to be used. With regard to this issue, the Department should seriously consider offering support to school systems for providing the needed data for the study. Where districts themselves have the capability of doing extracts and merges or record reviews, the use of local staff should be encouraged. In the long run, this would save costs and enhance the probability of cooperation.

Third, it is unlikely that the study will be able to include a control or comparison group of students against which to measure the progress of Chapter 1 participants. The problems in identifying and gathering the data needed on such students are significant. The study will have to employ other standards for assessing progress.

Fourth, it is unlikely that a national longitudinal study of the type envisioned by the Congress will provide adequate information on the practices or strategies which are most likely to result in success. Detailed data on Chapter 1 and other relevant program characteristics will not be available in easily obtainable form. If it is a priority to gain this type of program data, a different approach is needed.

Having concluded that a study is feasible, it is important to also consider what kind of a study will maximize the value of the information obtained. Some options are discussed below.

At first blush, it appears that what is needed is a large scale, nationally representative survey of Chapter 1 program impact. This is the strategy frequently adopted when Congress wants to know if a program is working. However, in the present circumstance, given the likely variations in data availability discussed above, carrying out such a study may not be economically possible. Some districts will not have usable data or the data may be available only through costly, and time consuming, manual record searches. In a study of this kind, attention must be given to the quality and accessibility of the retrospective data available. Trade-offs between satisfying the sampling statisticians and satisfying the budget watchers will have to be made.

In addition , a large scale, nationally representative sample of districts may also be disadvantageous from another point of view--namely, documenting successful or promising practices. To accomplish this goal, a more purposive sample is needed, one selected to maximize the chances of finding programs that work. This kind of approach would rely for sample selection more heavily on professional judgment from critical informants, than on specifications from a sampling statistician. Further, such a study would make far greater use of qualitative methodologies. And, while more restrictive in size, would provide for more indepth study of the sites included. It might even be possible to construct a quite credible picture of program impacts on older students from the detailed information obtained.

It may well be advisable to consider combining the two approaches, with more limited questions being addressed by the large scale study, and questions requiring more intensive data collection and site/program description left to the purposive study. In this case, it may take two stones to kill to birds.

Longitudinal Analysis of Student Achievement Data: Issues for Chapter 1 Evaluation

David Rogosa
Stanford University

0. INTRODUCTION

This paper is divided into three parts. The first part reviews my work on the failings of standard approaches to the analysis of longitudinal data. The second part describes some natural approaches to modelling and analysis of longitudinal data along with examples of applications to student achievement data (the analysis of student progress). The third part introduces models for the effects of interventions and considers some of the special technical issues in design and analysis that arise in the evaluation of Chapter 1 programs.

0.1 Purposes of longitudinal studies.

The most immediate question concerns the motivation for a longitudinal study, which can be divided into two types: (i) studies of growth and change (e.g. student progress) or (ii) studies of later outcomes. The first type have been the focus of my own technical research; such studies are characterized by repeated measurements (e.g. achievement measures) at multiple time points. The second type of study is longitudinal in that information is collected at various points in time, but interest is not overtly in the analysis of (individual or group) change. A good way to describe this would be "longitudinal data without longitudinal questions." An example would be collection of some background or school program information on individuals at an earlier time and linking that with some educational outcome (e.g. level of achievement) at a later time.

0.2 Research questions about growth and change.

All longitudinal studies do not have the same purposes; different types of longitudinal research questions arise throughout educational and behavioral sciences research. Some common flavors of longitudinal research questions are described below. One or more of these research questions may be addressed in the context of a particular research effort.

1. Individual and Group Growth. A basic type of question in longitudinal research concerns description of the form and amount of change. Such questions may be posed for an individual case or for the average of a group or subgroup of cases. Interest centers on the estimation of the individual (or group) growth curve, the heterogeneity (individual differences) in the individual growth curves, and the statistical and psychometric properties of these estimates.

2. Correlates and Predictors of Change. Questions about systematic individual differences in growtn are a natural sequel to the description of individual growth. A typical research question is given by "What kind of persons learn (grow) fastest?". The key quantities are the associations between parameters of the individual growth curves and the correlate(s) of change, which may be an exogenous individual characteristic (e.g. gender, IQ) or the initial status on the attribute measured over time.

3. Stability over Time. Questions about consistency over time are a natural complement to questions about change. In behavioral sciences many different research questions fall under the heading of "stability." Two key topics are the assessment of consistency over time of an individual and of consistency of individual differences over time.

4. Comparing Experimental Groups. The comparison of change across experimental groups is a standard, well-developed area of statistical methodology employing some form of repeated measures analysis of variance. When the effects of each

treatment (e.g. educational program) can be assumed identical for all members within each group (no individual differences in response to treatment), statistical comparison of the parameters of the group growth curves yields inferences about the "treatment effects."

5. Comparing Nonexperimental Groups. The comparison of of change among nonexperimental or nonequivalent groups has been a central topic in the methodology for the evaluation of social programs. The practical or political difficulties of random assignment of individuals to reatment are sometimes overwhelming in a field trial of a program. Yet the question of the relative efficacies of each program/treatment remains. However, the commonly employed statistical adjustment methods for pre-post data, often based on analysis of covariance, are inadequate.

6. Analysis of Reciprocal Effects. Questions about reciprocal effects are common and complex. Despite the complexity of these questions, empirical research has attempted to answer the oversimplified question, Does X cause Y or does Y cause X? from meager longitudinal data by casually comparing a couple of correlations (or structural regression coefficients). Hopefully, the simplistic cross-lagged correlation approaches have by now been fully discredited. Clearly, considerable empirical research on simpler longitudinal questions should precede attempts to assess reciprocal effects.

7. Growth in Multiple Measures. All questions about growth in a single attribute have natural extensions to multiple attributes. Natural questions include relative strengths and weaknessses in individual and group growth and associations of rates of growth across multiple attributes.

1. FAILURES OF TRADITIONAL ANALYSES

1.1. Myths About Longitudinal Research

Longitudinal research in the behavioral and social sciences has been dominated, for the past 50 years or more, by a collection of damaging myths and misunderstandings. These misconceptions have had large effects on the design and analysis of longitudinal research. The myths are (Rogosa, 1988):

1. Two observations a longitudinal study make.

2. The difference score is intrinsically unreliable and unfair.

3. You can determine from the correlation matrix for the longitudinal data whether or not you are measuring the same thing over time.

4. The correlation between change and initial status is
   (a) negative
   (b) zero
   (c) positive
   (d) all of the above

5. You can't avoid regression toward the mean.

6. Residual change cures what ails the difference score.

7. Analyses of covariance matrices inform about change.

8. Stability coefficients estimate
   (a) the consistency over time of an individual
   (b) the consistency over time of an average individual
   (c) the consistency over time of individual differences
   (d) none of the above
   (e) some of the above

9. Casual analyses support causal inferences about reciprocal effects.

The myths indicate some of the beliefs that have impeded doing good longitudinal research. Belief in these myths have served either to make the analysis of change appear prohibitively difficult or to direct research in unproductive directions.

The message of the myths is that models for collections of growth curves are the proper basis for the statistical analysis of longitudinal data. Research questions about growth and development make these models a natural, if not essential, starting oint. Rather simple approaches work well with longitudinal data, and much progress can be made using straightforward descriptive analysis of individual trajectories followed by statistical estimation procedures for collections of growth curves. Although only a small number of observations often are available in empirical research, the resulting difficulties in statistical estimation arising from these limited longitudinal designs should not alter the research questions or the proper statistical models.

The myths speak against what I call the "Avoid Change At Any Cost Academy of Longitudinal Research" which recommends analyses that try to draw complex conclusions about change over time without any examination of individual growth. That doctrine appears counter-productive, as these myths and my technical papers demonstrate. The doctrine of this Academy is sometimes justified by over-interpretations of the oft-quoted last sentence of Cronbach and Furby (1970): "Investigators who ask questions regarding gain [difference] scores would ordinarily be better advised to frame their questions another way." This statement could be regarded as a meta-myth. The factual basis for their conclusion is the shortcomings of the estimate of the amount of change from only two observations. But such facts do not support abandoning the framing of research questions about growth and change in a natural way. The suggested surrender to uninformative regression and residual change analyses is to be much lamented; the proper lesson to draw from difficulties with the difference score is that richer longitudinal designs and the application of appropriate statistical models for the longitudinal data are needed.

1.2 Causal Models and Longitudinal Data Analysis

My main message (also stated in Myth 7 above) is that the between-wave covariance matrix provides little information about change or growth. Thus, regardless of the sophistication of the modeling of the relations between manifest or latent variables, the causal model analysis is fatally flawed.

Path Regressions. Path analysis models for longitudinal data use the temporal ordering of the measurements to delimit the possible paths between the variables. Consider the example of a three-wave design with measures on X at times t_1, t_2, t_3. The path regressions for the unstandardized variables are:

$$X_2 = \alpha_2 + \beta_1 X_1 + e_2$$

$$X_3 = \alpha_3 + \beta_2 X_2 + \beta_3 X_1 + e_3$$

Thus the path analysis model includes direct paths from X_1 to X_2 and to X_3 (parameters β_1 and β_3, respectively) and from X_2 to X_3 (parameter β_2). The path coefficients are functions of the entries of the between-wave covariance matrix. An example of the use of this model is Goldstein (1979) in which X is a reading test score obtained on a nationwide British sample with measurements of ages 7, 11, and 16. This simple 3-wave path model was also discussed in a number of the early expositions of path analysis in the social sciences.

The properties of the path coefficients illustrate the perils of summarizing the longitudinal data by the analysis of the between-wave covariance matrix of the X_i or even the $\xi(t_i)$, thereby ignoring the analysis of individual growth. To cake the simplest situation let the true scores $\xi(t_i)$ ($i = 1, 2, 3$) be determined by a straight-line growth curve for each individual and assume perfect measurement of the X_i. For this specification the population partial regression (path) coefficients are:

$$\beta_3 = \frac{t_2 - t_3}{t_2 - t_1} < 0 \qquad \beta_2 = \frac{t_3 - t_1}{t_2 - t_1} > 0 \qquad .$$

Remarkably, the parameters depend only on the times at which the observations were taken; thus neither path regression coefficient contains any information about growth! One might think that because X_3 is perfectly predicted from X_1 and X_2 the analysis of relations among variables would be informative. Yet, under this simple structure estimates of either parameter are totally independent of the information in the data.

Latent-variable (LI^{\prime}REL) Regression Models. Latent variable regression models are a more sophisticated, but equally flawed approach to the analysis of longitudinal data. These structural equation models incorporate regression relations among latent variables (i.e., $\xi(t_i)$) with measurement models relating the observed indicators (X_i) to the latent variables. Estimation of these models is based on fitting the covariance structure implied by the structural equation model to the between-wave covariance matrix of the observations. Consider the simple structural regression model with one latent variaole ξ observed at times t_1 and t_2 and a latent exogeneous measure, W. Each latent variable has two indicators. This model is equivalent to the model for change in alienation that appears frequently as an example in Joreskog's papers. In Joreskog's examples ξ is alienation and W is socioeconomic status. The path from W to ξ_2 represents the exogenous influence on change. The structural parameter for that path is the regression coefficient for the latent variable at time 2 on the exogenous variable, with the latent variable at time 1 partialled out, $\beta_{\xi(t_2)W\bullet\xi(t_1)}$.

In terms of a simple straight-line growth model with individual rate of change θ_p , the parameters of interest for the relationship between the exogenous variable and change are the correlation between true rate of change and the exogenous variable, $\rho_{\theta W}$, or the analogous regression parameter $\beta_{\theta W}$. What does the regression parameter $\beta_{\xi(t_2)W\bullet\xi(t_1)}$ reveal about exogenous influences on growth? Not very much. For the simple case of a collection of straightline growth curves, this structural parameter has a complicated functional form which depends strongly upon the time chosen for the initial measurement. Rogosa and Willett (1985a, Section 3.2.2) gives mathematical results for the form of the structural regression parameter. For a specified relation between the exogenous variable and the individual growth parameter θ, the structural parameter may be pos.tive, negative, or zero depending upon the choice of time of initial status. Also, the structural parameter increases with the length of the interval between measurements. Numerical examples of the bizarre properties of the regression parameter are given in Rogosa (1988).

Simplex models . A third example of longitudinal analyses based on the between-wave covariance matrix is the simplex model, which specifies a first-order autoregressive process for true-scores. The numerical example of Rogosa and Willett (1985b) cautions against the propensity to base many analyses of longitudinal data on a simplex structure without careful consideration of the longitudinal data or of alternative growth models. Expositions of covariance structure analyses have encouraged such thinking; for example, Joreskog states "For one measure administered repeatedly to the same group of people, an appropriate model is a simplex model (Joreskog, 1979).

Rogosa and Willet (1985b) present an example of a 5 x 5 covariance matrix for observed scores X_{ip} over five occasions of observation. To the eye, the correlation matrix corresponds extremely well to a simplex. A simplex covariance structure marvelously fits this covariance matrix although it was generated by growth curves that maximally violate the assumptions of the simplex growth model. The consequences are far from benign because even when the simplex model fits wonderfully, the results of the covariance structure analysis can badly mislead. The covariance structure analyses usually go on to compute growth statistics and reliability estimates based on the simplex model, and these growth statistics (such as the correlation between true change and true initial status) estimated from the LISREL analysis can differ markedly from the actual values. Covariance structure analyses provide very limited information about growth in the sense that covariance matrices arising from very different collections of growth curves can be indistinguishable. Therefore, analyses of covariance structures cannot support conclusions about growth. Analysis of the collection of growth curves cannot be ignored.

2. STATISTICAL MODELS AND ANALYSES OF LONGITUDINAL DATA

2.1 Framework for Statistical Analysis

Statistical model for individual growth. Psychological learning theory and biological growth research provide a variety of complex models of individual growth, such as polynomial growth curves, logistic growth curves and simplex models. The simplest model is the straight-line growth curve,

$$\xi_p(t) = \xi_p(0) + \theta_p t \quad ,$$

where $\xi_p(t)$ is the true score of person p at time $t = 1, 2, \ldots, T$ and θ_p is the constant rate of change for person p. Thus, estimates of θ_p provide a simple index for individual rate of learning. The parameter θ_p is closely related to the amount of true change.

The straight-line growth model is useful for heuristic reasons because of its simplicity, as it yields a simple index for individual rate of progress. In addition, Rogosa and Willett (1985) point out that, "in applications, straight-line growth serves as a useful approximation to actual growth processes" (p. 205). Moreover, when observations at only a few time-points are available, such as $T = 4$ in our examples, the data may only justify the estimation of a constant rate of change.

Descriptive analyses of growth rates. When describing the learning of a group of individuals, the distribution, over individuals, of empirical rates of learning is informative. The five-number summary of empirical rates is one useful way to describe both typical rates of learning and the degree of variability in rates of growth among individuals. Also the variability in θ_p, σ^2_θ, is a key quantity. Similarly, we may want to describe the variability in level of performance at each time; $\sigma^2_{\xi(t)}$ has a functional dependence on time

(See Rogosa and Willett, 1985a).

Correlation of change and initial status. Another quantity of central importance is the correlation between change, θ, and initial status, $\xi(t_I)$, where t_I indicates initial time of measurement. As discussed in Rogosa and Willett (1985a), the choice of t_I is of critical importance because $\rho_{\xi(t)\theta}$ is functionally dependent on time. Our statistical procedures provide a maximum likelihood estimate of the correlation between true rate of change and true initial status; the correlation between observed change and observed initial status is well-known to have a strong negative bias (see Rogosa et al. 1982). The correlation is used to investigate whether those with lowest initial status make the most progress (negative value) or those with the highest initial status make the most progress (positive value).

Correlation of exogenous variables with growth. More generally, there is interest in describing systematic individual differences in growth, as indicated by the quantity $\rho_{\theta W}$ where W is some exogeneous background characteristic, for example, a characteristic of the school curriculum or a demographic characteristic of the student. The question addressed is whether students with certain values of W tend to exhibit more or less growth than students with other values of W. Our statistical procedures provide maximum likelihood estimates of this correlation.

In investigating systematic individual differences in growth, it is of course important to have a model for individual differences in growth. Rogosa and Willett (1985) state "Individual differences in growth exist when different individuals have different values of θ_p. Systematic individual differences in growth exist when individual differences in a growth parameter such as θ_p can be linked with one or more W's ." (p. 205) One simple representation is

$$E(\theta\ W) = \mu_\theta + \beta_{\theta W}(W - \mu_W) .$$

Thus non-zero values of $\beta_{\theta W}$ indicate that W is a predictor of growth. Alternatively, $\rho_{\theta W}$ is a u:eful summary quantity.

The cypical procedure is to correlate the value of the background demographic or curricular variable with performance at a giver time. That is, the cross sectional correlation is computed, sometimes for every occasion in time, and from these correlations conclusions about learning are attempted. Rogosa and Willett (1985) have shown that such cross-sectional correlations cannot inform about student progress. To illustrate, consider a situation where the correlation between true rate of change and the background variable is zero. Then the correlation between the true test score, $\xi(t)$, and the demographic variable, W, $\rho_{\xi(t)W}$, at any one slice in time could be big or small. The reverse is true also. The correlation between the background variable and a test score at a specific time can be positive, zero, or negative depending upon the time chosen for the cross-sectional correlation. Obviously, no useful conclusions about learning can be drawn from the cross-sectional correlations.

Consistency of individual differences. The index γ was proposed by Foulkes and Davis (1981) as an index of tracking, and is defined as the probability that two randomly chosen growth curves do not intersect. High values of γ indicate high consistency of individual differences over time. Thus γ indicates the stability of individual differences. If a collection of individual growth curves have a high value of γ, individuals that started out relatively high maintain that advantage and individuals starting out low retain that disadvantage (regardless of the overall growth rate).

2.2 Empirical Analyses Of Student Progress

School districts regularly assess students using group-administered achievement tests. Such testing represents a large investment in money and time for the schools, for administrators, for teachers and for students. Yet, local school agencies make relatively little use of the test data which they accumulate. In particular, test results are presented in a way that describes only the current status of students; the data are presented as a static "snapshot" of achievement without any link to prior levels of performance. Even the management of test data reflect these limitations. Whether the test results be stored as hard copy or electronically, the achievement data are typically organized as separate yearly files, which may be located on separate physical devices and even in separate geographical locations.

A key to the improved use of achievement test data is to use performance on repeated tests to describe student learning. A student's score at a single point in time cannot be used to measure learning; collecting together scores from previous testings is necessary for the analysis of student progress. A student's "cumulative folder" is organized in this manner, but these are rarely stored electronically nor uniformly maintained. Although traditional analyses and reports of test data are limited to the "snapshot" of current level of achievement, questions about student academic progress arise naturally and frequently. Such questions are separate, but not completely separable, from questions about current level of achievement. The statistical analysis of achievement histories of individual students can be highly useful in describing typical and unusual student progress and in understanding effects of instructional programs.

In the computer programs developed for the analysis of student progress, we investigate individual learning, individual differences in learning, and factors that might be related to learning, such as curricular variables, demographic variables, or other background variables. Ordinary least-squares is used to estimate the growth curve model from the longitudinal data for each student. The program analyzes The estimates of slope, squared multiple correlation and other properties of the straight-line fit are displayed and summarized. The output shows student ID; the estimates of rate of learning over the four years (i.e., the least-squares slope); the squared multiple correlation for that rate of learning; and finally, the scores at each time point. Plots of empirical rates and diagnostic listings are produced. Maximum likelihood is used to estimate properties of the collection of growth curves and key quantities describing systematic individual differences in growth.

SAN FRANCISCO HIGH SCHOOL DATA

Our initial data set was a large collection of hard-copy test scores obtained from the San Francisco Unified School District as part of the Stanford and the Schools project. We received these as four separate sets of yearly test reports, consistent with the manner in which most school systems maintain such information. The first reorganization was to form individual histories for each student consisting of their progress through high school: grades 9, 10, 11, 12. For each student this included demographic information, raw scores, derived scores and other information. The Comprehensive Tests of Basic Skills (CTBS) Form S Level 4 was administered at each grade level. The data are from the cohort that were freshman matriculating in the fall of 1979. The testing times were autumn 1979, autumn 1980, autumn 1981, autumn 1982. There are three main divisions of CTBS: reading (RTRS), language (LTRS), and mathematics (MTRS).

We examine the squared multiple correlation for the individual fits to see whether straight-lines are adequate descriptions of the four data points. We found, for example, that the median squared multiple correlation for the RTRS is .85, and generally the squared multiple correlations are very high.

The index of tracking $\hat{\gamma}$ is an index of consistency of individual differences. The estimate $\hat{\gamma}$ of .826 for RTRS indicates high consistency of individual differences. This value of $\hat{\gamma}$ is typical for most of the tests we analyzed. The stability of individual differences for each total test within gender subgroups is about .8. That is, four out of five pairs of growth curves don't intersect. The standard errors of these estimates show that we can estimate γ quite accurately for 200 people.

The estimated reliability of the rate for RTRS is .595. Thus individuals can be differentiated on the basis of their rates of change. The estimated reliability of $\hat{\theta}$ for various tests in the total group and gender subgroups range between .42 and .67. These values are certainly not consistent with the common "folklore" about the (UN)reliability of change measures.

The correlation between true change, θ, and true initial status, $\xi(t_I)$. with $t_I = 9$ for Reading Total has estimate -.15. For males and females, the estimates are -.21 and -.09, respectively.

Finally, one of the important issues we face when investigating growth is what kinds of people are growing fastest. Are they people in certain kinds of curricula? people from certain neighborhoods? males versus females? high SES or low SES people? The only background variable we had available was gender. The estimated correlation between true rate of change and gender was .04 for RTRS; for LTRS, the correlation is a little bigger, .134, which is reasonably large in terms of point-biserial correlations. In the five number summaries of $\hat{\theta}_p$ that there was a gap between males and females that was widening for language.

SAN DIEGO EVALUATION DATA

In this project we apply the methods for assessing student progress to achievement test data for minority (primarily Hispanic) students in the San Diego district and to investigate the effects on student progress of district programs for minority students. The two programs studied are briefly described below.

The San Diego district has extensive electronic data files on each student. The data central to this project resides in two separate locations: first a cumulative test file, containing the results of the achievement testing program, and second a demographic file containing essential information on individual background characteristics and school program and curricular experiences. The combination of these two files was far more difficult than anticipated. We constructed two main data analysis files, the first for 570 students over grades 5 through 9 and the second file is for 305 elementary school students over grades 1 through 4.

Equity in Student Placement. In March 1985, after an examination of tracking and placement practices, the San Diego Board of Education passed a policy for equity in student placement. In 1985-6 a five year longitudinal study was begun, with the generic research question: How are minority students doing in relation to majority students? Instead of the usual cross-sectional comparisisons between groups we focus on rates of progress and on variables that may be linked with rates of progress. Specific questions of interest include: Are there differential rates of progress by students in various tracks? What are the rates of progress by students in remedial courses?

Voluntary Ethnic Enrollment Program (VEEP). VEEP is one of the major integration programs in the San Diego City Schools. The specific objective of the program is to improve the racial balance at both the sending and receiving schools. Although student achievement is not a stated objective of the program, it is an implicit goal. Test results reported over the last two years revealed that students who participate in VEEP have lower scores on CTBS achievement tests than their ethnic counterparts at court-identified minority isloated schhols. The major question for this project is to assess student progress for VEEP students and the compare that progress with "comparison groups" at both sending and receiving schools. The results of the VEEP evaluation are reported in San Diego City Schools Evaluation Department Report #492, April 1988.

3. LONGITUDINAL MODELS FOR TREATMENT EFFECTS

The methods discussed in Part 2 are appropriate for passive observation studies and the analysis of natural maturation. However, in Chapter 1 and other educational intervention programs an additional component of any statistical model must represent the effect of the educational interventior. This part of the paper sketches some general approaches to modeling the effects of an intervention. In the last subsection additional aspects of Chapter 1 are incorporated into the formulation.

3.1 Models Combining Maturation and Response to Intervention

The statistical models and analyses for treatment-control group comparative desigrs are based on the use of (a) models for individual growth in the outcome variable and (b) models for individual differences in response to an intervention. Typically, two experimental groups are formed, measurements on individual characteristic(s) are obtained, the individuals in each group are exposed to an intervention (e.g., one of two different types of instruction), and subsequently, measurements on the outcome variable(s) are obtained. For convenience, the two experimental groups will be called the treatment and control groups. Three different specifications for the formation of these two groups are common: (i) The treatment and control groups are formed by random assignment of individuals to groups; (ii) the assignment of individuals to the treatment and control groups is by non-random (often unknown) mechanisms (selection processes); (iii) individuals are members of intact units (e.g., classrooms or schools) which are assigned to treatment and control groups.

The "natural maturation" of individuals in each of the two groups is represented by the use of models for individual growth. That is, in the absence of any intervention, a functional representation of change over time in an outcome measure Y is specified for each individual. In addition, simple models for individual response to the intervention are formulated. These models incorporate both a "main effect" between the treatment and control groups and an "interaction effect" representing systematic individual differences in response to the intervention.

Models for Individual Growth The first component of this approach is a model for the growth (change) of an individual's level or score on the outcome measure Y . (The outcome Y may be, for example, a score on an achievement test.) Individual differences in growth are represented by differences in the values of the parameters of the individual growth curves. Two functional forms for individual growth are considered: straight-line growth and asymptotic exponential growth.

Asymptotic Exponential Growth. An alternative model to straight-line grwoth curves, which may be a more realistic representation of individual growth, specifies that rate of change depends on the distance to the asymptote:

$$dY(t)/dt = \gamma_p(\lambda_p - Y(t)) \quad .$$

In this model, the parameter λ_p represents the ceiling or asymptote on Y for individual p . Thus, if $Y_p(t)$ represents the level of academic achievement for individual p at time t , then $\lambda_p - Y_p(t)$ represents the amount yet to be learned before the asymptote is reached. Here, γ_p (the learning rate constant) is specified to be identical for all p. The individual growth curve corresponding to this restriction is:

$$Y_p(t) = \lambda_p - (\lambda_p - Y_p(0))e^{-\gamma t} \quad .$$

Individual differences in growth result from differences in λ_p and in $Y_p(0)$.

Models for Response to Intervention

In conjunction with a representation for natural maturation, simple models for the (differential) effects of the intervention in the treatment and control groups complete the formulation. For each individual, the effect of the intervention is represented by an increment δ_p , which may depend upon both the group membership (treatment or control) of individual p and on individual characteristics. In this presentation, such individual characteristics are summarized by the value (assumed to be unchanging over time) of the variable A . For convenience, A_p may be termed the "aptitude" of individual p . A representation for δ_p which includes a "main effect" of group membership, and also allows for individual differences in response to the intervention is:

$$\delta_p = \begin{cases} \eta_1 + \kappa_1 A_p & \text{for } G_p = 1 \\ \eta_0 + \kappa_0 A_p & \text{for } G_p = 0 \end{cases}$$

Or, equivalently

$$\delta_p = [\eta_0 + \kappa_0 A_p] + [(\eta_1 - \eta_0) + (\kappa_1 - \kappa_0)A_p]G_p \quad .$$

The representation for δ_p specifies that all individual differences in response to the intervention are governed by values of A_p . For individual p (or any individual having the same value of A_p) the differential in incrementation between membership in the treatment vs. the control group is:

$$\eta_1 - \eta_0 + (\kappa_1 - \kappa_0)A_p \quad .$$

Consequently, if $\kappa_1 = \kappa_0$, the "treatment effect" (differential incrementation between the two alternative interventions) does not depend on individual characteristics (and is the same for all individuals). (Note that only $\kappa_1 = \kappa_0$, not $\kappa_1 = \kappa_0 = 0$, is required.) The condition $\kappa_1 \neq \kappa_0$ is consistent with the common interpretation of the term "interaction," where the interaction is between group membership and aptitude.

The average differential (or differential for a person of "average" aptitude) between membership in the treatment vs. control groups is:

$$\eta_1 - \eta_0 + (\kappa_1 - \kappa_0)\mu_A = \beta_{\delta G} \quad .$$

If $\kappa_1 \neq \kappa_0$ this average differential will change for populations or subpopulations having different μ_A . The increment δ is specified to be in the same metric as Y.

To complete this representation of the effect of the intervention, it is necessary to specify the quantity that δ_p increments. That is, given a δ_p , there are alternative answers to the question, What is the effect of the intervention? In this presentation, two types of incrementation will be considered: (a) increment directly to the status on the outcome measure (Y), and (b) increment to parameters of the natural maturation model. Specifically, the following combinations of the individual growth models and incrementation resulting from the intervention are of interest:

(i) Straight-line growth with increment to status

(ii) Asymptot˙: exponential growth with increment to status

(iii) Asymptotic exponential growth with in‹ ⁻ement to "learning potential" parameter, λ_p .

The chronology of a study provides a convenient scaffolding for the use of the models fo maturation and for response to intervention. The time of selection T^S marᵇs the division of the full sample into the two groups, treatment and control. The selection may be by random assignment of individuals, or by some systematic (or even haphazard) assignment of individuals, or even by assignment of intact classrooms to the two groups. The time t_1 marks the time of initial measurement (often termed the pretest) on ᴗne outcome Y or other individual characteristic X . The time T^I marks the time at which the intervention for both treatment and control groups (e.g., two different curricula or types of instruction) is initiated. And the time T^E marks the time at which the intervention ends. Thus, the effects of the intervention occur between T^I and T^E . Finally, t_2 marks the time of measurement of the outcome following the end of the intervention (often termed post-test).

The effects of the intervention are usually treated as occurring in a "black box," there being no attempt to model, or to otherwise investigate, exactly how or when the intervention affects the individual. For example, researchers have not explicity considered questions such as, Are the effects of the intervention instantaneous (at some time between T^I and T^E)? Or more plausibly, Are the effects gradual (spread out between times T^I and T^E or perhaps beyond)? And, if so, do these effects accrue uniformly, or at rates that vary over the interval?

Straight-line Growth with Increment to Status. For measurements at times t_1 and t_2 , the combination of the growth model and the effect of intervention can be written:

$$Y_{2p} = Y_p(t_1) + \vartheta_p(t_2 - t_1) + \delta_p \quad .$$

and the difference between the mean outcomes in the treatment and control groups is:

$$\beta_{Y_2G} = \eta_1 - \eta_0 + (\kappa_1 - \kappa_0)\mu_A$$

which equals $\beta_{\delta G}$. Incrementing the rate of change θ_p is equivalent to incrementing status and will not be considered separately.

Exponential Growth with Increment to Status. A representation of exponential growth towards an asymptote with an increment to status occurring in the interval $[T^I , T^E]$, but with no effect of the intervention on the asymptote, λ_p . For each individual status is incremented instantaneously at a time between T^I and T^E , and τ indicates the time interval between the instantaneous incrementation and t_2 , the time of post-test. For measurements at times t_1

and t_2 :

$$Y_{2p} = \lambda_p - [\lambda_p - Y_p(t_1)]\exp[-\gamma(t_2 - t_1)] + \delta_p\exp(-\gamma\tau)$$

(A gradual increment to stacus throughout the interval $[T^I , T^E]$ would yield just slightly different results.) The difference between the mean outcomes in the treatment and control groups is:

$$\beta_{Y_2G} = [\eta_1 - \eta_0 + (\kappa_1 - \kappa_0)\mu_A]\exp(-\gamma\tau) \quad ,$$

which is always less than $\beta_{\delta G}$. A consequence of the increment to status without any change in the asymptote is that the difference between the group means on Y_2 decreases as τ increases (t_2 more distant from the occurrence of the incrementation to status).

Exponential Growth with Increment to Asymptote. An alternative representation for the effect of the intervention is to increment λ_p , the asymptote for individual p. For measurements at times t_1 and t_2 , the combination of individual maturation and the incrementation to λ_p yields:

$$Y_{2p} = \lambda_p - (\lambda_p - Y_p(t_1)]\exp(-\gamma(t_2 - t_1)] + \delta_p[1 - \exp(-\gamma\tau)]$$

and the difference between the mean outcomes in the treatment and control groups is:

$$\beta_{Y_2G} = [\eta_1 - \eta_0 + (\kappa_1 - \kappa_0)\mu_A][1 - \exp(-\gamma\tau)] \quad ,$$

which is always less than $\beta_{\delta G}$. The difference between the group means on Y_2 is larger as t_2 is more distant from the occurrence of the incrementation.

Analyses for Non-Random Assignment cf Individuals

When individuals are assigned to the treatment and control groups by a non-random mechanism (haphazard or systematic), the assessment of individual differences in response to the intervention is far more difficult than with random assignment. Even the assessment of mean differences (main effects) using adjustment procedures such as analysis of covariance is nearly impossible without the use of precise information on the mechanism by which individuals are assigned to groups.

At T^S (time of selection), individuals are assigned to the treatment and control groups. If assignment is non-random, the two groups can no longer be consic .ed equivalent. In fact, the groups may differ on many attributes. The superscripts (1) and (0) are used to denote within-group moments for treatment and control groups, respectively. In particular, the mean aptitude may differ in the two groups; $\mu_A^{(1)}$ and $\mu_A^{(0)}$ denote the mean aptitudes in the treatment and control groups, respectively. Consequently, for non-random assignment, the average differential between the increments in the two groups is:

$$\beta_{\delta G} = \mu_\delta^{(1)} - \mu_\delta^{(0)} = (\eta_1 - \eta_0) + \kappa_1\mu_A^{(1)} - \kappa_0\mu_A^{(0)} \qquad (44)$$

For straight-line growth with increment to status, the difference between the group means on Y_2 is:

$$\beta_{Y_2G} = \beta_{Y_1G} + \beta_{\delta G} + (t_2 - t_1)\beta_{\theta G}$$

Thus the difference between the group means on Y_2 depends on three terms. For random assignment of individuals both β_{Y_1G} and $\beta_{\delta G}$ are zero because of the equivalence (on the average) between treatment and control groups prior tc the intervention.

Of special import is consideration of "pre-test equivalence." A common strategy in empirical research is to obtain reassurance about the validity of an analysis comparing two nonequivalent groups from a finding that the means on some initial characteristic(s) do not differ (significantly) across the two groups. However, this pre-test equivalence is not an adequate justification for

for the comparison of group means on Y_2 . Even if pretest equivalence holds, a nonzero value of $\beta_{\theta G}$ will cause the difference of outcome means to be affected by the nonequivalence of the groups.

3.2 REGRESSION DISCONTINUITY DESIGNS AND CHAPTER 1 EVALUATION

Even for the simple two-group comparative study with random assignment to groups, well-defined intervention (treatment), explicit representation of the effect of the intervention and specified outcome measures, modelling and analysis are non-trivial. (In fact, many aspects exceed current understanding and methods.) But the Chapter 1 evaluation (described by Sec 1462) introduces a slew of additional features that require acommodation in a technical formulation to guide design and analysis. The most I attempt here is to identify some of the important issues and speculate about their consequences for design. A Technical Appendix (in preparation) provides an initial look at the full formulation and its properties.

1. Regression discontinuity designs within schools or districts.

Students are not selected at random for participation in Chapter 1 programs. I do not know enough about the details of federal and local guidlines for eligability, but let us assume that the selection mechanism is knowable up to a random component, also presuming that any selection rule is not uniform across the nation. Then within a unit for which the selection rule is uniform (e.g. school or district) a comparison of Chapter 1 vs non-Chapter 1 students involves nonequivalent groups with the important structure of a known, systematic assignment rule to Chapter 1 or non-Chapter 1. Such nonequivalent groups designs are often described by the term "regression discontinuity" introduced by Don Campbell in the 1950's. More recently, statistical aspects of systematic assignment to nonequivalent groups have been analyzed by Don Rubin (work of Jim Heckman is also relevant). Rubin has shown that conditioning on the selection rule (e.g. regression adjustments) allows valid inferences about the effects of the intervention. Many aspects of regression discontinuity designs are reviewed in a 1984 book by William Trochim at Cornell. Complex probabilistic and multivariate selection rules can be incorporated into the basic regresiion discontinuity framework.

The formulation for a Chapter 1 evaluation could be described as "regression discontinuity when subjects are growing." The objective is to ascertain the effects of Chapter 1 over and above natural maturation by comparing the (nonequivalent) Chapter 1 and non-Chapter 1 groups of students. What is needed is to add the formulation of the selection process to the models for effect' of interventions and methods for the analysis of student progress, which are described in the previous parts of this paper. Such representations would be site-specific in that different selection rules may apply for different schools or districts.

2. Proximal and Distant Outcomes

Another important feature of the requirements in Sec 1462 is the inclusion of long-term (distant) outcomes in assessing the effects of Chapter 1. Proximal effects (i.e. improved school achievement, value-added to natural maturation) will be difficult enough to estimate; assessment of more distant impacts of the program (graduation, employment, earnings) is far more formidable. An intermediate strategy would be to assess proximal effects on attributes (e.g. motivation, attendence) that have obvious impacts on longer term effects.

3. What is the treatment (intervention) and How Much?

An important characteristic of Chapter 1 programs is that there is no such single thing. The amount of exposure to Chapter 1 is complex. At the very least the "treatment variable" is continuous in that different students are exposed to different amounts of Chapter 1 programs. Also, the nature of the programs vary from substitution for class content to external or extra-curricular programs. Clearly, variations in the amount and type of intervention each child receives cannot be ignored or "averaged out"; exposure to treatment must be explicit.

4. What might an effect of Chapter 1 look like?

Rudimentary formulations of effects of interventions are presented in the first section of Part 3. Those forms for δ_p and type of incrementation should be thought of as simple, abstract examples (i.e. "toy effects"). The realities of Chapter 1 should indicate more complex representaions for effects of the program (both proximal and distant). Successful detection of the effects of Chapter 1 requires at a minimum some explicit definition of what is being sought. Serious conceptual effort must precede the formulation of statistical models for the effect of the intervention.

5. Concluding Comments

Decentralize Design. Besides a common-sense revulsion for a single number, national estimate of the effect of Chapter 1, two specific features of the evaluation indicate a decentralized design. First is that selection rules for placing students in Chapter 1 vary, and second, the interventions (Chapter 1 programs) vary across schools and/or districts. Thus each site (e.g. district) included should be regarded as separate, with an accumulation of results across sites (formally or informally) providing the external validity. The imperitive is to do a solid job of assessing the efects of Chapter 1 at the individual site. Moreover, not all sites need employ the same methodology or designs.

Group Comparisons vs Understanding Chapter 1. Much can be learned from data on Chapter 1 participants that does not involve "as if by experiment" inferences as to the outcomes for the student if he/she had not been in Chapter 1 programs.

The natural history of academic progress for Chapter 1 students provides much information on many important questions. For example: What kinds of students appear to prosper in Chapter 1? In what type of program? What level of participation?

References

Cook, T.D., & Campbell, D.T. (1979). Quasi-experimentation: Design and analysis for field settings. Boston: Houghton Mifflin.

Goldstein, H. (1979). The design and analysis of longitudinal studies. London: Academic Press.

K.G. Joreskog & D. Sorbom (Eds.),(1979). Advances in factor analysis and structural equation models. Cambridge, MA: Abt Books.

Rogosa, D. R. (1980). A critique of cross-lagged correlation. Psychological Bulletin, 88, 245-258.

Rogosa, D.R. (1985). Analysis of reciprocal effects. In T. Husen and N. Postlethwaite (Eds.), International Encyclopedia of Education. London: Pergamon Press.

Rogosa, D. R. (1988). Myths about longitudinal research. In K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. M. Rawlings (Eds.) Methodological issues in aging research. New York: Springer Publishing Co.

Rogosa, D.R., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. Psychological Bulletin, 90, 726-748.

Rogosa, D.R., Floden, R.E., & Willett, J.B. (1984). Assessing the stability of teacher behavior. Journal of Educational Psychology, 76, 1000-1027.

Rogosa, D. R., & Ghandour, G. A. Statistical models for behavioral observations. To appear in Journal of Educational Statistics.

Rogosa, D.R., & Willett, J.B. (1983) Demonstrating the reliability of the difference score in the measurement of change. Journal of Educational Measurement, 1983, 20, 335-343.

Rogosa, D.R., & Willett, J.B. (1985a). Understanding correlates of change by modeling individual differences in growth. Psychometrika, 50, 203-228.

Rogosa, D.R., & Willett, J.B. (1985b). Satisfying a simplex structure is simpler than it should be. Journal of Educational Statistics, 10.

The National Longitudinal Study of Chapter 1: Design Considerations in Promoting Study Usefulness to Practitioners

Robert E. Slavin
Center for Research on Elementary and Middle Schools
The Johns Hopkins University

August, 1988

This paper considers two issues central to the design of the National Longitudinal Study of Chapter 1. One is the design of the main study, and the second is a proposal for a parallel demonstration study of alternatives in Chapter 1 design.

The Longitudinal Study

Fulfilling the congressional mandate and contributing something worthwhile will be difficult as things stand now. A retrospective study beginning with 18-year-olds and looking backwards to find out what Chapter 1 services they received may be impossible. First, most Chapter 1 services are provided in elementary school, so the study would have to find records going back 10-12 years. Finding accurate records of Chapter 1 participation and achievement that old may be difficult; in the 1970's, few districts had completely centralized and computerized their records and given students consistent ID's. Student names might be found, but confusion arising from identical names would be enormous. Even if Chapter 1 participation data and achievement data could be located, it would be extremely unlikely to find any more detail on the programs students received than whether or not they were program participants. The results would thus be of limited value to practitioners. Finally, any results obtained would reflect effects of Chapter 1 services as administered in the 1970's; they may already be obsolete.

An ideal longitudinal study would start with kindergarteners and follow them for many years, beginning with a large number of students and selecting sites representing a cross-section of Chapter 1 participants (and matched non-participants) and a full range of Chapter 1 delivery options (e.g., pullout, in-class, add-on, replacement, schoolwide). Such a study could substantially improve upon the now very dated Sustaining Effects Study. However, it would obviously not satisfy the intent of Congress to study effects of Chapter 1 on 18-25 year-olds.

I would propose a design that could satisfy Congress and still provide data that would be of more direct use to practitioners. The basic idea would be to follow 5-12 year-olds for six years. The 12 year-olds would satisfy the congressional mandate (to study effects by age 18), yet data on 12 year olds' Chapter 1 participation and early achievement should still be available within their school districts (and are likely to be centralized and computerized). Studying 5-11 year-olds in the same schools would allow for a better longitudinal study to take place under the overall rubric of the congressionally mandated study. Actually, studying 5-year-olds, 8-year-olds, and 12-year-olds would probably be enough (instead of studying every grade), but given retention policies now in effect in most districts, students at any given age are likely to be spread across many grades, so if all grades must be tested anyway, including all students may not add much to the study cost (and would add much to the study's contribution).

I would propose that school districts, district-university partnerships, or district-contractor partnerships be invited to compete for funds to participate in the longitudinal study. The funded districts could be asked to show that:

1. They had good data on the early achievement and Chapter 1 participation of current 12-year-olds.

2. They were willing to hold study participants receiving Chapter 1 services in the same service category as long as they qualified. That is, as long as students' achievement stayed below the 40th percentile, they would remain in the same pullout/in-class/add-on/replacement arrangement (this may require a waiver of Chapter 1 rules requiring that the lowest-achieving students be served first).

3. They had students who would qualify as a control group (e.g., the district has non-Chapter 1 schools in which there would be students who would qualify for Chapter 1 services in a Chapter 1 school).

4. Among districts which met these requirements, preference would be given to districts which had different service models within the same district (e.g., pullout, add-on, schoolwide) and to districts whose inclusion would create a better cross-section (e.g., rural districts, if most applicants were urban).

Of course, a national contractor would have to be sought to coordinate the activities of the local contractors.

In addition to using local tests, it would be important to use a common test with all participating districts. This test might use matrix sampling to assess a wide range of skills, including higher-order skills, an individual reading assessment, and a writing sample. Given the problems with districts' own standardized testing programs, use of such tests is essential. Incidently, testing younger children (in the same districts) with the common test would allow for translation of 12-year-olds' old standardized test scores into the metric of the common tests.

At the end of the process I've outlined, we'd have the following:

1. A true longitudinal assessment of the achievement effects on Chapter 1 participation per se in the elementary grades (from the 5-year-old and other young cohorts).

2. A pretty good longitudinal study of the effects of Chapter 1 participation on delinquency, dropout, early pregnancy, college entrance, and so on (from the 12-year-old data and older cohorts).

3. A very good longitudinal assessment of the achievement effects of alternative Chapter 1 service delivery models (pullout, in-class, add-on, etc.). This would be enormously useful for practice.

Along with these, several other sub-studies might be possible. One might be a process-product study to identify practices and program elements characteristic of outstanding Chapter 1 programs. Another might be one or more studies of such assessment issues as the use of standardized tests to determine program effectiveness, in particular the use of fall-to-spring and spring-to-spring NCE gains. The use of a reliable, broadly conceived common test other than the district's standardized tests would make this easy, and could answer many questions about use of tests to evaluate programs. This common test would tell us the degree to which district scores represent true achievement as opposed to teaching to the (standardized) tests.

A Demonstration Study of Alternatives in Chapter 1 Program Design

The study I've outlined above would do a pretty good job of comparing alternative Chapter 1 service models that are currently in widespread use. However, to make a significant impact on the capacity of Chapter 1 to meet the needs of at-risk students, efforts are also needed to expand the range of options, in particular to identify effective alternatives to traditional methods. The National Longitudinal Study could serve as a vehicle for promoting the development and evaluation of instructional methods which could be used under Chapter 1 funding to make a difference in the lives of children.

I would propose that a demonstration study of systematic alternatives in Chapter 1 service delivery be conducted along two parallel tracks, schoolwide and non-schoolwide. Because schoolwide models have been rare in the past, yet are sure to expand rapidly under new legislation, this is a particularly important area in which to do controlled evaluations of plausible alternatives. Examples of schoolwide alternatives which might be contrasted would include using Chapter 1 funds to reduce overall class size, any of the various continuous-progress models used in regular (not pullout) classes (e.g., Johnson City, Distar, PEGASUS, U-SAIL), Henry Levin's Accelerated Schools model, Margaret Wang's ALEM, James Comer's New Haven model, or our own Success for All and cooperative learning models.

Among non-schoolwide models worth studying might be traditional pullout and in-class models, CAI, peer tutoring, after-school programs, summer school programs, Reading Recovery or other early tutoring models (e.g., Early Prevention of School Failure, Wallach & Wallach's model, etc.), and specific curricula used within pullout models, such as High-Intensity Reading/Math and Corrective Reading.

What I have in mind is that program developers would be funded to implement their programs in several locations under stringent, pre-established conditions of experimental design. These would include use of random assignment to experimental o. control groups (or well-controlled matching).

In the case of schoolwide programs the control groups might be both schools using pullout models and schools using schoolwide funds to simply reduce class size. For non-schoolwide models, the control groups could be both pullout models and matched non-Chapter 1 students.

Districts which agreed to participate in the research would have to agree to keep their programs (includinc cheir control groups) for, say, three years without change, and (in the case of non-schoolwide studies) to keep individual children in the programs as long as they scored below the 40th percentile.

The measures of the program effectiveness could be the same as those developed for the longitudinal study. That is, they should probably use matrix ampling to get at reading, math, and language in the broadest sense, and should involve testing a subsample (at least) using individual reading measures. If possible, the tests themselves should be withheld from the developers to keep them from teaching to the test.

The funds provided to the developers would have to be adequate to allow for top-quality implementations. Site variation (i.e., failure to implement in many sites) cannot be allowed to occur, as it did in Follow Through Planned Variation.

A national contractor would have to be in charge of overseeing the developers' activities and conducting the actual evaluations. They would also prepare assessments of the true implementation costs of each program for use in dissemination.

Ideally, the Demonstration Study could provide a basis for a continuing process of federal support for development, pilot testing, developer evaluation, independent evaluation, and dissemination of alternative programs within Chapter 1. That is, there are some programs which already have adequate developer evaluations to justify independent evaluation in a demonstration study. Other programs are promising but need further developer evaluation and refinement before they ould be good candidates for independent evaluation. Still others don't even exist, or are in early stages of development. Federal monies could be provided to support movement of promising ideas through the development, pilot testing, and developer-evaluation stages so that in future years there would be a continuing supply of programs worth subjecting to independent evaluations.

Along with this process it might be a good idea to establish a national clearinghouse for good-quality program evaluations conducted by districts. A bounty might be established for districts which compared new programs to traditional control groups to send in their reports. For example, dozens of districts have done top-quality, control group evaluations of IBM's Writing to Read program, yet no one (except perhaps IBM) has a central file of such studies. It would be easy to identify districts which are experimenting with new programs under Chapter 1 funding and to solicit copies of program evaluations from these districts. The results of these evaluations could be reviewed and reported from time to time, and may be used to identify programs worth evaluating in the indepeudeut evaluation systam I've described.

October 14, 1988 M. Smith

Thoughts on the Chapter I Longitudinal Evaluation Design

This note reviews a variety of issues having to do with the logic and coherence of the Congressional mandate and the nature of a longitudinal evaluation design. The issues are dealt with independently even though they will all influence each other in the eventual design. No single design is advocated in this note -- instead I suggest that the government needs to establish a limited set of clear goals for the evaluation and then work toward the most parsimonious strategy possible to accomplish the work. The goals must be agreed to by both the executive and the legislative branches or the evaluation will run the risk of being of little practical value. I then consider a series of design issues that must be thought through no matter what goals are selected.

The first task is to figure out what issues you want to address in the final report. To do so you need to get some clarification of the language of Sec. 1461 of Public Law 100-297. Exactly what questions and what kinds of evidence does Congress want to consider? After that you need to figure out what kinds of questions you can address within the constraints of the possible budgets and time frames and what kinds of evidence will be convincing. Finally, you need to match the Congressional purposes with what is possible. The following are a set of somewhat random observations about the possible goals and design of the study.

1. Ambiguity (or clarity?) of Congressional Intent: Sec. 1481. There are lots of ambiguous terms in the section: for example, "eligible children", "significant participation", "qualified organization", "consider the correlations between participation ......" etc. But even with the ambiguity there arises out of the language a design that is coherent, though silly. As I read the language of the section I see the literal interpretation of the intent as implying the following design. (If you would like to go through my logic call but I think it is possible to follow if you look closely at the language of Sec. 1481.).

a. The first step would be the selection of a single agency or organization to conduct a survey of "eligible children participating in the program" of Chapter I. The language about "participating" must be wrong --in order to meet other parts of the intent the surveyed people would have to have been past participants (graduates) of the program. For one thing you will not know who has "significantly participated" until they do. For another, it is impossible to satisfy the "25 year olds" requirement within the Congressionally specified time limit with a design that samples participating students. The survey would be carried out in 1989. The children (graduates) would be spread out in years from 7 years old to 18 years old. The survey would also include a "comparison" group of "comparable" children who did not receive Chapter I services.

b. The survey would be stratified by region and place across the nation and be large enough to "assess and evaluate the effectiveness of the program in all regions".

c. The intent would be to find an "effect" of the program on achievement, an "effect" on delinquency, an "effect" on .... Presumably, the achievement "effect" would be the difference between the achievement of the "graduates" and the achievement of the "comparable" children who did not participate.

d. A report on this survey would be issued in 1993.

e. In the meantime there would be periodic followups to the original survey carried out by the same contractor. You have some freedom in the manner in which you followup. Let's say that there are 2 followups, one in 1992 and one in 1996. The latter followup would put the original survey's 18 year olds at age 25 meeting the intent in 1462(b). The followup surveys would also be designed to assess the effects of Chapter I.

f. A final report would be due in 1997.

2. What about the "Congressional design:" The design implied in Sec. 1462 is really cockamamie. The major problems have to do with the assumption of a single kind of "chapter I program" (indicated by the idea of a single "effect"), the retrospective identification of students as in Chapter I or not and for how long, the lack of base line data, the reliance on only one data point for the major results of the interim report, the meaning of "eligible" (which is relative to time and location), the difficulty in finding "eligible" children who did not participate, the meaning of "significant" participation, the requirement to follow study participants to age 25, the difficulty in obtaining any sort of representative sample of Chapter I students at age 18 (since many would have dropped out between ages 15 and 18), the time between the completion of the data collection and the final report, the fact that Chapter I changed significantly between 1978 and 1988 (the time periods when Chapter I students, among others. The chances of finding out

anything of importance with the apparently intended design are next to nothing.

3. You need to discover a deeper Congressional intent. My sense, also, is that you need to get the Congress to explicitly recognize it (the deeper intent). Because if you don't some wise guy is going to come along and point out the logic of the language in 1462(b) and put your procurements into jeopardy. This means that you have got to figure out what is important to know within this general area, convince the relevant Congressional staffers and either solicit a letter from the Committees clarifying their "intent" or, better yet, stimulate a colloquy on the floor that is agreed to by both houses. As part of this process of clarification you will need to obtain the flexibility to come up with the best design you can even though it could include some elements that are counter to the existing Congressional language.

4. A deeper Congressional intent. It seems to me that the underlying intent is to determine whether "significant participation" in Chapter I has the same kind of positive long term influence on behavior that participation in early childhood programs is believed to have. This stimulates a number of thoughts and suggestions:

a. Talk with someone who has made a careful study of the studies of the long term effects of preschool. As I recall there was no measurable effect of preschool on achievement or IQ. The effects were in the areas of the kinds of behaviors which are influenced by motivation and attitudes -- they seem to be particularly in areas allied with a sense of efficacy. Thus, preschool students were more likely to graduate, more likely to go to college, less likely to be in trouble with the law, less likely to be dependent on drugs, less likely to be pregnant etc.

b. The preschool longitudinal studies were extensions of earlier quite carefully conducted studies which had real "control" or at least similar "comparison" groups, pre-measures, immediate post-measures, in some instances a limited longitudinal design, and in almost every instance a close familarity between the investigator and families in the study. The design implied in the language of Sec. 1462 has none of this. An alternative design might embody some, but not most, of these elements.

c. The treatment (preschool) extended over an entire school years time and took up something on the order of 300 hours (3 hours a day for 100 days seems a conservative estimate). This ought to be checked. But it may give a clue about the meaning of "significant participation". It would be worthwhile to compare the amount of time a preschool student was in its "treatment" with the amount of time a Chapter I student was in its treatment.

d. The preschool did not substitute for another structured educational experience. The "comparison" group for these studies typically were not part of a structured educational group care experience. This condition will be practically impossible to achieve for Chapter I students. In fact, in many instances, the Chapter I educationally experience simply supplants the experience the Chapter I students would have if they stayed in their normal class. The only ways in which the condition could be reached would be through study of a sample of students whose Chapter I experiences were after school, on the weekends or during the summer.

e. Most of the preschool programs heavily involved the parents. One major hypothesis which would account for the long-term effects of the pre-school interventions (in the absence of short range major longitudinal effects on achievement) is that the parents increased their sense of efficacy and their knowledge of "the system" through their own participation in the program. This enabled them to better protect the rights of their children over the next 12 - 13 years of the life of the child as he or she worked their way through the school system. Again, this would be a difficult but not impossible condition to match in a Chapter I study since individual parents are typically nowhere near as involved in Chapter I as they are in preschool programs.

6. Implications of the Preschool Studies for the design of the Chapter I study: The above suggests that there are a large number of important differences in program design between the preschool programs which have shown long term effects and Chapter I. There are also a number of research design differences between the preschool studies and the Congressionally intended design for Chapter I. Although I would not slavishly follow the lead of the preschool experience there are a number of ideas that are worth pursuing.

a. Explore the use of measures of attitudes and motivation (particularly efficacy) as measures of interim states of students -- between the time of their "treatment" in Chapter I and their late adolescence. These measures might serve as proxies for the positive behaviors that you hope the Chapter I program will stimulate.

b. Use other interim proxies for the long-term behaviors. For example, grade retention (particularly double retention) is a terrific predictor of dropping out. Retention is also a powerful outcome in its own right. Early deviant behavior is a good predictor of later deviant behavior. Cuts, bad grades, and very low track assignments in the middle school years are good predictors of lots of negative behaviors. These proxies should be used in the aggregate sense -- there is always a lot of error in their predictive capacity and individuals should not be identified.

c. Look for previous studies of Chapter I. If you can alter the Congressional design look for earlier studies (conducted in the middle and late seventies and the early eighties which had pre-measures and comparison groups and which gathered post treatment data and perhaps even some longitudinal data. If you found some you might piggy back on them for your eventual assessment of effects at ages 18 and beyond. This suggests casting a wide net. Remember that the preschool studies were a pretty haphazard lot -- they were each designed to have internal validity -- it is the lot of them that suggests external validity.

d. Defining "significant participation". The preschool studies suggest some dimensions and there are others.

o Time 1. How long should the treatment be? A student goes to school roughly 900 hours a year for 12 years -- a total of some 11,000 to 12,000 hours. It seems absurd to imagine much long-term influence from a 40 minute a day, 150 day one year pull out program that substitutes for other instruction. Such a program amounts to roughly 100 hours -- less than 1% of a student's total time in school and only one-third of the time that the children had in the preschool programs. Is this "significant participation"? One way to extend it would be to define "significant participation" as being more than a one year program. The problem with this is that students who are successful in Chapter I leave after one year -- students in the population that participates in multiple years have a particularly difficult time with school and, therefore, may be less likely to show positive long term behaviors. This is not an easy problem.

o Time 2: The nature of the treatment.: Other things being equal, it makes sense that programs that replace other programs will have weaker marginal effects than will programs which increase educational exposure. The preschool programs are completely additive -- most Chapter I programs substitute for other, almost equivalent treatments. Think of a continuum of treatments from completely additive to fully substitutive in your definition of "significant participation". The dimension might have the following elements:

- Completely additive: Similar to Head Start. This would include Chapter I summer programs, preschool programs, after school or weekend programs.
- Partially additive: Pull out or within class programs that increase time spent in a subject matter area such as reading.
- Substitutive: Pull out or classroom subject matter (eg. reading) programs that supplant the existing program -- this is the large majority of programs.

The Whole School program is on a somewhat different dimension. I am not sure how to categorize it.

o Parental Involvement: It might be useful to include the degree of parental participation in the program in the definition of "significant participation". If you follow the preschool lead parental involvement is on the educational rather than the political side.

o Level of Resources: The level of resources committed to the Chapter I program should be taken into consideration in the definition of "significant participation".

o Best Practices: This maybe where you want to introduce the notion of "best practices". The quality of the practice as defined by theory and prior results might be one of the criteria defining "significant participation".

o Program goals: The purpose of the program should also be included -- you probably want to limit the sample to Chapter I programs which have clear academic goals.

7. Beyond a clarified intent what else do you need from Congress?

a. Multiple Contractors: You should be able to use different contractors to do different parts of the job. If you have multiple studies you will surely require multiple contractors. Even if you don't you may well want have different contractors for different parts of the work -- planning, data collection, analysis. One strategy might be to use a single contractor as a conduit for work similar to the consortium used for some of the OERI Centers.

b. Flexible ages: It will become harder and harder to track former participants as they get older. Tracking is very expensive as are the eventual interviews. Change 25 years old to 20 and put in language about "and older if possible".

c. Multiple studies: Take a page from the preschool book and obtain the flexibility to have multiple small well conducted studies rather than one big study. This would also allow you to look for pre-existing studies. Don't put all of your eggs in one basket.

d. Get rid of the state operated Chapter I programs: Make sure that Congress does not mean for you to include the N&D, the Handicapped and the Migrant programs in your study. They will add tremendous cost and complexity. If necessary arrange to do separate longitudinal studies.

8. The need for theory: Probably the most important thing that you can do in the study design period is to imagine a theory or theories that would explain how Chapter I might have a long-term effect on significant behaviors of past participants. One possibility is parental involvement. Another might be major gains in achievement which are either sustained or operate by reducing the probability of a student being retained in grade or placed in low tracks. Another possibility is that some Chapter I programs affect the motivation of students. Maybe there are other possibilities. The mix of plausible theories ought to help you define what you mean by "significant participation", what you mean by "best practices", what measures you will gather on an interim and final basis, your sample sizes in certain cells, the nature of your initial analyses and lots of other things. If you go into this study without a set of guiding plausible hypotheses there is nc hope for a successful study.

9. Multiple Studies for a single goal: I mentioned that the preschool studies were carried out by different researchers. There are great potential advantages in having a number of separate, small, internally valid Chapter I longitudinal studies carried out by different contractors operating in different parts of the country. External validity would be

generated by having a variety of studies carried out in a variety of settings. Multiple contractors might be cheaper since the need for travel would be reduced. They might also be able to take advantage of existing studies more easily than a single national contractor.

10. Explore different ways of conducting a retrospective study: Local data on the past Chapter I participation of students are weak and incomplete at best. It would be very difficult to classify students as having participated or not in Chapter I in the kind of broad based retrospective survey indicated by the language of Sec. 1461. It would be even harder to determine the nature of the Chapter I treatment, the intensity of the treatment and the kind of success that the student had in it. Since a retrospective study is necessary if you are going to have data on students at age 18 who have been in Chapter I before 5th grade you will have to think long and hard about how to get the data. Four different approaches seem plausible. All are important enough opportunities to warrant serious exploration.

a. Use a sample from the SES data: With the SES data you have a large sample of participants and non-participants, pre and post measures and even some further measures, and knowledge of the treatment. The trick would be to followup on the SES sample. By now in age they would be roughly 20-25 years old. This seems like a very best bet.

b. Use a sample from the HSB data: These students would now be 23-26 years old. We know where they are and we have lots of data on them spanning their years from age 15 to 22-26. The major problem with this sample would be to go back to the elementary years to find out about their participation or not in Chapter I and the nature of nature of the Chapter I program.

c. Use a sample from the new NLS data: These students are now in eighth grade (13-14 years old). By 1995 they will be roughly 20 years old. We will have lots of data on their late adolescent years. The central trick would be to go back and figure out whether they were in Chapter I or not and what the program was like. This also seems like a good bet to me.

d. Explore the availability of data from studies conducted by city and state evaluation agencies. There may well a number of very useful longitudinal studies out there in some SEA or LEA evaluation agencies. It would not be difficult to query the network of Chapter I evaluators to find out if there are any candidates.

11. Comparison Groups: As you work through the design remain initially flexible with respect to the unit of analysis and the nature of the comparison group. Differently formed comparison groups can serve to help tell us different things. Among the comparison groups that you should consider are groups formed from within the same school as the Chapter I students, groups formed from different similar schools within the same district (who may not have Chapter I in the same grades), groups formed from similar schools in different districts (who may not have Chapter I because of different district criteria). The comparison problem is not hopeless but it will take work.