DOCUMENT RESUME

ED 113 011	PS 008 000
AUTHOR	Smith, Marshall S.
TITLE	Some Short Term Effects of Project Head Start: A Preliminary Report on the Second Year of Planned Variation--1970-71.
INSTITUTION	Huron Inst., Cambridge, Mass.
SPONS AGENCY	Office of Child Development (DHEW), Washington, D.C.
REPORT NO	OCD-H-1926
PUB DATE	Jan 73
NOTE	277p.; For related documents, see PS 008 001-002 and ED 082 834-835
EDRS PRICE	MF-$0.76 HC-$14.59 Plus Postage
DESCRIPTORS	Achievement Gains; *Comparative Analysis; *Compensatory Education Programs; *Curriculum Evaluation; Data Analysis; Demography; *Early Childhood Education; Longitudinal Studies; *Program Evaluation; Race; Tables (Data)
IDENTIFIERS	*Project Head Start

ABSTRACT

This preliminary report evaluates the second year, 1970-71, of Head Start Planned Variation (HSPV), covering research methodology, description of the models, characteristics of the children, classrooms and sites, estimated overall effects of the Head Start experience, differences in the effects of PV and comparison classrooms, and short term effects of 11 Head Start program models. An attempt is made to answer the questions: (1) What are the short term effects of a Head Start experience on children? (2) Are there discernible differences between the effects on children of a HSPV experience and a conventional Head Start experience? (3) Do PV models differ in their effects on Head Start children? Five outcome measures were used: three measures of cognitive achievement, one of general intelligence, and one of motor control. Major findings indicated that: the Head Start experience substantially increased children's test scores on all five outcome measures; that children who had prior preschool experience gained less overall than children whose first year of preschool was in Head Start in 1970-71; and that there seemed to be no consistent differences among Mexican American, black, and white children in their Head Start gains on the five outcome measures. No differences in effects were found between the HSPV programs and the comparison Head Start programs. (GO)

SOME SHORT TERM EFFECTS OF PROJECT HEAD START: A PRELIMINARY REPORT ON THE SECOND YEAR OF PLANNED VARIATION -- 1970-71

MARSHALL S. SMITH

JANUARY, 1973

HURON INSTITUTE
CAMBRIDGE, MASSACHUSETTS

This document was prepared for Grant # H 1926 from the Office of Child Development, Department of Health, Education and Welfare, U.S. Government. The conclusions and recommendations in this report are those of the grantee and do not necessarily reflect the views of any federal agency.

The project director was Marshall S. Smith.

Project staff included:

Mary Jo Bane
Barbara Behrendt
Anthony Bryk
John Butler
Thomas Cerva
David Cohen
Jane David
Helen Featherstone
Nathan Fox
David Gordon
Deborah Gordon
Sharon Hauck
Gregory Jackson
Carol Lukas
Robert McMeekin
Anne Monaghan
David Napior
Ann Taylor
Deborah Walker
Herbert Weisberg
Jack Wiggins
Cicero Wilson
Cynthia Wohlleb
Joy Wolfe
Stanley Yutkins
Diane Zipperman

TABLE OF CONTENTS

SOME SHORT TERM EFFECTS OF PROJECT HEAD START: A PRELIMINARY REPORT ON THE SECOND YEAR OF PLANNED VARIATION -- 1970-71

I. INTRODUCTION

Background

During the early months of 1969 the Office of Child Development planned a three wave longitudinal study designed to assess the relative impacts of a variety of preschool curricula. The study was called Head Start Planned Variation (HSPV) and began in fall 1969. Plans called for the systematic assignment of ~~a~~ number of well-developed curricula, each to two or more sites throughout the country. Selected sites were to meet three criteria.

First, each site was to contain an on-going Head Start Program. No funds were allocated for serving children other than those already being served by Head Start.

Second, each site was to draw participant children from a preschool population living largely within the attendance area of a school or schools where older children attended a Follow-Through program.* By fall 1969

* Follow-Through is an intensive early elementary (K-3) compensatory program designed to enrich the experiences of economically poor children -- particularly poor children who have had Head Start experiences. Originally intended to be a national program, Follow-Through was designated as an experimental effort in 1968, one year after it was initiated. By 1969 there were over 170 school districts with Follow-Through programs.

most Follow-Through schools had adopted one of a number of well-defined educational curricula. These programs were being evaluated by the Office of Education. Children entering selected Follow-Through schools during the years 1969-1972 were to be tested at entrance and longitudinally followed and tested until they completed Follow-Through at the end of third grade.

Third, the selected Head Start site had to agree to adopt the curriculum model being used in the Follow-Through schools in its area. Aid in implementing the models was to be provided by consultants responsible to the original architects of the models. In addition, extra funds for purchasing equipment and for hiring teacher aides were to be provided to the selected Head Start classes. Overall, the cost of implementing the Planned Variations model is estimated to be $350.00 per child above the cost of conventional Head Start (see McMeekin, forthcoming, for a detailed estimate of the extra costs). Since many of the Follow-Through curricula were adopted from programs originally designed for pre-schools, the use of them in Head Start programs was appropriate.

Sixteen sites were selected for inclusion in the Planned Variations study during the 1969-70 school year. Eight curriculum models were represented, each by two sites, and formed the sample for the first wave of the study. The second wave of the study, the school year 1970-71, included thirty-seven sites. Since one of the original 16 sites had dropped out of the study, this meant that 22 new Planned Variation sites were added in 1970-71. Fourteen of the 22 new sites followed one of the original eight models and were located in a Follow-Through area. Of the remaining eight new sites three were in Follow-Through locations (one in each of three models) and five were located in sites without Follow-Through schools. The final five followed a curriculum designed by the parents and staff of the children in the site in collaboration with a consultant from the Office of Child Development. The third wave of the study (1971-72) involved the same sites as the second wave with two exceptions: two sites were dropped and one was added.

The design of the Planned Variations study called for children in all three waves to be tested at the beginning and end of their Head Start experience. Following Head Start, the children would enter the Follow-Through program in their community and be evaluated at the beginning and throughout their Follow-Through experience. The records of the Head Start and Follow-Through testings could then be linked. The linkage would provide data for a longitudinal assessment of the combined pre-school and early elementary experiences of the Planned Variation children.

Testing was also planned for other groups of Head Start children in every Planned Variation site. These children would attend Head Start classes without a designated curriculum component and serve as a local comparison group for the study of the Planned Variations Head Start classes. With some exceptions this strategy was followed for all waves of the Planned Variations study. The comparison children were also to be included in the Follow-Through evaluation.

Progress reports on Planned Variations were planned at three times during the course of the study: first at the end of the Head Start experience for each of the three waves of children; second, at various times during the Follow-Through experiences of the three waves; and third, in 1976, after the third wave of Head Start children had completed Follow-Through. A preliminary report on the first wave (1969-70) was prepared in 1971 by the Stanford Research Institute for the Office of Child Development.

The present report is one of a set of preliminary reports on the second wave of Head Start Planned Variations. Final reports, to be available in September 1973, will review the one-year data in all three waves of Planned Variations.

Purpose of the Report

This report attempts to answer three questions:

1. What are the short term effects of a Head Start

experience on children?

2. Are there discernable differences between the effects on children of a Head Start Planned Variations experience and a conventional Head Start experience?

3. Do Planned Variation models differ in their effects on Head Start children?

In all instances the measured effects we discuss here are narrowly defined. Specifically, we are concerned with three measures of cognitive achievement, one measure of intelligence and one measure of motor control. No attempt is made to introduce data about the many other areas which a preschool experience might influence.

This report has been prepared in conjunction with three other reports about Wave Two of Planned Variations. One report considers the process and success of implementing the Planned Variation curricula in the various sites. A second report presents a detailed summary of the various measuring instruments used in all three waves of the Planned Variations study. A third report explores the possibility that different characteristics of children interact with particular curricula to produce different results. This final report analyzes children in both Waves One and Two of the study.

No report in this series attempts an overall systematic review of the preschool literature. For this the interested reader should see Datta (1971), Stearns (1972) or White et al. (1972). And no report in the series attempts to provide a detailed description of the twelve Planned Variation models. For this the reader should see Maccoby and Zellner (1971), and the Rainbow series published by the Office of Child Development (1972)

Limitations of this Report

It is impossible for a report of this nature to capture the richness and complexity of a child's Head Start experience. The best we can do is to report summary estimates for a very narrow range of effects of presumably different preschool experiences. Four specific constraints on the value of this study should be noted at the outset.

1. Like almost all studies of school effects, we assume a production model of the preschool process. An analysis of this nature requires us to initially measure certain inputs of the child and his classroom which we think may be important, make assumptions about the homogeneity of children's experiences within a given classroom, and then gather some output measurements on the child at the end of his preschool experience. Then, after controlling for relevant initial differences among children, we compare groups of children on our output measures. For the most part we make no attempt to understand the diversity of experiences that children bring or have in their preschools. One reason for the

narrowness of our approach is the lack of a consistent theory of child development; another reason for it is the lack of a strategy of analysis which is sufficiently complex to deal with more than the skeleton of reality.

2. The lack of a consistent theory of child development is reflected in the sparsity and limitations of the measures used in this study. As we noted earlier we will report only on four measures of the cognitive area and one measure of motor control.* Though these measures are among the best to be found, they still have only questionable validity. (See Chapter 2 and the report "The Quality of the Data.")

3. In order to justify comparisons among curricula we have to make assumptions about the integrity of the various curricula in different sites. The initial assumptions are: first, that the various preschool curricula do create discernably different classroom environments; and second, that the curricula are exportable -- that is, that they can be implemented in various classrooms around the country. These

* Two other child output measures were included in the 1970-71 data collection. The reasons for their exclusion here are outlined in Chapter 2.

assumptions are discussed in the report on "Implementation in Planned Variations -- 1970-71". As that report points out, they are neither trivial nor always valid. By and large, however, we plunge ahead and accept them as valid, for given these assumptions we might be able to attribute differences in child outputs to the influence of the different models.

4. Problems also stem from the study design. First, as described above, Planned Variation sites were not randomly assigned to models. Rather, the sites were selected on two criteria unrelated to the requirements for an adequate experimental design, and then given the opportunity to accept or reject the assigned curriculur. Moreover, the local community had control over the specification of which classes within a site were to employ the Planned Variations (PV) curriculum. Since the selection of comparison classes within the PV sites occurred after selection of the Planned Variation classes, the treatment (PV) and comparison (NPV) classes cannot be assumed to be random samples drawn from the same population. Thus, randomization did not occur at either of the two critical design points -- at the level of assignment of curricula to sites or at the level of assignment of

treatment and comparison groups within sites.

Without randomization we must rely on statistical techniques to control for the influences of factors associated with the selection processes. Since we do not have a clear understanding of the motivations and mechanisms which guided the various among-site and within-site assignment procedures, our approach must be to control as much as possible the variables which might be relevant. If our developmental theory were adequate and the number of replications of each curriculum large enough, non-randomization might not be a serious problem. But our theory is not adequate to fully specify all possible important and uncontrolled influences. And even if our theory were adequate, the number of replications of each curriculum in the study is probably too small to accommodate all of the necessary controls. The relatively few replications for each curriculum leads to instances where treatments (curricula) are both partially and fully confounded with potentially important and measured control variables. The lack of a fully developed theory and the confounding of control variables with FT curricula forces us to deal with analytic models which have unknown specification biases.

Strengths of the Study

Granting the above, what particular strengths does this study bring to the analysis of the effects of preschools on economically poor children? To answer this we have only to look at previous research in the area. Three characteristics of the Planned Variations study stand out.

1. There is an attempt in Planned Variations to systematically vary the preschool environments of children in a number of locations around the country. Prior to this, national studies of preschool have looked only at naturally occurring differences among classroom environments. (See Westinghouse-Ohio, 1968 or "The Study of Natural Variations in Head Start," 1969.) There have been studies of systematically varied preschool environments in single locations (see Bissell, 1970 for a summary) but never before has there been a national study of this sort.
2. We have great confidence in the care and accuracy of the data gathered in this study. While many studies have gathered pre- and post-test data, and information on children and teachers characteristics, no data collection effort for a national study has been as carefully administered and conducted. For a review of the data collection procedures see "The Quality of the Planned Variations Data."
3. The Planned Variations study has multiple replications.

Among waves there is replication of the success of particular sites and curriculum models. Within waves there are generally two or more sites using the same model. Though models were not randomly assigned to sites, the fact that both form of replication exist serves to greatly increase our confidence in the validity of measured effects of the various preschool models.* As far as we know there is no other study of preschools with a planned strategy of curriculum replication.

Strategy for Analysis

The strategy for analysis is dictated in large part by the constraints on the study. First, we will focus principally on the analysis of cognitive growth. To do otherwise would be to seriously overplay the existing data. In doing this we recognize that we are not even attempting to capture the richness of a preschool experience or the largest part of the differences among preschools.

Second, we display the data in a more complete manner than is normally done. The limitation of theory that we bring to the analysis should not foreclose the possibility that other people could bring other theories and questions to put to the data. Though cumbersome, the intent is to let others explore, their favorite issues.

* In fact there is a paradox here. Were the models randomly assigned, implying that sites were forced to accept a particular model, we might find it hard to generalize from the results to a situation where sites had a free choice of which model to choose.

Third, the lack of a true experimental design puts the analysis of the data into a never-never land. Had we random assignment of curricula (treatments) to sites, then a comparison of treatments would yield us unbiased estimates. If we had random assignment of classes to PV and NPV groups within sites, then a comparison of the two sets of classes would yield unbiased estimates. If we had two random samples of children from the same population -- one going to Head Start and one not, then estimates of the general effects of a Head Start experience would be unbiased. But we have no random assignment, so all estimates are biased in some unknown fashion. Estimation of effects thus becomes an art instead of a science. There are numerous statistical techniques to help reduce bias (matching, covariance, blocking, crossed designs and standardization techniques). Each may be helpful depending on the adequacy of the structural model we are trying to fit. That is where the essential problem lies, for we have no a priori way of determining which is the best analytic model. Given this state of affairs, we follow Tukey's advice: "As in the famous discussion between Student and Fisher and the interjections by Sir Harold Jeffreys, it may not be a bad thing to use all the allowed principles of witchcraft and not just one set." (Tukey, in press, p. 112.)

We will not use all of the principles of methodological witchcraft but we do use a number. In particular, our strategies for removing bias in the data depend on (1) our choice of a statistical model; (2) our choice of variables; and (3) our assessment of the accuracy with which the data are measured. Different decisions in these areas of judgment lead to a variety of estimates of "effects." To some extent the variability of the estimates will aid in our determination of confidence about the magnitude of the effects. Thus the variability of results from different analyses gives us a sense of confidence limits for the reported effects. Such an approach will generally inspire caution in interpretation, for most of the effects found in this study are small. On the other hand, large effects which turn out to be robust -- insensitive to variations in analysis methods -- presumably should inspire confidence.

As we note later, estimates of statistical confidence in this study are compromised in a number of ways. First, we report a large number of comparisons. The effective significance level for any one comparison is thereby reduced. We also note that we carried out a larger number of comparisons that remain unreported. Second, the variety of methods and models used on the same data lead to statistical estimates which are not independent of another. Third, the lack of random assignment at any level leads us to make ex post facto arguments about the representativeness of the data for any given population. To the extent that our arguments are inaccurate, our inferences to larger populations based on sample estimates may be both biased and statistically imprecise.

Organization of the Report

This report contains seven additional chapters and assorted appendices. With the exception of Appendix A, all appendices are in a separate volume. Chapter II sets out an overview of the study design, describes the data collection instruments and procedures, and contains a brief discussion of each of the twelve curriculum models. Chapter III describes characteristics of the sample of children, classrooms, and sites. We consider two issues in detail. First, we explain our reasons for reducing the original total sample of over 6300 children to an analysis sample of 2235 children. Second, we focus on problems posed by the final analysis sample. A number of exemplary tables are included in Chapter III.

Chapter IV attempts to estimate the average effects of a Head Start experience on children. The procedure used is primarily descriptive. Our strategy is three-fold. First we present actual gains for various groups of children. Second we estimate what would have happened to the scores of children had they not been exposed to Head Start. Third we present comparisons of the expected to the actual gains for the various groups of children. Appendix B contains supplemental tables for this chapter.

Chapter V discusses a variety of analysis problems. First it considers the issue of an appropriate unit of analysis. Second, it generally describes strategies for overcoming biases in estimating differences among non-randomly selected groups. Third it describes in detail two analysis strategies used in Chapters VI and VII.

Chapter VI considers the question of differences between the effects of Planned Variation Head Start and conventional Head Start experiences. To answer this question we disregard differences among Planned Variations curricula and contrast children in PV classes with children in NPV classes. Various subsets of these two groups are also contrasted. Appendix C supplements Chapter VI with extra tables.

Chapter VII focuses on the issue of differences among curricula. The data for this issue are gathered from preceding chapters and new summary statistics are generated by other analytic techniques. Appendix D supplements Chapter VII.

Chapter VIII summarizes the preceding discussions. Though a summary, it also pinpoints the major findings of the analyses and raises questions regarding their importance and stability. In particular we focus on four major results of the study: (1) the magnitude of the overall estimated effects of the Head Start experience; (2) the overall similarity of the effects among the different programs; (3) the strength of one model in imparting specific information; and (4) the extraordinary success that one curriculum seems to have in raising the IQ level of children. Appendix E contains an extensive discussion of the fourth result.

Chapter II

DESIGN, DATA COLLECTION ACTIVITIES AND THE CURRICULUM APPROACHES

Overview:

This chapter describes the overall design of the Planned Variation Wave Two study. It also includes a brief description of the measures used in the study and of the 12 curriculum approaches. The next chapter describes the characteristics of the children and their classrooms.

Design:

Thirty-seven sites had Planned Variation (PV) curricula in 1970-71. Twelve curricula (models) were represented. There were comparison classes at 14 of the 37 sites (on-site comparison) and at seven locations not having Planned Variation classes (off-site comparisons). Table II-1 displays this information. Columns 1, 2 and 3 of Table II-1 show the names of the twelve curricula, a site code for each site, and the location of the site. The first two digits of the site code refer to the model (e.g. all Bank Street sites have codes beginning with 05). The second two digits specify the site within the model (Tuskegee is site 0510). With the exception of the Enabler model, which is unique to the PV study, the model and site codes were assigned as part of the Follow-Through evaluation and contain no information other than identification of model and site.

Column 5 of Table II-1 contains the year of entry of the site into the Planned Variation study. Fifteen of the thirty-seven sites were also in Planned Variation in 1969. Columns 6 and 7 show the number of classrooms in the site. Column 6 shows the number of Planned Variation classes. Column 7 shows the number of comparison classes. Blanks in column 6 indicate that the site was an "off-site comparison" site. Note that the off-site comparisons are paired with Planned Variation sites and are given the same site code number as a Planned Variation site. Blanks in column 7 indicate that there were no comparison classrooms at that site.

A few things should be noted from the table. First, three of the twelve models have only one site (Pittsburgh, REC, and N.Y.U.). Though these models are included in analyses when possible, confidence about their effects will necessarily be less than for the other models. Since there is no site level replication for these models, effects of the model and of the specific site cannot be separated. Second, there is an uneven distribution of Planned Variation and comparison classrooms among the sites. As we point out in the next chapter, however, not all of the classes were tested -- the tested sample levels out the number of classrooms per site. Third, only 14 of the 37 Planned Variation

TABLE II-1

HEAD START COMMUNITIES 1970-71

CURRICULUM MODEL SPONSOR	SITE CODE	SITE COMMUNITY	TESTING LEVEL	YEAR SITE JOINED STUDY	NUMBER OF PV CLASSES	NUMBER OF COMPARISON CLASSES
Nimnicht (Far West Laboratories)	02,02	Buffalo	I	70	11	
	02.04	Duluth	III	70	9	
	"	St.. Cloud	III	70'		2
	02.05	Fresno	III	70	4	
	02.09	Salt Lake	I	69	6	
	02.13	Tacoma	II	70	7	
Henderson (Tucson)	03.08	LaFayette	III	69	17	
	"	Albany	III	69		4
	03.09	Lakewood	I	69	8	
	03.16	Lincoln	III	70	7	
Bank Street	05.01	Boulder	III	70	4	1
	05.10	Tuskegee	I	69	12	
	05.11	Wilmington	II	69	9	
	"	DeLaWar	II	69		4
	05.12	Elmira	III	70	7	3
Becker & Englemann (Oregon)	07.03	E. St. Louis	III	69	9	4
	07.11	Tupelo	III	69	4	4
	07.14	E. Las Vegas, NM	II	70	5	
	"	W. Las Vegas, NM	II	70		4
Bushell (Kansas)	08.02	Oraibi	III	69	7	
	"	Acoma	III	69		4
	08.04	Portageville	III	69	4	4
	08.08	Mounds, Ill.	II	70	5	2
Weikart (Hi-Scope)	09.02	Ft. Walten Bch.	III	69	5	
	"	Pensacola	III	69		3
	09.04	Central Oz	I	69	16	
	09.06	Greeley	III	70	4	3
	09.10	Seattle	II	70	6	3
Gordon	10.01	Jacksonville	I	69	3	
	10.02	Jonesboro	III	69	3	3
	10.07	Chattanooga	III	70	9	4
	10.10	Houston	II	70	7	4
EDC	11.05	Washington	III	69	5	4
	11.06	Paterson	II	70	4	4
	11.08	Johnston Co.	III	69	6	4
Pittsburgh	12.03	Lock Haven	III	70	7	
	"	Mifflenburg	III	70		4
REC	20.01	Kansas City	III	70	8	
N.Y.U.	27.01	St. Thomas, VI	I	70	4	4
Enablers	27.04	Billings	II	70	5	
	27.05	Colorado Spr.	II	70	6	
	27.03	Bellows Falls	II	70	6	
	27.02	Newburgh	I	70	8	
	27.01	Puerto Rico	I	70	6	

sites have on-site comparison groups. Though 7 more Planned Variation sites have matching off-site comparisons, 17 Planned Variation sites have no matched comparison groups at all. Fourth, inspection of Table II-1 reveals that while the sites are generally spread around the country, for some models there is little spread in site location. For example, all of the Gordon sites are in the South.

Each of these observations serves to complicate the analysis. Thus, while the structure of the design--sites nested within models and Planned Variation and comparison classes nested with sites--appears relatively straightforward, there are complications involved in carrying out a conventional analysis.

Data Collection Activities

Column 4 in Table II-1 indicates the level of testing and evaluation carried out in the various sites. Primarily because of economic constraints, not all children in all sites were tested on the full range of measures. There were three levels of evaluation activities. Table II-2 describes the activities at each of the three levels. Level I is the most basic. Nine Planned Variation sites fit this category. Only one comparison site is in Level I. No data gathering at this level involved the children. Teachers completed demographic information forms and filled out the California Social

Table II-2

Three Levels of Planned Variation Evaluation Activities

Level of Evaluation	Data Collection Period		
Level I	Fall		Spring
1) Teacher completed classroom information forms -- for child demographic data	X		X
2) Teacher completed California Social Competency Scale -- one for each child	X		X
3) Sponsor ratings of Level of Implementation	X	X	X
4) Head Start Directors ratings of Level of Implementation	X	X	X
5) Teacher and Teacher Aide survey			X
Level II (includes all activities in Level I and the following)			
6) Classroom observations	X		X
7) Basic Child Test Battery a. Preschool Inventory b. NYU Book 3D c. NYU Book 4A d. Motor Inhibition Test	X		X
8) Child completed Ethnic Heritage Test	X		X
Level III (includes all activities in Levels I and II and the following)			
9) Stanford-Binet testing on random one-half of children in all tested classes	X		X
10) 8-Blocks Sort Task -- given to other random one-half of children in all tested classes			X
11) Parent Interviews -- administered to parents of children taking the 8-Block Sort Task			X
12) Intensive Case Studies (U. of Maryland)			X

Competency Scale for each child in their classrooms in both the fall and spring, and both teachers and teacher aides responded to a questionnaire requesting information about their own backgrounds, teaching experiences and attitudes. In addition, model Sponsors and Head Start Directors rated the level of implementation in the classrooms in each site.

All data collected at Level I was also collected at Level II. In addition, three other sets of data were gathered at Level II. Classroom observations were made in both the fall and spring by observers using the SRI Classroom Observation Instrument (see report on "Implementation in Planned Variation--1970-71"). All children in tested classrooms were administered the Basic Test Battery in both the fall and spring. Four tests were included in the Basic Battery--the Caldwell Preschool Inventory (PSI), NYU Booklet 3D, NYU Booklet 4A and the Motor Inhibition Test. Finally, black and Spanish children whose parents were willing took a test assessing their knowledge of their ethnic heritage. Ten Planned Variation sites were classified as Level II. Of the ten sites, four had on-site comparison classes which were also tested at Level II. Finally, two of the ten Planned Variation sites had off-site comparison classes tested at Level II.

Level III sites had all the data collection carried out in Level I and Level II sites and, in addition, four other activities were added. One randomly chosen half of the children in each tested Level III classroom were administered the Stanford-Binet in both the Fall and the Spring -- the same children received the test both times. The children in the other half of the class, along with one of their parents or guardians, were administered the 8-Block Sort Task in the Spring. Additionally, the parents or guardians of the children in this group completed a parent questionnaire which asked about attitudes toward Head Start, their child and the Planned Variation model used in their child's classroom. Finally, a small number of children in each of the Planned Variation Level III sites formed the sample for an intensive case study carried out and reported by the University of Maryland (see Head Start Planned Variation Case Studies -- 1970-71). Eighteen Planned Variation, ten on-site comparison, and five off-site comparison sites were assigned to Level III.

Descriptions of Data Collection Instruments

"The Quality of Planned Variation Data" describes in detail many of the instruments used in this study. The report on "Implementation in Planned Variation -- 1970-71" describes in detail most of the rest. We urge readers who want more than a cursory description to refer to those reports. In this section, we merely indicate the principal intent of the instruments, briefly describe how they were used in the data collection and note whether we will be using data from them in this report.

Before describing the instruments, some mention should be made of the strategy for data collection used by the Stanford Research Institute. For questionnaires completed by teachers, teacher aides and Head Start directors, the approach was to request that the forms be filled out and to pay the respondents a small stipend for their time. If the forms were incomplete or patently inaccurate they were returned to the person who filled them out with a request for clarification. In some instances, as with the Classroom Information Forms filled out by the teacher, this process was repeated a number of times. Generally a site co-ordinator was present to encourage teachers and Head Start directors to finish their forms quickly and accurately.

The site coordinator was also responsible for hiring, training, and overseeing a staff of local testers. Site coordinators were local people initially trained at SRI headquarters in Menlo Park, Cal. The reason for using local testers was to insure that sufficient rapport existed between teachers and children. By and large, SRI personnel visited every site during the testing period to answer questions about procedures and to evaluate the quality of the testing. The main exception to this general procedure was the Stanford-Binet testings. Here, certified testers from as near the local sites as possible were hired to do the testing.

The classroom observations were also carried our by locally hired personnel after they had been extensively trained by SRI personnel. The Case Studies were completed by students and faculty from the University of Maryland.

Testers and observers were instructed to fill out a short questionnaire after they had tested each child, indicating problems with the test session. The responses to these questionnaires have proved to be very useful in the data cleaning effort preceeding data analysis. When the data were gathered from the sites, they were returned to SRI and subjected to a careful screening before being placed on IBM cards and subsequently on magnetic tape.

By and large, we have been very pleased with the quality of the data. For an evaluation of the data gathered in 1971-72, the reader is referred to "The Quality of the Data in Planned Variation -- 1969-72".

The following briefly describes each of the instruments used in 1970-71. We also indicate the extent of each instrument's use in this and other reports in this series.

1. Classroom Information Form: This instrument was used to gather information about the background and family characteristics of every child in the sample. Teachers completed the instrument by gathering information from Head Start application blanks and interviews with parents. A validity study of selected items from a similar form used in 1971-72 yielded encouraging results (see "The Quality of the Data"). Information from this form is heavily used in this report and in the report "Cognitive Effects of Preschool Programs on Different Types of Children".

2. California Preschool Social Competency Scale: This is a teacher completed rating scale of 30 items designed to "measure the adequacy of preschool children's interpersonal behavior and the degree to which they assume social responsibility" (Levine et al., 1969, p.3). An extensive description of the measure is included in "The Quality of the Planned Variation Data". This measure is only briefly analyzed in this report. Completion of the scale by teachers suggested to us that among classroom and among site comparisons would be illegitimate. The reason is simply that teachers may consider their own classrooms as the reference group for rating students. Since the compositions of classrooms vary greatly, the ratings may lose comparability when they are taken out of the immediate context of their classroom.

3. Sponsor Ratings of Implementation: This rating form is fully described and analyzed in the report on "Implementation".

4. Head Start Directors Ratings of Implementation: This form is similar to the Sponsor Rating except that it was completed by the Head Start Director. It is discussed in the report on "Implementation".

5. Teacher and Teacher Aide Survey: These forms assess teacher and teacher aide background, teaching experiences and attitudes towards the Planned Variations model. They are extensively analyzed in the "Implementation" report. In this report, we .use some items taken from these surveys.

6. Classroom Observation Instrument: This is a broad range objective observation instrument developed at the Stanford Research Institute to assess the degree of implementation of classroom processes and child outcomes in the various programs. Trouble with the coding on the classroom observation tape limited our use of this important instrument. An analysis of some results from it are included in the

report on "Implementation" and an extensive analysis of its use in 1971-72 is under preparation by SRI.

7. Basic Child Test Battery: Four tests are included in this battery. The results from these tests are extensively analyzed in this report. Additionally, results from one of the tests, the Caldwell Preschool Inventory, are used in the report on "Cognitive Effects of Preschool Programs on Different Types of Children". Complete descriptions of the tests are in "The Quality of the Planned Variation Data".The four tests are:

a. Caldwell Preschool Inventory. (PSI) The PSI was developed to assess general achievement in preschool in areas deemed necessary for later success in school. Specifically developed for preschool populations, 64 items tap areas of general knowledge, listening and word meanings, listening and comprehension, writing, copying, quantitative skills, and speaking and labeling. Though the test was originally designed to have four factors, factor analyses of our data revealed only one factor which seemed to cut across all areas tapped by the test. Consequently, we simply summed the items to create a score on the test. Internal (KR-20) reliability is roughly .90. (See "The Quality of the Planned Variation Data".) By and large, we consider this test a measure of general achievement in preschool.

The scoring procedure for the test is not normed for age and as a consequence, pre-scores on the PSI are highly and positively correlated with the age that the child enters the program. The PSI also correlates roughly 0.50 with the Stanford-Binet, which in turn has a slightly negative correlation with age. The Stanford-Binet IQ score is obtained by dividing a calculated Mental age by chronological age -- the division by age makes the IQ score comparable across ages. The Mental age score taken alone can be thought of as the Binet score uncorrected for age. Mental age on the Binet correlates roughly .75 with the PSI. Assuming both tests have a reliability of .9, we find that the correlation among the "true score" parts of the PSI and the Binet score unadjusted for age is roughly .83*. Though this correlation is far from perfect, it suggests that the two tests are tapping somewhat the same domain.

b. NYU Book 3D. The NYU booklets were designed to measure areas of specific preschool achievement. Book 3D

*The sample used for these estimates and other estimates on following pages of this chapter was the same sample used for the correlation matrix on Page 120 of "The Quality of the Planned Variation Data" for estimates of the reliabilities of the two tests.

is designed to tap achievement in pre-math (seven items), pre-science (seven items), and linguistic concepts (five items assessing knowledge of prepositions). Both NYU booklets (3D and 4A) were extracted by SRI from the Early Childhood Inventories developed by A. Collier and J. Victor at the Institute for Developmental Studies at the NYU School of Education. Two scoring systems are used in the analyses in this report. First, a simple summary score obtained by adding together all correctly answered items is used. A factor analysis of the Book 3D suggested that there was only one stable, interpretable factor.* Estimates of internal reliability for the total score are generally in the range of 0.60-0.70. In this report we use 0.65 as a reliability estimate for individual scores. Moreover, the single score seems to have a ceiling problem for some groups of older children on the post-test results. See "The Quality of the Planned Variation Data" for discussions of these issues. Second, a set of scores is obtained by considering the three sub-tests as criterion-referenced measures. Using these measures, we report the percentages of children in various sites

*A factor analysis of Books 3D and 4A together convinced us to keep the tests separated for analytic purposes.

and models for each sub-test who obtain either a perfect score or only one item incorrect at post-test time. We also report the percentages of children in these groups who fail to get more than one item correct on each sub-test.

A score derived from a summing of correct items for Book 3D bears a very strong relationship to the PSI. By and large, different sub-samples of the data reveal correlations of about 0.70 at pre-test time (see page 176 of "The Quality of the Planned Variation Data"). Adjustment of this correlation for the reliabilities of the two tests (PSI reliability is roughly 0.90 and Book 3D reliability is roughly 0.65) yields a corrected correlation coefficient of roughly 0.95 indicating that the two tests are tapping almost entirely the same domains.

c. NYU Book 4A. This test is designed to tap achievement in three areas: knowledge of alphabet names (nine items); knowledge of numeral names (six items); and knowledge of shape names (three items). The development of scores for this test was similar to the development of scores for the Book 3D. A single summary score is analyzed in this report along with three criterion-referenced measures. With the exception of the third sub-test we follow the same rules for creating our criteria, as we did for Book 3D. In the third sub-test, we required that the student answer all three questions correctly to meet the criterion. The single score on Book 4A has an internal reliability of roughly 0.65 for the pre-test. To some extent this reliability is reduced by a minor floor problem in the Fall testing. For all groups the Book 4A scores were positively skewed in the Fall and more normally distributed in the Spring. Pre-test scores for Book 4A and the PSI correlate roughly 0.45-0.50, with the Book 3D the correlations are roughly 0.40-0.45 and with the Stanford-Binet, the correlations are roughly 0.40. Overall, then, though the Book 4A is assessing somewhat similar areas as the PSI, Book 3D and the Stanford-Binet, there is considerable residual unique variance associated with the test.

d. Motor Inhibition Test. This test was developed by Hagen and Degerman (see Maccoby et al., 1965) to measure a child's ability to inhibit movement when the task demands it. Three tasks are used to assess inhibition, the Draw a Line slowly task, the Walk slowly task, and the Pull Truck slowly task. Four preliminary items assess the child's understanding of the concepts of slow and fast. A substantial proportion of the sample of children in this study (over 50%) failed to answer two or more of the four pre-test items correctly, in either the Fall or Spring, indicating that these children did not understand the two concepts. The scores on the Motor Inhibition test were not analyzed for these children. Analyses of the three sub-tests indicated that the first two tasks yielded scores that correlated roughly 0.46. Correlations of the first two tasks with the third task were roughly 0.24. The low correlations with the third sub-task indicated to us that it was either unreliable or was measuring something other than the other two subtasks. Consequently, we formed a measure of the Motor Inhibition by summing the amounts of time in seconds taken to complete the first two sub-tasks. Following Maccoby's lead and an inspection of the data, the log of this score was then taken. The log transformation removed the strong positive skewness from the new scores. This final score correlates in the 0.30 to 0.40 range with the NYU 3D and PSI and in the 0.15-0.20 range with the Book 4A and the Stanford-Binet.

8. Ethnic Heritage Test: Two tests were actually used here. The Ethnic Identity Questionnaire (EIQ) was developed by Manuel Ramirez III at the University of California, Riverside, to investigate the ethnic identity of Mexican-American children and the Children's Cultural Awareness Scale (CCAS) was developed by Edward J. Barnes at the University of Pittsburgh to explore the cultural awareness of Black children in the Head Start Planned Variation Study. Scores from neither test are used in this report.

9. Stanford-Binet: The Stanford-Binet Intelligence Scale is a well-known measure of "general intelligence". The 1960 revision was used in this study. A single measure of IQ is used i1 this report. After extensive checking for matched pre- and post- birthdates and valid items, the score was calculated by dividing a child's Mental Age derived from the test by his chronological age in months and then corrected for age-related fluctuations in variance using the revised Pinneau tables (see Terman and Merrill, 1960).

10. 8-Block Sort Task Test: The Eight Block Sort Task is a measure of maternal teaching style and interaction styles between mother and child. The score used in this report ranges from 0-8 points and indicates the success of the mother in teaching the sorting tasks to the child. (See "The Quality of the Planned Variations Data" for an extensive discussion of this measure.)

11. Parent Interviews: This measure assesses parents' attitudes toward their children, Head Start and Planned Variations curricula. Although the interviews were conducted with only a small number of children, some of the items in the interview are analyzed in this report.

12. Intensive Case Studies: In all three years of Planned Variations, students and staff of the University of Maryland's Institute for Child Study did extensive case studies of a few selected children in Planned Variations. (See Dittman and Kyle, in press, for a report of these efforts.)

The Curriculum Approaches (Models)

This section briefly describes the twelve models used in the Planned Variations study in 1970-71. As we noted earlier, each of the approaches, with the exception of the Enabler model, has been developed and is sponsored by some group of people in a University or private corporation. The descriptions are intended to reflect the goals and expectations of the sponsors rather than to be a critical analysis. As presented, they are idealized descriptions of the twelve treatments. These sponsored approaches were included in Head Start Planned Variations because they were considered to be promising methods for working with disadvantaged children and families and because they were unique in some significant way. Nevertheless, the sponsors share common orientations. All of them seek to develop children's learning abilities. All are convinced of the importance of individual and small group instruction and frequent interchange between children and concerned adults. All attempt to make learning interesting and relevant to the child's cultural background. All believe that the child's success in learning is inseparable from his self-esteem, motivation, autonomy, and environmental support, and all attempt to promote successful development in these domains while fostering academic goals. The sponsors differ among themselves chiefly in the priorities which they assign to these objectives and in the sequences through which they pursue them.

It is important to recognize that the concept of Planned Variation was not intended as a means of finding a single "best" method for educating disadvantaged children. A wide variety of groups of children are included in this study, and a program that is appropriate for some may not be appropriate for others. Some approaches, for example, are primarily concerned with parental involvement and community control, while others place primary emphasis on the curriculum, the teacher, and the classroom. The following paragraphs briefly attempt to capture the emphasis of each model.

EDC Open Education Curriculum
Educational Development Corporation (EDC)

Sponsor Contact: George Hein

EDC has an open classroom approach derived from the British primary school model and theories of child development. It believes that learning is facilitated by active participation in the process. The classroom provides a setting in which there is a range of materials and activities from which the child can choose. Academic skills are developed in a self-directed way through classroom experiences. The role of the teacher is one of leading the child to extend his own work and generally involves working with an individual child or small group.

The Systematic Use of Behavioral Principles Program
(Engelmann-Becker)
University of Oregon

Sponsor Contact: Wesley Becker

The primary focus of the Engelmann-Becker program is on promoting skills and concepts essential to reading, arithmetic and language achievement, with particular emphasis on remedying language deficiencies. The main techniques are programmed materials, structured rapid-fire drills, and positive reinforcements of rewards and praise to encourage desired patterns of behavior. Small study groups of five to ten children are organized by teachers according to ability levels in order to facilitate presentation of patterned learning materials and to elicit verbal responses from children.

The Bank Street College of Education Approach
Bank Street College of Education

Sponsor Contact: Elizabeth Gilkeson

The Bank Street approach emphasizes both learning and social-emotional development of children on the premise that they are intertwined. The teacher functions as a supportive adult whom the child can trust, and teaches by relating and expanding upon each child's response to his experiences. The classroom is viewed as a stable environment and workroom for the child in which he is encouraged to explore, make choices and carry out plans. Academic skills are presented in the context of classroom experiences.

The Behavior Analysis Approach
Support and Development Center for Follow-Through, University of Kansas

Sponsor Contact: Don Bushell

The Behavior Analysis approach has three predominant aspects. First it emphasizes academic and social skills. Individualized programmed materials are the primary teaching mode. Second it makes systematic use of positive reinforcement. A token exchange system is used to support children's learning efforts. Third it employs parents as members of the instructional team as well as behavior modifiers. They receive training and work in the classroom in shifts throughout the year.

Individually Prescribed Instruction and the Primary Education Project (IPI)
Learning Research and Development Center, Univ. of Pittsburgh

Sponsor Contact: Lauren Resnick

The IPI approach provides an individualized program of instruction for each child which teaches him academic skills and concepts in the areas of language, perceptual motor mastery, classification, and reasoning. The materials are sequenced to reflect the natural order in which children acquire key skills and concepts. Diagnostic tests determine each child's strengths and weaknesses and are used by the teacher to prescribe instructional materials appropriate to his needs. Positive reinforcement, both social and concrete, is given continually for success in learning.

The Responsive Environments Corporation Model (REC)
Responsive Environments Corporation

Sponsor Contact: Lori Caudle

The REC model uses specially designed, self-correcting multi-sensory learning materials which strengthen school readiness skills in language and reading. They are designed to teach basic concepts while allowing children to make choices, work independently, and set goals for themselves. Teaching machines in the form of "talking typewriters" and "talking pages" involve children in learning by seeing, tracing, typing, imitating and discriminating among sights and sounds and by recording and listening to their own voices.

The Florida Parent Educator Model
University of Florida

Sponsor Contact: Ira Gordon

The Florida approach is not a specific classroom instructional model but is designed to work directly in the home. It focuses on the parent, believing that the parent is the key agent in a child's development. The major goals of the program are to develop educational competence in the child and to develop an atmosphere in the home which will foster continued growth. An important role is played by paraprofessionals called parent educators. The parent educator spends half-time with the teacher in the classroom and the other half making home visits. The home visit involves bringing tasks into the home and instructing the mother how to teach them to the child.

The Tucson Early Education Model
University of Arizona

Sponsor Contact: Ron Henderson

The Tucson model has a flexible child-oriented curriculum which focuses simultaneously on four areas of development: language competency, intellectual skills, motivational skills and societal skills. Emphasis is placed more on learning to learn skills than on specific content. The content is individually determined by a child's environment and interests. The classroom is arranged in interest centers for small groups. The teacher's role is to work on a one-to-one basis with the child, arrange the classroom setting and encourage interactions between the child, his environment and others.

Responsive Educational Program
Far West Laboratory for Educational Research and Development

Sponsor Contact: Glen Nimnicht

The Responsive Educational model emphasizes self-rewarding learning activities and a structured environment responsive to a child's needs and interests. The model encourages the child to make interrelated discoveries about his social world and physical environment and stresses the importance of the development of a healthy self-concept. The classroom is a controlled environment in which the child is free to explore various learning centers, games and activities. Problem solving and concept formation as well as sensory and perceptual acuity are stressed and the pace of all learning activities is set by the child for himself.

Cognitively Oriented Curriculum
Hi/Scope Educational Foundation

Sponsor Contact: David Weikart

The Cognitively Oriented Curriculum combines Piagetian theory and an open classroom approach. It uses a cognitively oriented curriculum and emphasizes the process of learning rather than particular subject matter. It stresses a child's active involvement in learning activities. The teacher takes an active role. Additionally, home training is seen as part of the program and the teacher suggests tasks for the mother to present to the child at home.

The Enabler Model
Office of Child Development

Sponsor Contact: Jenny Klein

The Enabler Model does not involve affiliation with a particular instructional approach. It is build on goals prescribed by each community for itself. The development and implementation of this model are facilitated by the assistance of an OCD consultant who takes a very active role in all aspects of the program. Thus projects with the Enabler Model may differ considerably in the approach and style of their educational tactics, but all share a commitment to high levels of parent participation in policy making, program planning and classroom operation.

The Independent Learner Model
New York University

Sponsor Contact: Don Wolfe

In the Independent Learner model, learning occurs principally in structured small-group instructional "games" where children of different ability levels teach one another and become relatively independent of the teacher. The verbal interactions among children are implicit in the process and are a direct stimulus to language development. Experiences in phonics blending and decoding skills stimulate reading ability and language-math-logic games such as Cuisenaire rods and matrix boards promote mathematical comprehension.

CHAPTER III

CHARACTERISTICS OF CHILDREN, CLASSROOMS AND SITES IN THE 1970-71 PLANNED VARIATIONS SAMPLE

Introduction:

This chapter has six sections. The first section describes the 37 Planned Variation sites. After criteria for the selection of sites are discussed, characteristics of the children in the sites are summarized for each site. Additionally, data on location and structural characteristics of the sites are shown. The second section describes the comparison sites in the same way. Section three describes the strategy used to reduce the total sample to a working analysis sample. Section four describes the analysis sample by child characteristics and classroom characteristics. Section five contrasts the Planned Variation and Comparison analysis samples. Section six reports analyses comparing pre-scores in the Planned Variation and Comparison samples.

I The Thirty-Seven Site Planned Variations Sample

The selection of the thirty-seven Planned Variation sites had a large part in determining the overall design of the study. Although we briefly mentioned the distribution of sites by models and testing level in the second chapter, it is useful to give an overview of this information before we discuss the criteria used to select the sites.

Table III-1 shows the number of sites for each of the twelve models in the study. The models are cross-classified by the level of testing used in the study. Four things should be noted in Table III-1. First, for analysis purposes there are really only 11 models. The Virgin Islands is the only NYU site and it is a Level I. Second, two other models (Pittsburgh and REC) have only one site. Although we carry out extensive analyses of the outcomes of the programs in these sites, inferences about the effects of the models are weakened by the fact that there is no experimental replication. Third, there is no Level III Enabler site. Thus, we have no Stanford-Binet or 8-Block Sort data for this model. Fourth, for the remaining eight models an attempt was made to have at least two Level III sites and one Level II site per model.

TABLE III-1

Distribution of Planned Variation Sites Within Models by Testing Level

Models		Far West	Tucson	Bank St.	Oregon	Kansas	Hi-Scope	Gordon	EDC	IPI	REC	NYU	Enablers	Total
Testing Levels	I	2	1	1			1	1				1	2	9
	II	1		1	1	1	1	1	1				3	10
	III	2	2	2	2	2	2	2	2	1	1			18
Totals		5	3	4	3	3	4	4	3	1	1	1	5	

In Chapter I we described three criteria that Planned Variation sites had to meet in order to be included in the study. Briefly, each site had to have existing Head Start classes, each site had to be located in an area that fed into the Follow-Through school or schools, and each had to adopt the curriculum model used in the Follow-Through school. With the exception of the Enabler model sites, the same criteria applied in 1970-71 as in 1969-70. Other criteria, however, were also used in 1970-71. The expansion from 16 to 37 sites reflected a variety of design and political constraints.

In addition to the three previous constraints, the choice was influenced by an attempt to have three or more sites for each of the original eight models, by an attempt to expand the number of models from the original eight, and for reasons of geographic representation. A final constraint was imposed by the budget of the study. One result of these often conflicting constraints was that the characteristics of sites within models differed from model to model. Table III-2 displays the 37 sites grouped by model and contains some summary structural and demographic information about them. As noted earlier, at least demographic and the California test information were gathered in each of these sites.

A few things are clear from Table III-2. First, there is wide geographic diversity. All regions of the nation are represented. Second, large Northern cities are clearly unrepresented. There are no sites, for example, in New York City,

TABLE III-2

Characteristics of Head Start Planned Variation Sites
Total Sample 1970-71

SPONSOR	CODE	SITE	Testing Level	Region* Location	No. of Children	Entry Level	% Previous Preschool	Mean Age in Months	% Black	% P.R.	% Mex.-Am.	% Am. Indian	% White
Nimnicht	02.02	Buffalo	I	NU	178	K	46.3	52.6	75.7	6.8	0.0	0.0	17.5
	02.04	Duluth	III	NU	140	K	11.7	55.5	6.5	0.0	0.0	8.0	85.5
	02.05	Fresno	III	NU	70	K	2.9	53.0	85.7	0.0	14.3	0.0	0.0
	02.09	Salt Lake	I	NU	136	K	0.7	56.0	11.0	0.0	22.8	0.7	64.0
	02.13	Tacoma	II	NU	138	K	20.4	56.1	31.2	0.0	1.4	7.2	58.7
Tucson	03.08	LaFayette	III	SR	399	1	40.1	59.2	30.0	0.0	0.0	0.0	69.8
	03.09	Lakewood	I	NU	125	K	0.8	55.3	59.2	25.6	0.0	0.0	15.2
	03.16	Lincoln	III	NU	170	K	4.2	55.0	7.7	0.0	5.3	4.1	82.8
Bank Street	05.01	Boulder	III	NU	66	K	0.0	55.6	1.5	0.0	37.9	3.0	56.1
	05.10	Tuskegee	I	SR	262	1	16.0	64.8	89.3	0.0	0.0	0.0	10.7
	05.11	Wilmington	II	SU	158	K	12.0	52.4	98.1	0.6	0.0	0.0	1.3
	05.12	Elmira	III	NU	136	K	36.8	46.3	39.3	0.0	0.0	0.0	59.3
Becker & Engle-mann	07.03	E. St. Louis	III	SU	237	K	0.0	52.9	100.0	0.0	0.0	0.0	0.0
	07.11	Tupelo	III	SU	105	1	25.3	64.6	67.3	0.0	0.0	0.0	31.7
	07.14	E. Las Vegas, NM	II	NR	114	1	4.4	64.3	0.0	0.0	90.4	0.0	9.6
Bushell	08.02	Oraibi	III	NR	93	K	16.3	56.0	0.0	0.0	0.0	98.9	0.0
	08.04	Portageville	III	NR	74	K	2.7	55.7	40.5	0.0	0.0	0.0	59.5
	08.08	Mounds, Ill.	II	NR	89	K	5.6	53.5	69.7	0.0	0.0	0.0	30.3
Weikart	09.02	Ft. Walton Bch	III	SU	90	K	0.0	53.0	68.9	0.0	0.0	0.0	31.1
	09.04	Central Oz	I	SR	282	K,1	21.7	60.1	0.0	0.0	0.0	0.0	99.6
	09.06	Greeley	III	NU	77	K	16.9	57.7	1.3	5.2	76.6	0.0	16.9
	09.10	Seattle	II	NU	104	K	15.5	54.5	49.5	0.0	0.0	5.8	35.0
Gordon	10.01	Jacksonville	I	SU	60	K	13.3	51.7	93.3	0.0	0.0	0.0	6.7
	10.02	Jonesboro	III	SU	68	1	2.9	67.0	30.9	0.0	0.0	0.0	69.1
	10.07	Chattanooga	III	SU	170	K,1	1.8	61.6	83.9	0.0	0.0	0.0	16.1
	10.10	Houston	II	SU	75	K	0.0	55.5	63.5	0.0	35.1	0.0	1.4
EDC	11.05	Washington	III	SU	85	K	17.6	46.3	89.3	2.4	1.2	0.0	4.8
	11.06	Paterson	II	NU	135	K	0.8	53.4	92.5	6.0	0.0	0.0	1.5
	11.08	Johnston Co.	III	SR	117	1	32.5	66.4	53.0	0.0	0.0	0.0	47.0

TABLE III-2
(cont'd)

SPONSOR	CODE	SITE	Testing Level	Region* Location	No. of Children	Entry Level	% Previous Preschool	Mean Age in Months	% Black	% P.R.	% Mex.-Am.	% Am. Indian	% White
Pitts-burgh	12.03	Lock Haven	III	NU	120	K	15.8	51.4	0.0	0.0	0.8	0.0	98.3
REC	20.01	Kansas City	III	NU	178	K	7.4	54.3	34.5	0.0	42.4	0.6	22.0
N.Y.U.	26.01	St. Thomas, VI	I	-R	191	K	11.1	51.5	92.0	4.3	0.0	0.0	3.7
Enablers	27.04	Billings	II	NU	82	1	9.8	66.3	3.7	0.0	23.2	6.1	64.6
	27.05	Colorado Sp	II	NU	108	K	6.5	55.8	33.3	0.0	41.7	0.9	24.1
	27.03	Bellows Falls	II	NR	103	K,1	10.7	55.6	0.0	0.0	0.0	0.0	100.0
	27.02	Newburgh	I	NU	135	K	0.7	52.5	79.3	5.2	0.0	0.0	15.6
	27.01	Puerto Rico	I	-U	104	K	4.8	64.7	0.0	100.0	0.0	0.0	0.0
		TOTAL			4974		14.2	56.6	47.0	3.6	8.2	2.7	37.9

* The first code is either N or S, standing for North or the old South; the second code is either U or R, standing for either Urban or Rural.

Philadelphia, Chicago, San Francisco or Detroit. The cities representing the North -- e.g., Buffalo, Salt Lake, Elmira, Wilmington, and others -- should probably not be thought of as similar to the large metropolises.

Third, sites and models differ in aggregate characteristics of children. Some sites are predominantly Black while others are predominantly white or Mexican American. In some sites most of the children have had preschool experience -- in others none have. This holds for models as well; in the Florida model, for example, very few of the children had prior preschool experience while in other models a considerable percentage of children had previously attended preschool. By and large, children in the South are older and the probability of a child having had previous preschool experience is slightly greater if he is in the South. A number of Southern sites* and one Northern site (Billings) send their children directly into first grade from preschool -- we call these Entering First sites (E1 sites). Chattanooga, Central Ozarks and Bellows Falls graduate children into both Entering First and Entering Kindergarten classes. The majority of Chattanooga children enter E1 classes and the site is classified as E1 for analysis purposes. Central Ozarks is a Level I site and is not included in the analysis (see next section). The majority of children from Bellows Falls enter EK classes and the site is classified EK for analysis purposes. Variation in age among the sites is very highly correlated with elementary grade entering level (E1/EK). Older children attend E1 sites and younger children attend EK sites. At the classroom level the correlation between

*LaFayette, Tuskegee, Tupelo, E. Las Vegas, Jonesboro, and Johnston County.

mean class age and an E1/EK site level dichotomous variable is 0.922*. The reason for this high correlation is clear -- by and large, the Head Start experience directly precedes elementary school and most first grades enroll children at age six while most kindergartens enroll children at age five.

We have focused on the variation among sites in ethnic background, preschool experience and E1/EK because these variables are used extensively for control and stratifying purposes in the analyses described in later chapters. Each of the variables has a powerful relation to test scores and test score gains. Analyses contained in the report "Cognitive Effects of Preschool Programs on Different Types of Children" suggest that there may be important model interactions with both preschool experience and ethnic background.

Finally, there are both logical and empirical reasons for distinguishing E1 sites from EK sites in the analysis. Since the E1 sites are primarily Southern and involve older children, entering level might be viewed as a proxy for region and age. Moreover, it is not difficult to imagine that preschool teachers in E1 sites go about their jobs somewhat differently than do preschool teachers in EK sites. The fact that children in E1 sites will directly enter first grade might make the teachers conscious of a responsibility to prepare their children for beginning reading and arithmetic. Teachers in EK sites might

* The sample used here is the analysis sample of 166 classes described in a later section of this chapter.

well end up relying on kindergarten to share part of that responsibility.

II Description of the Comparison Sample

Table III-3 describes the comparison sample sites -- in the same terms as the PV sample in Table III-2. There are 22 comparison sites described in this table. All but one (St. Thomas) are either Level II or Level III tested sites.

The remarks in the preceding section about the diversity among Planned Variation sites in ethnic composition, proportion of children who have previously attended preschool, and age apply equally to the comparison sites. Although there is great diversity among the sites, however, a brief comparison of Tables III-2 and III-3 suggests that there is considerable similarity within locations between Planned Variation classrooms and comparison classrooms. The similarity between mean ages of the paired Planned Variation and comparison sites is particularly great, and, while the pairing by sites does not always eliminate differences in racial composition and preschool experience, it clearly has some effect on these variables.

The selection of comparison sites deserves some discussion. By and large, there was an attempt to obtain a comparison site for each of the Planned Variation sites. The idea was to find Head Start classes not funded by Planned Variation in a nearby location for each PV site. In theory, the comparison classes could exist in the same centers as the PV classes, though in practice this did not occur in 1970-71. When a reasonably

TABLE III - 5

Characteristics of Head Start Comparison Sites
Total Sample 1970-71

[illegible]	[illegible]	[illegible]	[illegible]	[illegible] *	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
N [illegible]	[illegible]	[illegible]	III	NU	69	K	11.8	57.4	0.0	0.0	0.0	97.1	2.9
[illegible]	[illegible]	[illegible]	II	SR	100	1	36.0	65.4	82.0	0.0	2.0	16.0	0.0
Penn.	[illegible] 01	[illegible]	III	NU	17	K	25.5	54.4	29.4	0.0	29.4	35.3	5.9
[illegible]	[illegible]	[illegible]	II	SU	83	K	1.2	52.6	57.8	2.4	0.0	38.6	[illegible]
	[illegible]	[illegible]	III	SU	50	K	0.0	54.3	32.0	0.0	0.0	65.0	[illegible]
[illegible]	[illegible]	[illegible]	II	SU	85	K	0.0	55.5	94.0	0.0	0.0	4.8	[illegible]
[illegible]	[illegible].11	[illegible]	II	SR	63	1	57.1	65.3	61.9	0.0	0.0	38.1	0.0
[illegible]	07.14	[illegible]	II	NR	84	1	7.1	61.6	0.0	0.0	98.8	1.2	0.0
[illegible]	[illegible]	[illegible]	II	NR	51	K	2.0	56.7	0.0	0.0	0.0	0.0	100.0
	[illegible]	[illegible]	III	NR	66	K	25.8	55.2	54.5	0.0	0.0	45.5	0.0
	[illegible]	[illegible]	II	NR	28	K	62.5	53.9	25.0	0.0	0.0	75.0	[illegible]
[illegible]	[illegible]	[illegible]	III	SU	66	K,1	23.1	59.2	84.6	0.0	0.0	15.4	0.0
	[illegible]	[illegible]	III	NU	59	K	8.5	56.1	0.0	0.0	45.8	50.8	3.4
	[illegible]	[illegible]	II	NU	44	K	74.3	52.6	79.5	0.0	2.3	15.9	[illegible]
[illegible]	[illegible]	[illegible]	II	SU	61	1	4.6	67.6	19.7	0.0	0.0	80.3	0.0
	10 [illegible]	[illegible]	II	SU	83	K,1	20.5	66.1	98.8	0.0	0.0	1.2	0.0
	10 [illegible]	[illegible]	II	SU	103	K	67.0	53.0	80.6	0.0	19.4	0.0	[illegible]
[illegible]	11.[illegible]	[illegible]	III	SU	70	K	14.3	51.1	100.0	0.0	0.0	0.0	[illegible]
	11 [illegible]	[illegible]	II	NU	37	K	0.0	52.2	83.8	16.2	0.0	0.0	0.0
	11.[illegible]	[illegible]	III	SR	78	1	26.0	67.1	70.5	0.0	0.0	29.5	0.0
[illegible]	12.03	Pitt[illegible]	III	NU	61	K	20.0	52.4	0.0	0.0	1.6	98.4	0.0
N.Y. [illegible]	[illegible] 01	St. [illegible]	[illegible]	[illegible]R	68	K	14.8	50.7	100.0	0.0	0.0	0.0	0.0
					1423		21.9	58.0	58.0	0.6	10.0	30.[illegible]	3.[illegible]

* The first code is either N or S, standing for North or the old South, the second code is either U or R, standing for either Urban or Rural.

comparable set of Head Start classes were found in the Planned Variation community, up to four of the classes were chosen to have the tests and measures administered to them. When it was impossible to find a comparable set of Head Start classes within a community there was an attempt to go outside of the community to nearby locations to find comparison children. In this manner, region and locality could be generally controlled in the analysis. If this tactic failed, no comparison classes were selected. No attempt was made to find comparison classes for the Enabler model sample of sites.

The strategy for selection of comparison sites within PV communities is based on the assumption that the comparison classes will be similar to the PV classes in all respects except those influenced by the PV curricula. This assumption has two problems. First, there were probably sources of bias introduced in the selection of Planned Variation classes. This could mean that the classes selected to use a PV model were originally different from the classes selected for comparison purposes. Second, the comparison classes may have been influenced by the existence of PV in the community and thus reflect a different situation than would have existed had there not been PV classes in the community.

With regard to the first problem, the constraint of selecting classes for PV from within Follow Through attendance areas provides a likely source of bias. In some instances this constraint entirely restricted the choice of PV classes. In this case, whatever biases existed in the initial selection of Follow Through schools were communicated to the PV classes. In other instances, the Head Start and CAP personnel could choose from among a variety of Head Start centers all within Follow Through attendance zones. Since no clear guidelines existed to determine their choice of PV classes, their selection could have introduced bias into the study. Speculation in this area is very difficult. Since we have no direct way of understanding what biases exist in the sample, our strategy is to control for as much as possible and hope that we are directly controlling the biases or that our control variables are strongly associated with the bias.

The second problem with the argument that the comparison classes may not be different from the PV classes except with respect to differences created by the PV curriculum arises from the effects of contamination. This issue only arises in places where both model and comparison classes are tested. Briefly, it stems from the fact that facilities and consulting services available to only a select group of classes within a community and not to other classes may create a situation that is intolerable morally and politically for Head Start directors and other supervisory personnel. In the case of the Planned Variation study, the PV classrooms were receiving extra equipment and the model teachers were receiving extra in-service and pre-service training beyond that available with the normal level of Head Start funding in the community. In these circumstances, it would not be unnatural for a Head Start director to let some of the equipment intended for PV model classes make its way into comparison classes. And it is natural for Head Start directors to let comparison teachers attend some pre-service and in-service training sessions. This situation might be aggravated in a community where the Head Start director was enamoured of the particular model being used in the model classes and not particularly impressed by the importance of the evaluation. Over the course of the year 1970-71, reports from the OCD consultants indicated that some contamination was occurring. When this evidence was known to SRI and OCD before sample selection for 1970-71, care was taken to exclude heavily contaminated classes from the comparison sample.

It is, however, practically impossible to estimate even roughly the effect that the contamination had in the various communities so, by and large, we ignore the problem in this report. If this type of study is to be done again, some systematic way of estimating the influence of contamination should be devised.

III Generation of the Analysis Sample

One problem faced in all analyses of large data bases is the creation of the sample used for analysis purposes. For a variety of reasons, all data is not usable in all analyses. Throughout this report, we will focus on one particular sample of 2,235 children. The reduction from the original sample of

6,297 rostered children to the sample of 2,235 children had two main steps. First, we eliminated certain entire sites from the total sample and also eliminated all children in classes which were not tested in the evaluation. Second, we eliminated some children in the remaining sites and classes because of missing data.

The focus for analysis in this report is on objective measures of the effects of different pre-school experiences on children. The major measures assessing these effects are the four cognitive tests and the Motor Inhibition test. The California measure, as noted in The Quality of the Data, should be viewed as a subjective child assessment. As such, it presumably has within-classroom validity but lacks across classroom validity. Since the children in Level I tested sites were not administered any of the four cognitive tests or the Motor Inhibition test, there was no reason to include them in an analysis prepared to assess the effects of the cognitive tests. The elimination of the Level I sites reduced the sample from 6,397 children to 4,864 students. Another reduction in size resulted from the elimination of two PV sites and one comparison site. Specifically, we eliminated from the general analysis sample both Oraibi and Fresno. The reason for eliminating Oraibi and its comparison site Acoma was simply that we felt they were not comparable either with other sites or with each other. Both Oraibi and Acoma are American Indian reservations in the Southwest United States. We felt, for reasons of different languages, cultures and experiences that neither site was comparable to the other sites in the analysis sample, and we felt for reasons of different languages and cultures that the two were not comparable to each other.

The second site eliminated from the analysis was Fresno. Fresno underwent considerable controversy during the school year over its Planned Variation model and at year's end, decided not to continue the model in the third year of study. This controversy not only seemed to affect the nature of the pre-school program as reported by the OCD consultant, but also influenced the quality of the data collection. After deliberation with SRI personnel responsible for the data collection, we decided that too many unknown biases existed in the Fresno data to make it a legitimate candidate for inclusion in the general analysis. There were no comparison classes in Fresno.

Both of these sites were excluded on the basis of intuition and subjective analysis rather than empirical data, thus there is room for argument on the validity of the decisions.

The elimination of these two sites and the comparison classes in Acoma reduced the sample from 4,684 children to 4,650 students.

The next step was to eliminate the non-tested classrooms from the analysis. This reduced the sample to 3,131 children. After this step, the number of classrooms was not reduced, though the number of children was reduced. We then eliminated all children who did not have a valid pre-test or post-test score on at least one test in the Basic Battery. One result of this was to eliminate all children who either left or entered the classes during the year. Thus a child was retained at this stage if he had a valid Fall or Spring score for either the PSI, Book 3D, Book 4A or the Motor Inhibition Test. A valid score was determined by the tester.

The next step was to eliminate all children who did not have a valid pre-test and post-test score on at least one of the following measures: PSI, Book 3D, Book 4A, Motor Inhibition, California, Stanford-Binet. Three final steps were taken. First, we eliminated all children who did not have a legitimate code for the background variables sex, age, preschool experience and race. These variables were necessary as key stratifying variables and would be difficult either to treat as missing values or to impute scores to. Second, we eliminated all children with an ethnic or racial origin other than Black, white or Mexican-American. Specifically, we removed Puerto Ricans, American Indians, Orientals and other non-Caucasian children from the sample. Our reasoning was that there were too few children in these groups for which to make reasonable comparisons. There were a total of only 47 American Indians, 31 Puerto Ricans, and less than 10 Orientals and other non-Caucasians in the sites included in the analysis sample. Third, we eliminated 22 children whose ages were under 44 months or over 74 months. These children were seen as distinct outlyers and not at all representative of the rest of the sample.

This concluded our sample reduction and left us with 2,235 children.

It is reasonable to expect that had other analysts been responsible for the analysis, they would have developed different decision rules. Our justification for those we developed was that they seemed at the time to be reasonable. One indication that our sample reduction was not extreme comes from the fact that only 2,567 children received the basic battery (in our selected sites) in the fall of 1970. Of these children, we retained 2,235 children, or 87%. Thus only 13% of the possible candidates for inclusion based on Fall tests alone were eliminated for one of the following reasons: (1) they did not remain in the class during the entire school year; (2) they did not recieve a Spring Basic battery; (3) they did not have valid scores in both the fall and the spring on one of the tests; (4) they were missing data on sex, pre-school experience, race or age; (5) they were in under-represented minority groups; (6) they were outlyers in terms of age. This seems like an extraordinarily low percentage of missing data-eliminated cases for a study of the size and complexity of the 1970-71 HSPV study. Another indication comes from the fact that there were 166 classes in the retained sites tested in the fall of 1970 and there are 166 classes retained in the analyses reported here. Thus, from the point of view of using the classroom as the unit of analysis, no data was lost -- although the classroom aggregates were computed on less than the overall possible number of cases. One class, for example, had only 3 eligible children, 3 classes had 4 eligible children, and 3 classes 5 eligible children. Over 70% of the classes, however, had over 10 eligible children, and the average class size of eligible children was 13.46.

IV Characteristics of the Analysis Sample

A. Child Characteristics

Table III-4 shows aggregate percentages and means for a variety of characteristics by site in the final analysis sample of Planned Variation children. Table III-5 shows the same data for the comparison sample of children. The child background characteristics shown are those which were found to have the strongest relationships to the test variables used in this report. They can be divided into three groups. The first group are child characteristics -- specifically, age, race and sex. The second group are family background characteristics -- family income, size of household, and extent of mother's education. The third group contains only one variable -- the child's prior experience in preschools. Also shown in the tables are the number of children in the site and the testing level of the site.

B. Classroom Characteristics

Tables III-6A and III-7A show means of classroom means by site for selected variables in the PV and NPV analysis sample respectively. Two groups of variables are shown. The first

TABLE III-4

Characteristics of Planned Variation Children in the Final Analysis Sample with Children as the Unit of Analysis

SPONSOR	CODE	SITE	Test. Level	# of children	% pre-sch.	mean age	% Mex.-Amer.	% White	% Black	* mean mother's educ.	* mean household size	mean income* (in hundreds)	% female
Nimnicht	02.04	Duluth	III	33	12.1	56.2	0.0	87.9	12.1	10.2	5.5	56.5	51.5
	02.13	Tacoma	II	47	27.7	55.3	2.1	53.2	44.7	11.1	5.4	42.1	51.1
Tucson	03.08	LaFayette	III	66	71.2	64.8	0.0	75.8	24.2	11.1	3.9	48.6	39.4
	03.16	Lincoln	III	70	2.9	54.9	5.7	84.3	10.0	11.2	5.5	46.2	40.0
Bank Street	05.01	Boulder	III	49	0.0	55.3	40.8	57.1	2.0	10.5	5.3	39.8	51.0
	05.11	Wilmington	II	56	10.7	52.5	0.0	1.8	98.2	10.4	4.9	28.1	48.2
	05.12	Elmira	III	35	74.3	54.0	0.0	51.4	48.6	11.4	5.2	46.7	45.7
Becker & Englemann	07.03	E.St.Louis	III	77	0.0	53.6	0.0	0.0	100.0	11.9	5.2	93.7	51.9
	07.11	Tupelo	III	81	27.2	64.5	0.0	28.4	71.6	9.5	6.1	39.3	54.3
	07.14	E.LasVegas, NM	II	55	1.8	64.4	90.9	9.1	0.0	9.7	6.2	35.3	43.6
Bushell	08.04	Portageville	III	63	3.2	55.4	0.0	58.7	41.3	9.7	5.7	34.3	55.6
	08.08	Mounds, Ill.	II	50	8.0	54.8	0.0	30.0	70.0	10.6	5.8	40.1	48.0
Weikart	09.02	Ft.Walton Beach	III	58	0.0	53.0	0.0	24.1	75.9	10.1	6.4	33.1	48.3
	09.06	Greeley	III	39	10.3	56.1	84.6	15.4	0.0	9.4	5.1	35.4	35.9
	09.10	Seattle	II	41	14.6	54.6	0.0	41.5	58.5	11.4	4.7	69.8	53.7
Gordon	10.02	Jonesboro	III	39	0.0	67.1	0.0	64.1	35.9	8.7	6.1	34.8	38.5
	10.07	Chattanooga	III	50	0.0	66.2	0.0	12.0	88.0	10.1	5.7	27.7	56.0
	10.10	Houston	II	44	0.0	56.1	22.7	2.3	75.0	10.1	3.7	31.3	47.7
EDC	11.05	Wash'ton.	III	20	30.0	51.3	0.0	5.0	95.0	11.0	4.5	41.3	65.0
	11.06	Paterson	II	60	0.0	53.1	0.0	0.0	100.0	10.0	5.5	48.8	35.0
	11.08	JohnstonCo.	III	70	35.7	66.6	0.0	47.1	52.9	9.3	5.9	36.5	55.7
Pittsburgh	12.03	Lock Haven	III	48	37.5	53.1	0.0	100.0	0.0	10.4	6.3	40.5	50.0
REC	20.01	Kans. City	III	56	10.7	54.4	42.9	19.6	37.5	9.8	6.1	65.9	48.2
Enablers	27.04	Billings	II	41	4.9	66.5	29.3	63.4	7.3	10.1	6.1	37.3	48.8
	27.05	Colo.Spgs.	II	47	4.3	56.0	42.6	21.3	36.2	10.2	5.6	37.1	43.8
	27.03	Bellows Fls.	II	41	9.8	58.1	0.0	100.0	0.0	10.5	5.0	80.4	51.2
				113.	15	[illegible]	[illegible]	[illegible]	[illegible]	10.2	5.5	45.1	48.3

*These means are not always based on n's equal to the number of children because of missing data.

TABLE III-5
Characteristics of Comparison Children in the Final Analysis Study with Children as the Unit of Analysis

SPONSOR	CODE	SITE	Test. Level	# of children	% pre-sch.	mean age	% Mex.-Amer.	% white	% Black	mean mother's educ.	mean household size	mean income (in hundreds)	% female
Nimnicht	02.04	St. Cloud	III	51	11.8	57.8	0.0	100.0	0.0	10.7	6.8	48.1	45.1
Tucson	03.08	Albany	III	63	33.3	64.7	3.2	4.8	92.1	10.2	5.9	35.2	49.2
Bank Street	05.01	Boulder	III	12	25.0	53.7	25.0	41.7	33.3	11.3	5.9	42.9	50.0
	05.11	DeLaWar	II	51	2.0	52.2	0.0	37.3	62.7	10.9	5.2	39.1	52.9
	05.12	Elmira	III	45	0.0	54.2	0.0	64.4	35.6	10.2	5.2	**99.0	51.1
Becker & Englemann	07.03	E.St.Louis	III	59	0.0	53.9	0.0	5.1	94.9	12.4	5.3	82.9	50.8
	07.11	Tupelo	III	51	62.7	65.0	0.0	29.4	70.6	9.2	6.3	30.5	49.0
	07.14	N.LasVegas, NM.	II	70	7.1	64.7	100.0	0.0	0.0	10.6	6.0	52.6	57.1
Bushell	08.04	Portageville	III	51	31.4	55.4	0.0	37.3	62.7	9.8	5.2	33.6	45.1
	08.08	Mounds, Ill.	II	17	58.8	54.8	0.0	82.4	17.6	10.4	5.6	64.1	58.8
Weikart	09.02	Pensacola	III	51	27.5	59.3	0.0	9.8	90.2	9.8	5.7	28.9	41.2
	09.06	Greeley	III	37	5.4	55.2	51.4	48.6	0.0	9.7	5.6	55.5	54.1
	09.11	Seattle	II	27	77.8	53.3	0.0	18.5	81.5	12.9	3.6	86.9	51.9
Gordon	10.02	Jonesboro	III	36	5.6	67.3	0.0	83.3	16.7	8.8	6.7	31.8	47.2
	10.07	Chattanooga	III	64	25.0	66.5	0.0	0.0	100.0	11.0	5.9	34.6	51.6
	10.12	Houston	II	48	81.3	55.6	16.7	0.0	83.3	10.6	4.9	41.5	39.6
EDC	11.05	Wash'ton.	III	39	17.9	52.1	0.0	0.0	100.0	9.7	6.7	37.8	43.6
	11.08	Paterson	II	28	0.0	52.6	0.0	0.0	100.0	11.0	4.9	37.5	60.7
	11.09	JohnstonCo.	III	64	31.3	66.8	0.0	29.7	70.3	9.1	6.0	34.2	51.6
Pittsburgh	12.03	Mifflenburg	III	35	25.7	54.2	0.0	100.0	0.0	10.1	5.3	50.0	54.3
				899	24.9	59.0	11.3	30.0	58.6	10.2	5.7	47.0	49.8

*These means are not always based on n's equal to the number of children because of missing data.
**This is hard to believe.

TABLE III-6-A
Characteristics of Planned Variations--Children in the Final Analysis Sample with Classroom as the Unit of Analysis

SPONSOR	CODE	SITE	Test Level	# of classes	% pre-sch.	mean age	% Mex.-Amer.	% White	% Black	mean mother's educ.	mean household size	[illegible]come (in hun-dred)	% [illegible]
Nimnicht	02.04	Duluth	III	4	10.6	56.2	0.0	87.1	12.9	10.0	5.5	50.3	51.3
	02.13	Tacoma	II	4	28.7	55.3	1.8	49.8	48.5	11.0	5.5	43.2	50.2
Tucson	03.08	LaFayette	III	4	70.6	64.6	0.0	77.2	22.8	12.4	3.9	36.9	39.4
	03.16	Lincoln	III	4	2.9	55.0	6.0	84.5	9.5	11.2	5.5	46.1	39.9
Bank Street	05.01	Boulder	III	4	0.0	55.3	40.7	57.6	1.8	10.5	5.3	38.9	50.8
	05.11	Wilmington	II	4	10.1	52.5	0.0	1.8	98.2	10.4	4.9	34.3	48.3
	05.12	Elmira	III	3	73.5	53.6	0.0	50.7	49.3	11.4	5.2	42.8	45.6
Becker & Engle-mann	07.03	E.St.Louis	III	4	0.0	53.8	0.0	0.0	100.0	11.9	4.9	62.0	51.0
	07.11	Tupelo	III	4	27.3	64.6	0.0	28.3	71.7	9.4	6.1	31.3	54.1
	07.14	E.LasVegas, NM	II	4	1.9	64.4	91.7	8.3	0.0	9.6	6.2	31.1	43.2
Bushell	08.04	Portageville	III	4	3.1	55 4	0.0	58.4	41.6	9.7	5.7	34.8	55.8
	08.08	Mounds, Ill.	II	4	7.6	54.7	0.0	32.6	67.4	10.6	5.8	45.3	48.2
Weikart	09.02	Ft.Walton Beach	III	4	0.0	53.0	0.0	23.8	76.2	10.1	6.4	33.0	49.2
	09.06	Greeley	III	4	9.1	56.1	87.0	13.0	0.0	9.3	5.0	36.3	37.7
	09.10	Seattle	II	4	12.2	54.3	0.0	41.8	58.2	11.5	4.7	37.0	55.7
Gordon	10.02	Jonesboro	III	3	0.0	67.1	0.0	64.1	35.9	8.7	6.1	31.5	38.5
	10.07	Chatta-nooga	III	4	0.0	65.9	0.0	25.0	75.0	9.9	5.6	26.9	53.5
	10.10	Houston	II	4	0.0	56.4	32.5	2.1	65.4	9.5	3.9	27.7	50.0
EDC	11.05	Wash'ton.	III	4	40.6	51.6	0.0	8.3	91.7	11.2	4.1	41.3	59.6
	11.06	Paterson	II	3	0.0	53.2	0.0	0.0	100.0	10.0	5.4	48.5	36.5
	11.08	JohnstonCo.	III	4	35.3	66.6	0.0	47.6	52.4	9.3	5.9	36.3	54.7
Pitts-burgh	12.03	Lock Haven	III	4	37.6	53.1	0.0	100.0	0.0	10.4	6.3	52.8	49.4
REC	23.01	Kans. City	III	4	10.6	54.4	42.8	21.7	35.5	9.8	6.1	43.4	45.7
Enablers	[illegible]	Billings	II	4	7.1	66.5	27.4	66.7	5.8	10.1	6.2	34.6	47.[illegible]
	[illegible]	Colo.Spgs.	II	4	3.7	55.8	43.1	18.0	38.9	10.4	5.6	36.2	51.6
	[illegible]	Bellows Fls.	II	4	7.7	58.3	0.0	100.0	0.0	10.5	5.0	39.7	52.3
				[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	39.3	48.3

TABLE III-7A.

Characteristics of Comparison Children in the Final Analysis Sample with Classroom as the Unit of Analysis.

SPONSOR	CODE	SITE	Test. Level	# of classes	% pre-sch.	mean age	% Mex.-Amer.	% White	% Black	mean mother's educ.	mean household size	[illegible]	% fe-male
Nimnicht	02.04	St. Cloud	III	2	8.6	57.0	0.0	100.0	0.0	10.7	6.9	47.5	41.3
Tucson	03.08	Albany	III	4	35.9	64.8	7.1	4.2	88.7	10.1	5.9	31.3	52.5
Bank Street	05.01	Boulder	III	1	25.0	53.7	25.0	41.7	33.3	11.3	5.9	42.2	50.0
	05.11	Del.Mar	II	4	2.1	52.2	0.0	38.9	61.1	10.9	5.2	39.4	53.1
	05.12	Elmira	III	3	0.0	54.0	0.0	71.1	28.9	10.2	5.3	63.4	55.0
Becker & Engle-mann	07.03	E.St.Louis	III	4	0.0	53.9	0.0	5.1	94.9	12.4	5.3	63.8	51.1
	07.11	Tupelo	III	4	62.9	65.0	0.0	29.1	70.9	9.2	6.3	29.5	49.4
	07.14	E.LasVegas, NM	II	4	7.7	64.7	100.0	0.0	0.0	10.9	6.0	36.1	57.6
Bushell	08.04	[illegible]	III	4	33.7	55.4	0.0	36.9	63.1	9.8	5.3	34.5	44.4
	08.08	Mounds, Ill.	II	2	57.6	54.8	0.0	82.6	17.4	10.5	5.6	53.4	59.7
Weikart	09.02	Pensacola	III	3	27.5	59.3	0.0	9.8	90.2	9.7	5.7	28.7	41.2
	09.06	Greeley	III	3	5.3	55.2	50.9	49.1	0.0	9.7	5.6	55.6	53.8
	09.10	Seattle	II	3	71.4	53.5	0.0	15.3	84.7	12.9	3.6	82.0	49.0
Gordon	10.02	Jonesboro	III	3	6.3	67.7	0.0	77.1	22.9	8.9	6.8	29.1	47.2
	10.07	Chatta-nooga	III	4	21.4	66.4	0.0	0.0	100.0	11.1	5.9	30.5	50.7
	10.10	Houston	II	4	76.1	55.7	30.6	0.0	69.4	10.3	4.1	36.0	41.3
EDC	11.05	Wash'ton.	III	4	18.3	52.1	0.0	0.0	100.0	9.7	6.7	37.2	44.1
	11.06	Paterson	II	1	0.0	52.6	0.0	0.0	100.0	11.0	4.9	35.4	60.7
	11.08	JohnstonCo.	III	4	28.7	66.7	0.0	30.2	69.8	9.1	6.0	35.1	51.0
Pitts-burgh	12.03	Mifflen-burg	III	4	25.2	54.2	0.0	100.0	0.0	10.2	5.2	[illegible]	50.4
				65	26.7	58.5	11.2	31.6	57.2	10.3	5.6	39.2	50.1

are site level means of classroom means, computed by equally weighting each of the classrooms in a site. For those variables which are common to Tables III-4 and III-5, the means of classroom means will vary slightly from the means computed on individuals, since the number of children per classroom varies within sites. By and large, however, inspection of the two sets of tables suggests that the differences are small. A set of variables not common to Tables III-4 and III-5 are included in Tables III-6B and III-7B. These are variables which refer specifically to classrooms. In particular, we include here the percentage of white teachers, the percentage of certified teachers and the mean years of experience of the teachers in Head Start and of the teacher aides. Also included are a mean index of the classroom levels of implementation in February and May, 1971, as seen by the sponsor and a mean rating of the staff working conditions by the teachers. Finally, the administrative arrangement of the Center is included (whether the Center is administratively run by a CAP agency or by the public schools) and where the center is located (in a public school or CAP location).

Although these variables are certainly not sufficient to paint a complete picture of the various sites and classrooms, they should give the reader some feeling of the variations among sites on a number of classroom-relevant variables.

Table III-10

Mean Teacher and Site Characteristics for Planned Variation Classes in the Final Analysis Sample

MODEL	CODE	SITE	Testing Level	Years Teacher Experience in Head Start	Percent Teachers Certified	Percent Teachers White	Average Staff Working Conditions	Administration by CAP or Public School	Sponsor Rating in February	Sponsor Rating in May	Housing Arrangement CAP or PS	Teacher Aide Years in Head Start
mnicht	02.04	Duluth	III	4.33	0	100	1.75	PS	5.3	6.0	PS	1.67
	02.13	Tacoma	II	2.75	0	100	1.71	PS	7.0	7.0	PS	2.50
Tucson	03.08	LaFayette	III	4.25	100	50	2.54	PS	NA	5.75	PS	3.00
	03.16	Lincoln	III	1.50	0	100	2.33	PS	NA	6.75	PS	1.00
ink Street	05.01	Boulder	III	2.75	75	67	2.63	CAP	5.5	6.25	CAP	2.75
	05.11	Wilmington	II	2.25	0	50	2.46	CAP	6.75	6.25	CAP	4.75
	05.12	Elmira	III	2.33	100	100	2.45	CAP	6.0	6.0	CAP	3.67
ecker & Engel-mann	07.03	E.St.Louis	III	6.00	100	0.0	2.0	CAP	5.0	5.5	CAP	NA
	07.11	Tupelo	III	3.00	75	50	2.48	CAP	7.75	7.0	CAP	3.67
	07.14	E.LasVegas NM	II	5.25	100	33	2.96	PS	5.5	6.75	CAP	2.50
ushell	08.04	Portageville	III	3.00	100	75	3.17	CAP	NA	NA	CAP	2.80
	08.08	Mounds, Ill.	II	5.50	100	50	1.83	CAP	NA	NA	CAP	4.25
Welkart	09.02	Ft.Walton Bch	III	2.25	75	25	2.42	CAP	6.0	5.5	CAP	1.50
	09.06	Greeley	III	3.25	25	75	1.96	PS	6.25	6.5	PS	2.25
	09.10	Seattle	II	5.25	25	25	2.81	CAP	5.25	5.5	PS	3.67
Gordon	10.02	Jonesboro	III	1.00	33	67	2.89	PS	6.0	6.33	PS	1.00
	10.07	Chattanooga	III	2.75	0	100	2.37	PS	7.5	7.25	PS	2.33
	10.10	Houston	II	3.75	75	0	2.38	CAP	3.0	4.25	CAP	1.25
EDC	11.05	Washington	III	2.00	100	0	2.30	CAP	NA	NA	CAP	4.50
	11.06	Paterson	II	5.50	100	0	1.92	CAP	NA	NA	CAP	1.67
	11.08	Johnston Co.	III	3.50	25	0	2.48	CAP	NA	NA	PS	4.00
Pitts-burgh	12.03	Lock Haven	III	2.00	25	67	2.13	CAP	5.75	6.0	CAP	1.50
EC	20.01	Kansas City	III	2.67	67	100	1.33	PS	8.0	7.5	PS	5.00
nablers	27.04	Billings	II	4.50	75	75	1.87	CAP	5.25	7.0	CAP	1.50
	27.05	Colorado Sp	II	4.00	50	67	2.56	CAP	5.67	6.0	CAP	2.50
	27.03	Bellows Falls	II	1.50	75	75	2.63	CAP	7.75	7.75	CAP	2.33

TABLE III - 16

Head Teacher and Site Characteristics for Comparison Classes in the Final Analysis Sample

Head Start Communities
1970-71

SPONSOR	CODE	SITE	Testing Level	Years Teacher Experience in Head Start	Percent Teachers Certified	Percent Teachers White	Average Staff Working Conditions	Administration by CAP or Public School	Housing Arrangement CAP or PS	Teacher Aide Years in Head Start
Nimnicht	02.04	St. Cloud	III	2.00	50	100	1.25	PS	PS	1.50
Henderson	03.28	Albany	III	4.25	100	0	2.04	PS	PS	3.67
Bank	05.01	Boulder	III	9.00	100	0	2.33	CAP	CAP	4.00
Street	05.11	DeLaWar	II	3.00	75	0	2.46	CAP	CAP	1.50
	05.12	Elmira	III	3.00	0	100	1.55	CAP	CAP	3.50
Becker &	07.03	E.St.Louis	III	3.67	100	0	2.28	CAP	CAP	2.00
Englemann	07.11	Tupelo	III	4.50	20	75	2.25	CAP	CAP	5.50
	07.11	W.LasVegas NM	II	3.00	67	33	2.17	PS	CAP	4.00
Bushell	08.04	Portageville	III	3.25	75	75	2.34	CAP	CAP	3.67
	08.08	Mounds, Ill.	II	4.50	50	0	2.83	CAP	CAP	3.50
Weikart	09.02	Pensacola	III	2.33	0	67	2.67	CAP	CAP	1.50
	09.06	Greeley	III	1.00	0	100	2.22	PS	PS	2.33
	09.10	Seattle	II	NA	33	33	2.61	NA	NA	NA
Gordon	10.02	Jonesboro	III	2.33	33	67	2.33	PS	PS	1.33
	10.07	Chattanooga	III	2.40	0	25	2.31	PS	PS	3.00
	10.10	Houston	II	4.67	75	0	2.73	CAP	CAP	1.67
EDC	11.05	Washington	III	3.00	50	25	2.83	CAP	CAP	4.00
	11.06	Paterson	II	5.00	100	0	3.50	CAP	CAP	2.00
	11.08	Johnston Co	III	2.25	75	50	2.04	CAP	PS	1.50
Pittsburgh	12.03	Mifflenburg	III	2.50	50	100	2.21	CAP	CAP	2.33

V Differences between the Planned Variations and Comparison Analysis Samples

Tables III-6 and III-7 indicate that the Planned Variations and Comparison analysis samples are, overall, essentially equivalent on Mean Age in the classroom, sexual composition of the classrooms, and on the three measures of family background (Income, Household Size and Mother's Education). There are, however, overall differences in the ethnic composition and the preschool experience of children in the two samples. Specifically, the PV sample has proportionally fewer Black and more white children as well as fewer children with pre-school experience. The differences between the PV and comparison group means are not fully eliminated when only paired Planned Variation and comparison sites are contrasted. Albany (0308), for example, is the comparison site paired in the design with LaFayette. The mean classroom percentage of Blacks in Albany is 88.7%, while the mean classroom percentage for LaFayette is only 22.8%. A rather large number of examples like this could be shown -- even for the variables which show no overall differences between the Planned Variation and Comparison analysis samples.

Another way of looking at the difference in inputs between the PV and comparison groups is to contrast the two groups of classrooms within models. Using the classroom as the unit of analysis within models and using only those sites which have both PV and comparison classes within them, we used a two way analysis of variance-- a model factor was crossed with PV/comparison. Table III-8 shows the amount of variation in the background and teacher characteristic variables attributable to models, PV/comparison, interaction, and within cells. Four things should be noted about this table. First, for all but two of the variables (Percent Mexican-American and Household Size) there are statistically significant model to model differences. This suggests that the composition of the sites within models differ rather radically from model to model. Second, there are no overall statistically significant differences between the PV and comparison groups. Taken as a whole for locations with both PV and comparison classes, the PV and comparison groups are remarkably similar on the variables described in these tables. Third, for only two variables are there statistically significant interactions between the models and the PV/comparison factor in the table. This suggests that the PV and comparison groups within models are remarkably similar. Fourth, it should be mentioned that most of the variation on each of these variables lies within cells. Model to model variation plus PV/comparison variation plus the variation attributed to interactions between the two main factors never accounts for more than 40% of the total variation and generally accounts for less than 30% of the total variation.

One implication of these findings is that insofar as these variables are important determinants of achievement, a matching of PV and comparison groups within models will go a long way

TABLE III-8

Percentages of Variation:

(1) Among models (both PV and comparison classrooms together);

(2) Between PV and comparison groups pooled across models;

(3) Due to interaction between models and PV/comparison groups;

(4) Within cells.

Classrooms are the unit of analysis. The design is as unweighted means crossed model by PV/comparison using only those sites in the analysis sample which have both a PV and a comparison group of classes. The sum of the four sources of variation for each variable is 100%.

	PERCENTAGES OF VARIATION			
Variable	(1) Among Models	(2) Bet. PV & Compar. Grps.	(3) Interaction	(4) Within Cells
Age	38.5*** df=8	0.18 df=1	0.29 df=8	61.03 df=124
Preschool Experience	15.24** df=8	0.50 df=1	15.29** df=8	68.96 df=124
Mexican-American	11.32 df=8	0.06 df=1	0.83 df=8	87.79 df=124
Black American	29.57*** df=8	0.66 df=1	5.69 df=8	64.07 df=124
Household Size	8.82 df=8	1.99 df=1	12.97* df=8	76.22 df=124
Income	18.90*** df=8	0.04 df=1	3.42 df=8	77.64 df=121
Mother's Education	4.17 df=8	0.05 df=1	3.78 df=8	92.00 df=115
El/Ek	48.58*** df=8	0.01 df=1	0.06 df=1	51.36 df=124
Pct. Females	15.24** df=8	0.50 df=1	15.29** df=8	68.96 df=124

Table III-8

(Cont'd)

	PERCENTAGES OF VARIATION			
Variable	(1) Among Models	(2) Bet. PV & Compar. Grps.	(3) Inter-action	(4) Within Cells
Teacher Headstart Experience	13.00* df=8	1.98 df=1	6.49 df=8	78.53 df=108
Teacher Certification	19.35** df=8	0.03 df=1	5.26 df=8	75.37 df=111
Staff Working Conditions	16.69** df=8	1.15 df=1	3.72 df=8	78.44 df=115
Teacher Aide Year in HS	15.66** df=8	0.15 df=1	4.44 df=8	79.75 df=100

* Statistically significant beyond the .05 level

** Statistically significant beyond the .01 level

*** Statistically significant beyond the .001 level towards equalizing the PV and comparison groups on important input factors. A second implication is that the composition of classrooms within models differs dramatically from model to model and therefore that even pooling sites within models will not make the models equivalent on these variables.

VI Pre-Test Score Differences Between the Planned Variations and Comparison Samples

Still another way of looking at initial differences between the PV and Comparison groups is to directly contrast the two groups on their pre-test scores. In order to give the reader a feel for the pre-test data we carried out these comparisons in a number of ways. First, we show the overall mean differences in pre-test scores and their variances for the two groups. Second, we divide the children into twelve groups (by ethnicity, preschool experience and entering level) for each sample and present mean and variance differences for each of the twelve groups. Third, we move to the classroom level and show overall means and variances for the two groups on each of the tests. Fourth, we present the results of regression analyses using the pre-test scores as dependent variables, with a PV/comparison group dummy variable and a series of background characteristics as independent variables. Fifth and finally, we present results from a multivariate analysis of variance with three pre-test scores as dependent variables, and a series of background variables as covariates, for a design with PV/comparison crossed with models. The overall conclusion from these *analyses is that after the introduction of a few controls there are almost no differences between the PV and comparison groups on the pre-test variables.*

Table III-9 below shows the simple contrast between the means and variances for the overall PV and comparison groups for five pre-test variables. Table III-9 indicates that there are a few significant differences between the overall PV and comparison groups on pre-test means and variances. Specifically, three of the ten statistical tests revealed differences at the .05 level. There are no significant differences on the PSI, the Stanford-Binet and the Motor Inhibition tests. The variances for the PV and Comparison groups on the Book 3D and the Book 4A tests are statistically different, with the comparison group each time having the largest variance. It must be noted that while the variances for these two tests are significantly different, the ratios of the two sets of variances are very small -- the large number of degrees of freedom made the statistical tests very sensitive to small differences. Finally there is a statistically significant difference between the overall PV and comparison means on the Book 4A favoring the comparison group. Again, however, the difference is small (roughly one tenth of the pooled standard deviation) which indicates the power of a large sample in detecting small differences. We earlier pointed out clear overall differences between the samples in input characteristics which might be causing these differences -- specifically in the overall percentages of children who are in their first and second

Table III-9

Differences between the PV and Comparison group analysis sample in means and variances of 5 pretest scores. Children are the unit of analysis. Only children with a valid pre- and post-test on the particular variable being compared were used in the analysis.

	Test Variable				
	Book 3D	Book 4A	PSI	Stanford-Binet	Motor Inhibition
PV N	1188	1178	1197	389	465
Comparison N	805	803	806	297	300
PV mean	11.851	5.548	35.498	90.511	5.047
Comparison mean	12.103	5.889	35.835	90.042	5.116
Difference between PV and Comparison mean	-0.252	-0.341*	-0.337	0.469	-0.069
PV variance	9.746	10.063	151.221	184.289	0.294
Comparison variance	11.092	11.403	142.014	176.839	0.274
Ratio of PV and[F] Comparison variance	1.1381*	1.1331*	1.0648	1.0421	1.0729

* Statistically significant beyond the .05 level

[F] The largest variance was the numerator for this test.

year of pre-school. We might imagine that this difference (which favored the comparison children) could easily explain the few differences in pre-scores that we see on the uncontrolled pooled samples.

In a second set of analyses we disaggregated the children in the classrooms for both samples and divided each sample into twelve groups. Our stratification procedure took two levels of prior preschool experience (no and yes), three ethnic categories (Mexican-American, Caucasian and Black) and two levels of entering grade (E1 and EK). We then separately compared the Planned Variation and Comparison samples for each of ten groups (two groups were left out due to sample sizes less than ten) on means for all four of the cognitive tests and the Motor Inhibition test (see Table III-10). This gave us a total of 50 independent comparisons. We found only 4 comparisons to be statistically significant beyond the .05 significance level -- none were significant beyond the .01 level. Specifically we found that (1) white children with no preschool experienc in Entering Kindergarten classes in Planned Variations scored higher than their corresponding comparison group with Stanford-Binet; (2) Black children with no preschool experience in Entering Kindergarten classes in the comparison sample scored higher than the corresponding Planned Variations group in both the Book 3D and the Stanford-Binet tests; (3) Black children with preschool experience in Entering first classes in the comparison group scored higher than the corresponding Planned Variations group on

Table III-10

Pre-Test Mean Differences and Variance Ratios for 12 Groups of PV and Comparison Children on 5 Pre-Tests[F]

	Group		Tests									
Ethnicity	Pre-Sch. Exp.	Ent. Gr.	BX4A		BK3D		PSI		SB		MI	
			Mean Diff.	Var. Ratio	Mean Diff.	Var. Ratio	Mean Diff.	Var. Ratio	Mean Diff.	Var. Ratio	Mean Diff.	Var. Rati
Mex.-Amer.	no	E1	0.058 df=109	-1.21	0.522 df=109	-1.04	1.291 df=109	-1.20	---	---	-0.036 df=61	-1.50
Mex.-Amer.	no	EK	-0.064 df=99	1.43	0.803 df=104	1.14	-0.389 df=104	1.78	-2.193 df=38	1.17	0.180 df=20	1.06
White	no	E1	-0.429 df=156	1.12	-0.068 df=156	-1.02	4.207* df=155	-1.17	-3.518 df=63	1.48	0.144 df=101	1.13
White	no	EK	-0.129 df=416	-1.06	-0.209 df=420	-1.05	1.772 df=424	-1.21	1.872 df=145	-1.49*	-0.006 df=167	1.08
Black	no	E1	-0.699 df=248	1.06	-0.290 df=248	1.07	0.551 df=250	1.08	0.578 df=127	1.48	-0.096 df=114	1.63*
Black	no	EK	-0.088 df=554	-1.04	-0.657* df=555	1.19	-0.629 df=559	1.13	-4.193* df=162	-1.04	-0.178 df=134	-1.27
Mex.-Amer.	yes	E1	---	---	---	---	---	---	---	---	---	---
Mex.-Amer.	yes	EK	---	---	---	---	---	---	---	---	---	---
White	yes	E1	-1.832 df=67	-1.08	-0.468 df=67	1.07	-1.037 df=67	-1.13	-5.382 df=29	-1.85	-0.105 df=45	1.50
White	yes	EK	-1.770 df=72	-1.26	-0.792 df=73	1.04	-3.872 df=73	1.94*	8.179 df=25	1.01	0.001 df=31	-1.75
Black	yes	E1	-1.711* df=109	-1.24	0.532 df=109	1.29	-1.437 df=109	-1.05	0.623 df=46	-1.21	-0.118 df=48	1.19
Black	yes	EK	0.386 df=111	-1.19	-0.263 df=112	1.69*	-2.312 df=113	1.02	5.922 df=27	-4.81*	-0.295 df=14	-2.27

[F] All differences are expressed with the PV groups as positive and the Comparison group as negative Similarly, the variance ratios, while all greater than 1.00, are given a sign -- a positive sign indicates that the PV variance is larger and a negative sign that the Comparison group is larger.

* statistically significant at the .05 level

Book 4A. No differences in means were found for the Motor Inhibition test.

The large number of statistical tests carried out and the very few number of statistically significant results leads us to suspect that these differences might have occurred by chance. The fact, however, that two of the statistically significant findings occurred for one group and that the tendency for the other tests for these groups to go in the same direction for all three tests though the t's are less than one, suggests that there may be a difference between the comparisons and Planned Variation samples for Black children with no preschool experience in Entering Kindergarten classes. Overall, though, a general conclusion is that the Planned Variation and Comparison groups are essentially similar on pre-test mean scores.

We then compared the variances for the groups. Five of the fifty variance ratios were statistically significant at that .05 level, with two of the five having greater variances in the comparison group. There doesn't seem to be any pattern to the differences in variance. Book 4A is the only test for which there is not significant difference in variances. For Blacks with prior preschool and EK, where there are two statistically significant variance ratios, the Comparison group has the larger variation for one group and the smaller for the other. Since no patterns exist and there are only five significant differences, none reaching a higher level of significance than .05, our conclusion about differences among variances is the same as our conclusion about differences among means -- that the PV and Comparison groups are overall very similar.

The third approach was to contrast classroom means and variances for the overall PV and Comparison samples. These data are shown in Table III --11. There are no statistically significant differences between the means and variances of the overall PV and Comparison groups on any of the five pre-test variables.

Although no differences exist between classroom means in an uncontrolled state, there may be differences after controlling for some relevant variables. This situation was tested in the final two analyses presented here. In the first analysis, presented in Table III-12,we ran a set of five regressions with the five pre-test variables as the dependent variables. In each equation the variable of principal interest was a PV/Comparison dummy variable. There were 7 major control variables -- Mean Classroom Age, Percent Mexican-American, Percent Black, Percent with Previous Preschool Experience, Mean Income, Mean Household Size and Mean Mother's Education. In no equation did the dummy variable representing membership in the PV or comparison group enter with a statistically significant coefficient.*

The fifth and final attempt to see whether there are sig-

* We also ran these equations letting there be different slopes for the control variables for the PV and comparison groups. There were no serious differences between the presented results and the results of these runs. In particular, in no instance was the dummy PV/comparison variable statistically significant.

Table III-11

Differences between the PV and Comparison total analysis sample in overall means and variance on each of five pre-tests. Classrooms are the unit of analysis. Only classrooms with valid pre- and post-test scores on the particular variable being compared are used in the analysis.[F]

	Test Variable				
	Book 3D	Book 4A	PSI	Stanford-Binet	Motor Inhibition
PV N	101	101	101	61	87
Comparison N	65	65	65	47	59
PV Mean	11.792	5.572	35.081	90.591	5.002
Comparison mean	11.993	5.788	35.384	90.299	5.060
Diff. between PV and comparison mean	-0.201	-0.216	-0.303	0.292	-0.058
PV variance	2.981	2.263	54.194	66.359	0.149
Comparison variance	2.436	2.534	47.484	54.466	0.184
Ratio of PV and comparison variance	1.2237	1.1197	1.1413	1.2183	1.2348

[F] There are no statistically significant differences in mean or variance between the PV and Comparison groups.

Table III-12

Standardized and raw regression coefficients for a PV/Comparison group membership dummy variable for five pre-tests. The overall classroom means analysis sample was used. Seven control variables were also in the equation (Percent Previous Preschool, Percent Mexican-American, Percent Black, Mean Age, Mean Income, Mean Household Size and Mean Mother's Education).[F]

Test	N's		Zero-order r between PV/comparison and test variable	PV/Comparison Regression Coefficients		Standard error of raw coefficient	Percentage of variance explained in total equation
	PV	Comparison		Standardized	Raw		
PSI	101	65	-.021	.058	0.8516	0.900	46.78***
Book 3D	101	65	-.059	.002	0.0066	0.253	32.62***
Book 4A	101	65	-.068	-.047	-0.1477	0.245	14.59**
S-B	61	47	.019	-.091	-1.4316	1.421	28.08***
Motor Inhibition	87	59	-.070	-.028	-0.0251	0.069	13.69**

[F] The PV/Comparison dummy variable is coded 1/0 where 1 stands for membership in the PV group.

*** Statistically significant beyond the .001 level.

** Statistically significant beyond the .01 level.

nificant differences between the PV and comparison samples utilized a multivariate analysis of variance approach. We used four dependent variables (the Stanford-Binet, PSI, Book 3D and Book 4A). These tests were chosen because they were used in all Level III tested sites, most of which also had comparison classes. Classrooms were the unit of analysis. The design was crossed with PV/Comparison as one factor and models as a second factor. A variety of [illegible] (see Table III-13) were used. Only locations with both PV and Comparison classes were included in the analysis. We are primarily interested in the overall test of the PV/Comparison factor and whether there are model by PV/-Comparison interactions. The method used was an unbalanced exact least squares solution. Table III-13 presents the estimated combined means for the analysis, F-ratios for the overall multivariate tests of significance and F-ratios for the univariated tests.

Table III-13 clearly shows the similarity between the PV and comparison groups overall (all univariate F's are non-significant as is the multivariate F for this factor), and the similarity between the PV and comparison groups within sites (the univariate [illegible] F's for the interaction term are all non-significant as is the multivariate interaction F). It also clearly shows the strong model to model differences (all univariate F's but the [illegible] are statistically significant, as is the multivariate F for model to model differences. These findings would suggest that it may be difficult to [illegible]

Table III-[illegible]

Estimated Combined Means for the Stanford-Binet, PPVT, Deep 50, Deep [illegible] as Pre-Tests for [illegible] and for Participation groups. Also shown are F-ratios for the overall multivariate test of significance between [illegible], Participation and the Interaction as well as the univariate F's. Classrooms were the unit of analysis. Covariates used were Mean Age, Percent Black, Percent Mexican-American, Percent Female, Percent Prior Preschool, Mean Income and Mean Mother's Education.

Estimated Combined Means for Pre-Tests

	Stanford-Binet		PPVT		DEEP		DEEA	
	PV	Comp	PV	Comp	PV	Comp	PV	Comp
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12.55	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	12.19	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	5.28	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.84	11.55	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	10.08	[illegible]	[illegible]
Grand Mean Unweighted	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	11.08	[illegible]	[illegible]
N's	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
Univariate F's [illegible]	[illegible]		[illegible]		[illegible]		[illegible]	
Univariate F's [illegible]	[illegible]		[illegible]***		[illegible]***		[illegible]**	
Univariate F's [illegible]	[illegible]		[illegible]		[illegible]		[illegible]	

(next page)

Table III-12 (cont'd)

b) Multivariate F for model to model differences = 2.99 with 12 and 241.53 df, p < .001

c) Multivariate F for P/Comparison differences = 0.5012 with 4 and 71 df, p > 0.75

d) Multivariate F for Interaction = 1.1285 with 32 and 261.87 df, p > 0.30

* Statistically significant at .05 level

*** Statistically significant at .001 level model to model differences but that it looks as though PV/Comparison differences within models may be easily eliminated with appropriate controls.

For the Stanford-Binet, however, inspection of the table of combined means reveals that there are some fairly large within model differences between PV and Comparison groups -- not enough to bring the overall F to a statistically significant level but enough to worry about.

Taking all of these analyses together implies a variety of things about analyses in later chapters. First, on the basis of present [illegible] it looks as though we are on quite safe grounds in contrasting overall PV and comparison groups since their pre-test means differ little, as do their pre-test variances. What [illegible] differences do show up seem to be essentially [illegible] by the simple introduction of linear controls. It appears as if it is more appropriate to contrast the overall PV and Comparison groups for the sites where we have both PV and Comparison [illegible], but only slightly more than comparing them without [illegible].

Second, [illegible] PV models are [illegible] the models appear to differ on [illegible] variables. We do not know, of course, whether [illegible] for [illegible] we have no way of [illegible].

comparisons, we will [illegible] to the fact that there appear to be [illegible] aggregate differences among the FV models and that the differences appear to be of some importance. This is unfortunate because [illegible] that we will have to rely heavily on [illegible] contrasts among the models. In this strategy we [illegible] model [illegible] FV classes with model A Comparison classes and model [illegible] FV classes with model B Comparison classes [illegible] the differences in the results to infer differences among the models.

Then, [illegible] FV [illegible] models, even this strategy fails since they have no comparison classes. In these instances we either have to rely [illegible] on direct contrasts among models or use a third analytical strategy. Our approach is to do both. The third analytical [illegible] taken in later chapters is matching. Briefly we will be matching FV classes from all models with [illegible] comparison classes. Then FV-Comparison [illegible] on the matched samples. This approach has the [illegible] of [illegible] out [illegible] problem [illegible] comparison. [illegible] matching without [illegible] functional form. [illegible] possibility [illegible] [illegible] in this chapter [illegible] means and [illegible] comparison groups [illegible] this problem. In [illegible] we will [illegible] further [illegible] problem [illegible] and on the [illegible] of the [illegible] sample.

Chapter IV

[illegible] THE HIGH STREET [illegible]

In [illegible], we [illegible] the changes in students' [illegible] of the High Street program and [illegible] attributed to students' [illegible] program other than [illegible]. We will not [illegible]. The importance of [illegible] overall value of the impact of the [illegible]. In later chapters, where the [illegible] of the [illegible]. In this chapter, we will [illegible] narrowly defined by the [illegible]

The [illegible]. In the first we present summary [illegible] for the overall [illegible] the [illegible] students [illegible] of Chapter III. We then [illegible] in [illegible] way by [illegible] [illegible] program [illegible] different [illegible] In the [illegible] [illegible] not attended Head Start. We then use these estimates to estimate the amount of change attributable to the Head Start experience.

I. Overall [illegible] changes in Head Start children during the 1970-71 preschool year.

Table IV-1 presents summary changes for the entire group of tested children in the 1970-71 Head Start analysis sample and by 12 subgroups of children. The 12 groups in Table IV-1 were formed by [illegible] the two categories of prior preschool experience (yes/no), by three categories of ethnicity (Mexican American, whites, and blacks), and by two categories of the entering public school grade [illegible]. The changes are presented for five measures: PSI, Book 3D, Book 4A, Motor Inhibition, and Stanford-Binet.

Although these data are primarily descriptive, the statistical significance of the changes are indicated for heuristic purposes. [illegible] in most cases the changes are very large—often approaching a standard deviation—suggesting that practical significance [illegible] for the [illegible]. In particular for the PSI and Book 4A tests, the average gains are [illegible] standard deviations. For the Motor Inhibition [illegible] standard deviations. For the Book 3D, the overall average [illegible] standard deviations and for the Stanford-Binet, [illegible] of a standard deviation. A number of patterns emerge from these tables.

First, [illegible] as much as children [illegible] prior preschool experience. [illegible]

TABLE IV-1

Overall Average Changes in Mean Test Scores for Children in the 1970-71 Final Analysis Sample

2,2[illegible] children are represented in the table. Each cell contains the mean gain and the number of children in the group. (Blank cells indicate insufficient N to estimate mean).

Groups

Group No.	Ethnicity	Prior Pre-School Exper.	Entering Grade	PSI Gains	Book 3D Gains	Book 4A Gains	MI Gains	Stanford Binet Gains
1	Mexican American	No	Fl 1	x=12.[illegible] n=111	2.6 111	4.0 111	0.3[illegible] 6[illegible]	—
2	Mexican American	No	Fk 2	9.7 106	1.7 106	2.9 101	0.46[illegible] 71	7.[illegible] 40
3	White	No	Fl 3	10.[illegible] 1[illegible]	2.[illegible] 15[illegible]	5.8 158	0.2[illegible] 10[illegible]	7.[illegible] 6[illegible]
4	White	No	Fk 4	12.[illegible] 4[illegible]	2.7 42[illegible]	3.7 418	0.37[illegible] 168	[illegible] 147
5	Black	No	Fl 5	11.[illegible] 2[illegible]	2.8 255	6.0 250	0.36[illegible] 115	2.[illegible] 12[illegible]
6	Black	No	Fk 6	12.8 [illegible]	2.1 557	2.6 556	0.4[illegible] 17[illegible]	6.[illegible] 16[illegible]
7	Mexican American	Yes	Fl 7	—	—	—		—
8	Mexican American	Yes	Fk 8	—	—	—		—
9	White	Yes	Fl 9	7.1 65	1.9 65	5.4 65	0.0[illegible]* 3[illegible]	1.[illegible] 32
10	White	Yes	Fk 10	7.9 72	1.8 72	2.7 74	0.4[illegible] 32	2.[illegible] * [illegible]
11	Black	Yes	Fl 11	10.[illegible] 111	2.5 111	5.1 111	0.[illegible] 3[illegible]	0.[illegible] * 4[illegible]
12	Black	Yes	Fk 12	[illegible] 11[illegible]	2.2 114	2.[illegible] 111	0.[illegible] 1[illegible]	[illegible] [illegible]
	Total			x=11.[illegible] n=2,[illegible]	2.4 1,5[illegible]	[illegible] 1,5[illegible]	0.[illegible] [illegible]	4.[illegible] [illegible]
			SD_{gain}	=[illegible]	[illegible].9	4.[illegible]	0.[illegible]	11.[illegible]
			SD_{pre}	=1[illegible]	[illegible]	[illegible]	0.[illegible]3	11.9
			SD_{post}	=10.[illegible]	[illegible]	4.[illegible]	[illegible]	11.[illegible]

*[illegible] not statistically significant beyond the [illegible] level

for just about all tests and all sub-groups.

Second, children who enter first grade directly from preschool gain more on the Book 4A than children who will enter kindergarten. There is some tendency for EI children to show more change on the PPVT, Book 3B and Stanford-Binet while EK children show more on the Motor Inhibition. The larger difference on the Book 4A, a test which assesses knowledge of letters, numerals and shape names may be explained by at least two factors both of which probably contribute to the effect. On the one hand, children in the EI group are older and thus may be more ready to handle the academically oriented instruction offered in preschools. On the other hand, teachers in EI groups may feel a greater responsibility to prepare their children for first grade. Consequently they may spend more time on the readiness skills deemed necessary for first grade while the teachers in EK groups might rely more on the public school kindergarten teachers to teach these children the readiness skills.

Third, there is [illegible] consistent pattern of differential change among [illegible] groups.

[illegible] the Stanford-Binet [illegible] [illegible] for the other groups [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible]

[illegible] therefore, indicates that there has been an increase in the rate of change in mental [illegible] which might be attributed to the [illegible].

Fifth, [illegible] on the total sample the variance of the pre-test is significantly different from the variance of the post-test. In three of the [illegible] the [illegible] variance is significantly smaller than the pre-test variance. Only for the [illegible] is there a significant increase in the [illegible] pre to post-test. In the other [illegible] the variances are not significantly different [illegible], the post-test variance is [illegible] smaller than the pre-test variance.

These [illegible] directly to an hypothesis [illegible] by [illegible] in the [illegible] field -- the fan spread hypothesis. [illegible] hypothesis the variance from pre- to post-test [illegible] increase [illegible] evidence [illegible] initially high [illegible] children. [illegible] the [illegible] data. [illegible] a divergence [illegible] from pre to post [illegible] initially low [illegible] children [illegible] initially high.

scoring children. The one instance where this does not hold (Book 4B) may be an anomaly explained by a pre-test floor effect (see "The Quality of the Data").

Another way of expressing change is in establishing criterion referenced standards. The Book 4A test has two subtests which are amenable to this. Subtest 1 for Book 4A ~~assesses~~ knowledge of letters (7 items) and Subtest 2 has 6 items on the knowledge of numerals. We designated the arbitrary criterion that children have mastery over letters and numerals if they answered either all or all but one of the items correctly. Table II-4 shows the percentage of children in each of the twelve schools reaching these criterion for both the Fall and Spring testings. Data for both Book 4A and [illegible] are shown. Again, there are large overall gains particularly for [illegible] children. Roughly one-eighth of all children achieve "mastery" on the knowledge of letters criterion and [illegible] reach mastery on the knowledge of numerals.

We [illegible] for the three Book [illegible] mathematics results [illegible] (7 items), and [illegible] For each subtest, the "mastery" criterion is all or all but one item correct. [illegible] and [illegible] children overall, and finally [illegible] on the [illegible] the Book 4B subtests. Roughly

TABLE [illegible]

[illegible] Final Analysis Sample [illegible] of Book 4A. Columns 1 and 2 show for the pre and posttest the percentage of children answering correctly 8 or 9 out of 9 items [illegible] Columns 3 and 4 show for the pre and posttest the percentage of children answering correctly [illegible] items assessing knowledge [illegible] indicate insufficient N's to [illegible]

[illegible]	[illegible]	[illegible]	1 [illegible]	2 [illegible]	3 Numerals pre	4 Numerals post	N
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.2	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	28.[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	21.4	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	16.8	[illegible]
[illegible]			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

TABLE [illegible]

[illegible] of children in 1970-71 [illegible] Analysis Sample "[illegible]" on three subtests of Pack 3D. Columns 1 and 2 show for the pre and post-test the percentages of children answering correctly 6 or 7 out of the 7 item [illegible] readiness subtest. Columns 3 and 4 show pre and post-test percentages of children answering 6 or 7 items correctly out of the 7 item Science subtest. Columns 5 and 6 show the percentage of children in the pre and post-test answering 4 or 5 items correctly out of the 5 item proposition subtest (Blank [illegible] indicate insufficient children to estimate percentages)

[illegible]	[illegible]	[illegible]	[illegible]	1 Math Pre	2 Math Post	3 Science Pre	4 Science Post	5 Propos. Pre	6 Propos. Post	N's
	[illegible]	[illegible]	[illegible]	79.7%	64.9	29.6	52.3	22.5	44.1	111
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
[illegible]	White	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	158
[illegible]	White	[illegible]	[illegible]	[illegible]	48.[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	422
[illegible]	Black	[illegible]	[illegible]	[illegible]	[illegible]	44.4	70.2	34.8	67.2	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	44.9	22.6	43.1	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	----	----	----	----	----	----	----
[illegible]	[illegible]	[illegible]	[illegible]	----	----	----	----	----	----	----
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	85.1	69
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	58.[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	48.[illegible]	[illegible]	[illegible]
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]
	[illegible]			[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

one-half of the children have reached criterion on each of the subtests at post-test time. For the entering first grade children who have attended preschool, the percentage reaching criterion is closer to [illegible] on the average though it must be noted that the percentages for these children are initially quite a bit higher than for the other children. Once again, there is a clear pattern of greater gains for the EI children with a less consistent pattern of greater changes for the children who have not had a prior preschool experience.

The use of percentages of children reaching criterion is an exploratory and descriptive way of presenting the data gathered in this study. By and large, its utility rests on the adequacy of the criteria used and the feel for the data that it gives to the reader. There are a number of statistical tests that we might have applied to these data but since our principal purpose was descriptive and our criteria arbitrary, we felt that applying inferential statistics might not be justified.

II. [illegible]

That [illegible] to the Head Start [illegible] experienced [illegible] expected [illegible] as intended [illegible] received and [illegible] one-half of the children have reached criterion on each of the subtests at post-test time. For the entering first grade children who have attended preschool, the percentage reaching criterion is closer to 75% on the average though it must be noted that the percentages for these children are initially quite a bit higher than for the other children. Once again, there is a clear pattern of greater gains for the El children with a less consistent pattern of greater changes for the children who have not had a prior preschool experience.

The use of percentages of children reaching criterion is an exploratory and descriptive way of presenting the data gathered in this study. By and large, its utility rests on the adequacy of the criterion and on the feel for the data that it gives to the reader. There are a number of statistical tests that we might have applied to these data but since our principal purpose was descriptive and our criteriia arbitrary, we felt that applying inferential statistics might not be justified.

II. Estimated Effects of the Head Start Experience

This section attempts to estimate changes attributable to the Head Start experience. In Table IV-4 "gains" are expressed in terms of a difference between an actual and an expected score on the particular test. The "expected" score is intended to represent the score the child would have received had he not attended Head Start. Since an appropriate

TABLE IV-4

Overall Gain for children in the Planned Variation Study in 1970-71. "Gains" are computed by subtracting an "Expected post score" from an "Observed Post Score." (Blank cells indicate insufficient numbers of children to estimate mean "gains.")

Group	Ethnicity	Prior pre-school exp.	Entering Grade	PSI "Gain"	Book 3D "Gain"	Book 4A	Motor Inhib.	Stanford-Binet
1	M-A	No	El	Gain= 6.7 n = 111	1.8 111	2.6 111	0.14* 63	---
2	M-A	No	Ek	0.40* 106	0.2* 106	1.9 101	0.19 22	9.3 40
3	Whte.	No	El	4.7 157	1.1 158	4.7 158	0.14 103	8.0 65
4	Whte.	No	Ek	4.0 426	1.1 422	1.9 418	0.28 169	7.0 147
5	Blk.	No	El	8.4 252	2.4 250	4.8 250	0.40 116	9.6 129
6	Blk.	No	Ek	6.4 561	1.0 557	1.9 556	0.38 136	10.4 164
7	M-A	Yes	El	---	---	---	---	---
8	M-A	Yes	Ek	---	---	---	---	---
9	Whte.	Yes	El	7.1 69	0.4* 69	3.4 69	0.05* 47	12.2 31
10	Whte.	Yes	Ek	8.8 75	2.4 75	2.7 74	0.47 33	14.5 27
11	Blk.	Yes	El	7.3 111	1.7 111	4.8 111	0.03* 50	8.4 48
12	Blk.	Yes	Ek	-0.7* 115	-0.2* 114	1.1 113	0.47 16	-2.0* 29
			TOTAL	5.4 2003 SD_{gain}=9.2	1.2 1993 2.8	2.7 1981 4.5	0.26 765 0.56	8.7 686 12.2

* Indicates gain not statistically significantly greater than zero beyond the .05 level.

control group did not exist in the study -- that is, a group of children who did not attend Head Start -- we had to estimate expected maturational growth by using age variations in the children in the study. Briefly, the procedure was as follows. First, the children were divided into the 12 groups represented in the tables. Second, the pre-test scores of these children on all of the five test variables were used as dependent variables in 60 separate regression equations (1 for each of the 5 pre-test variables for each of the 12 groups). This gave us the relationships among the test score and the independent variables prior to the time the children entered Head Start. The independent variables for each analysis were age, sex, family income, household size, mothers' education and appropriate dummy variables to control for missing data. By using the coefficients for these equations and the original data, we arrived at an "expected" pre-score for each child. Within each of the 12 sub-groups, the mean of the expected pre-scores equals the mean of the actual pre-scores.

The regression analysis estimates the effect of age with controls for the three stratifying variables and their interactions as well as for the other variables in the regression equations (sex and family background). Therefore, with respect to the relationship between age and the score on a particular test, it can be argued that the coefficient for age (expressed as the average change per month in a group) in each equation reflects the rate of growth for the children in each group prior to their entering Head Start. (The assumptions for this argument are discussed in later paragraphs). In other words, the average expected difference between a child at 48 months and at 56 months without Head Start is reflected in the coefficient for age for his particular group. We then can estimate what we would expect a child's score on a test to be 7 months after entering a Head Start program *if* the program has no effect at all. This assumes that the relationship between the test scores and the independent variables remains during the Head Start program as it was prior to the Head Start experience.

The analysis required two more steps: estimating an expected post-test score and finding the difference between the expected and observed post-test scores. First, for each child in each group on all five tests we estimated an expected post-test score by adding to his pre-test score the product of the number of months he was in the program and the age coefficient for his group. This expected post-test score reflects an estimate of the child's change assuming *no* Head Start effect. We then subtracted this expected post-test score from his observed post-test score and computed group means. The mean differences between expected and observed post-test scores are then interpreted as the effect of Head Start above and beyond the effect expected by maturation alone.

In somewhat more precise terms the procedure was as follows: (1) Divide the children into the 12 groups. Consider now one group (Black children with no prior preschool experience who will enter public school kindergarten) and one test (the PSI). The procedure for this group and test was the same as for all other tests and groups. (2) A simple linear regression using PSI pre-score as the dependent variable on age, ~~sex~~, size of household, income of family, mothers' education, a dummy variable for missing cases on mothers' education, a dummy variable for missing cases on income, and a dummy variable for missing cases on household size. Complete data were available for age, sex and PSI pre-score. Following Cohen (1970), the other independent variables were given their observed values unless there was no information for the child on a variable. In this instance, the variable was given the value of zero. Dummy variables were then computed for each of the three variables with missing values. The dummy variables were coded with a 1 if the data were missing and a zero if the data were present in the original variable. Thus, if a child had a value for mothers' education, he would be assigned that value on his variable 'mothers' education' and a zero on the 'mothers' education'dummy variable. If, on the other hand, he had no observed value for mothers' education he would be assigned a zero on the mothers' education variable and a one on the mothers' education dummy variable. For the group and test we are considering, the final equation was as follows:

PSI pre = b_0 + B_{age}*Age + b_{sex}*Sex + b_{HHsize}*HHsize + b_I*Income = b_{ME}*Mothers' Education + b_{MED}*Mothers' Education Dummy + b_{ID}*Income dummy + b_{HHD}*Household Size Dummy

or the "expected PSI pre-score" for a child equals a constant (b_0) plus a coefficient for age (b_{age}) times the child's age plus a coefficient for sex (b_{sex}) times the child's sex (1=male/2=female) plus a coefficient for size of household (b_{HHsize}) times the number of persons in the child's household plus a coefficient for the child's family income (b_I) times the family income etc. The coefficients for the group of 620 children were:

b_0 = -32.606	b_{age} = 0.7652***	b_{sex} = 1.9911*
b_{HHsize} = 0.0119	b_{Income} = 0.0738***	b_{ME} = 1.4338***
b_{MED} = 18.2855***	b_{ID} = -1.2671	b_{HHD} = -0.8753

One * indicates statistical significance at the .05 level; Three *'s indicates statistical significance at the .001 level.

The equations were run on all children in the pre-test analysis sample within a group. The key to the generation of an expected post-test score for a child is in the coefficient for age (here it is 0.7652 and is statistically significant beyond the .001 level). The interpretation of this coefficient is that on the average for this group, children's scores increase by 0.7652 points for every month of age. In other words, in this sample, children who are 60 months old; score 5 x 0.7652 or 3.826 points higher than children who are 55 months old. If we assume that other things are equal then a child's score would increase naturally over the period of time that he is attending preschool -- specifically, it would increase naturally 0.7652 points per month while he is attending Head Start.

Granting the assumptions that other things are roughly equal and that the relationships hold for the various age levels,we can compute an expected post-test score for each child -- a test score which reflects only natural growth and does not reflect the Head Start intervention. To do this we calculated, for each child, the number of months between pre and post-test. We then took his pre-test score and added to it the number of months between pre and post-test times the coefficient for age (0.7652). For analyses which retain the original twelve group composition, we could have used the alternative procedure of taking a child's "expected" pre-test score and adding to it the product of the number of months he was in the program times the coefficient for age. Since there was roughly 8+ months between pre and post-test time, the average gain attributable to natural maturation was roughly 6.4 points. This procedure was carried out for each of the children in the group with a valid post-test score (561 children). Each of these children's predicted post-test scores was then subtracted from his observed post-test score and a mean for the entire group of children who had both valid pre and post-test scores was calculated. Group means for these differences were then calculated and the results are presented in Table IV-4.

A number of things should be noted about Table IV-4. First, almost all of the estimated gains in Table IV-4 are statistically significantly greater than zero. Only nine of the forty-nine comparisons shown in the Table do not reach significance. Second, for the PSI and Book 3D tests, the estimated "gains" attributable to Head Start (see Table IV-4) are roughly one-half the total gains shown in Table IV-1. This indicates that one-half of the total gain is estimated to be attributable to Head Start while the other half is attributable to maturation. Thus, in effect, the children double their rate of growth on these tests during their months in Head Start. For the Book 4A and Motor Inhibition tests the Head Start experience accounts for roughly 70% of the total gain. For these tests the children are tripling their rate of growth during Head Start.

Third, by and large, the estimated gains shown in Table IV-4 for the Stanford-Binet are greater than the gains in Table IV-1. On the average, the estimated gains are 85% larger than the actual gains. This indicates that the coefficients for age for the Stanford-Binet are generally negative. In other words, older children at pre-test time on the average have lower Stanford-Binet scores than younger children. If the assumptions for this estimation procedure hold, then, it appears as if Head Start arrests a deterioration in Stanford-Binet scores and additionally accelerates the rate of growth of Mental Age as assessed by the Binet. The arresting plus the acceleration appears to be on the order of two-thirds of a standard deviation.

Fourth, there seem to be no consistent differences in estimated gains between children with and without a prior preschool experience. Fifth, there are greater estimated gains for El children on the PSI, Book 3D and Book 4A tests. There are no differences between El and Ek children on the Stanford Binet and Ek children tend to gain more on the Motor Inhibition. Sixth, there seem to be no consistent patterns of differential gains for the three ethnic groups.

A variety of assumptions were made in this analysis. First, we have no way of controlling for the effect of pre-test sensitization on the children. It may be that the specific effect of taking the pre-test contributes to the post-test score. Second, we have to make the assumption that there is no differential selection of older and younger children within groups -- that is, we must assume, for example, that the older children in a group were not more nor less clever than the younger children. There is no way of controlling for this. Third, we must make the assumption that the coefficients for age are unbiased. We have no assurance of this aside from the fact that we have physically controlled for ethnicity, preschool experiences, and entering grade and their interactions as well as the variables in the equation. Yet even these rather extensive controls do not assure that the age coefficients are unbiased.

If there were pre-test sensitization, the "estimated gains" in Table IV-4 would be overestimated. If there were selection effects into Head Start programs favoring more clever younger children and less clever older children, the "estimated gains" would again be overestimated. If, however, the selection procedure operated in the other direction, the estimated gains would be underestimated. Bias in the age coefficients could lead to either under or over estimation of "gains". Our best guess is that the combination of these influences probably leads to a slight overestimate of the "gains" shown in Table IV-4. Yet even if the "gains" were halved the overall increased growth rate would still be on the order of 25% for the PSI and Book 3D tests, over 33% for the Book 4A and the Motor Inhibition tests and the natural loss on the Stanford-Binet would be arrested.

Interpretations and Conclusions:

The two central purposes of this chapter were first, to describe the overall changes in test scores for the total sample of Head Start children and second, to estimate to what degree the changes can be attributed to the Head Start experience. Data summarizing these efforts are contained in Chapter IV-5. Column 1 of that table shows the average total gain for children in the overall analysis sample for the five outcome measures. Column 2 shows the portion of the total gain attributable to natural maturation (the estimated amount of gain that would have occurred had the children not been in Head Start). Column 3 shows the estimated amount of gain attributable to Head Start. All estimates in in this table are expressed in standard deviations of the pre-scores of the tests.

TABLE IV-5

Gains for the total analysis sample on 5 outcome measures. Observed gains are partitioned into two components -- gains attributable to maturation and gains attributable to an Head Start experience. All gains are expressed in individual level pre-test standard deviations.

Test	Observed gain (total)	Attributable to maturation (estimated)	Attributable to Head Start (estimated)
PSI	0.942	0.496	0.446
Book 3D	0.727	0.363	0.364
Book 4A	1.151	0.333	0.818
Motor Inhibition	0.36	0.10	0.26
Stanford-Binet	0.348	-0.296	0.644

Conclusions from this table are straightforward. Observed gains for the five tests varied from a low of about 0.35 standard deviations on the Stanford-Binet to 1.15 standard deviations on the Book 4A test. In all instances the gain attributable to the Head Start experience indicated that during Head Start the children at least double their normal rate of growth. For the PSI and Book 3D tests the total gains are estimated to be evenly divided between maturation and the Head Start experience. For the Book 4A and the Motor Inhibition tests the Head Start experience accounts for over two thirds of the total gains. Finally, for the Stanford-Binet the estimates indicate that the Head Start experience arrested a decline of roughly 0.3 standard deviations and additionally increased children's scores by another 0.35 standard deviations.

The text of the chapter points out a variety of untested assumptions underlying the procedures used to reach these estimates. We suspect that the procedures may have produced slight overestimates in the effect of the Head Start experience. Yet even halving the gains attributable to Head Start would result in effects of a substantial magnitude indicating at least the powerful short term effect of Head Start on the measured outcomes.

In the course of reaching these estimates three other issues were addressed in this chapter. The first set of issues had to do with differential gians by a few different types of children in the sample. Specifically we found that:

1. Children with prior Head Start experience averaged lower overall gains than children without prior Head Start experience. Thus, the overall effect of a second year of Head Start seems to be less than the effect of the first year. An indication that we must be careful in making this inference stems from the fact that our estimates of the gains attributable to Head Start for the children enrolled in a first and second year of preschool seem to be roughly equal (see Table IV-4). Thus, the difference in overall gains for the two groups may be attributed to differences in the gains expected from natural maturation. Our interpretation of this is that the first year of Head Start acted as an homogenizing experience on children (at least with regard to measured outcomes). In our analysis such an effect would reduce the differences between the prescores of children of different ages who have had a prior Head Start experience thereby lowering our estimates of the rate of growth that such children would have had without a second year of Head Start.

2. Children who will enter first grade directly from Head Start tend to gain more overall than children who will enter kindergarten though the effect is pronounced only for the Book 4A. When the gains attributable to

the Head Start experience are considered it appears as if children in El sites profit more from Head Start on the PSI, Book 3D and Book 4A. We can speculate that two things are occurring here. First, teachers of children in El sites may feel a strong obligation to prepare their children academically while teachers of future kindergarten children may not feel such an obligation. Second, since the children who will enter first grade are generally older than the children who will enter kindergarten the effect may simply be due to greater maturational readiness for instruction.

There are no differences between El and Ek children on the Stanford-Binet while the Ek chidlren tend to gain more on the Motor Inhibition test.

3. There are no discernable patterns of differences among the gains for the three ethnic groups studied here.

The second set of issues briefly addressed in this chapter has to do with ways of presenting gains. Tables IV-2 and IV-3 present data for the Book 4A and Book 3D subtests structured as criterion referenced tasks. The procedure is exploratory.

Finally, we examined the pre and post-test variances for each of the five outcome measures used in the study. With one exception the post-test variance was smaller than the pre-test variance indicating that the fan spread hypothesis is probably incorrect for these data.

Chapter V

SOME METHODOLOGICAL CONSIDERATIONS

Introduction:

For many "true" experiments there is a clearcut "best" method for analysis. The analysis strategy flows logically from the structure of the experimental design and the hypotheses of the experimenter. When, however, the experimental design is compromised as is Planned Variation's, the choice of an analysis strategy becomes less obvious. This chapter considers issues in selecting analysis strategies for the Planned Variation study. It is divided into three parts. The first part discusses what unit of analysis is appropriate. The second part considers strategies for reducing bias in estimates of differences between groups. The third part describes three analysis procedures used in later chapters of this report.

I. The Choice of a Unit of Analysis

One issue prior to the selection of appropriate analysis strategies is that of choosing a unit of analysis. In this study the choice is among models, sites, classrooms, and children. Three considerations in making this decision are: practical considerations (what is needed to answer certain questions); constraints imposed by the experimental design; and the conceptual framework (how the application of the treatment is perceived). These considerations are discussed below.

(1) First, we wanted to select a unit of aralysis that would be common to most of the questions we were asking. The possible units of analysis are models, sites, classrooms and children. For each of these we could distinguish among PV and comparison groups and thus any could be used for the analyses presented in Chapter VI (analyses contrasting overall PV and comparison effects). But for analyses presented in Chapter VII (comparisons among curricula), we could not use the model as the unit of analysis since we would have no error term for testing the significance of differences among the models. Thus, we needed to choose among sites, classrooms and children.

(2) A second consideration had to do with the sampling design used in the study. In order to obtain estimates of the variability of model to model differences, we need more than one observation for each model. The natural level of replication in this design is the site; however, there are two serious problems with this choice. First, the original design was conceived on a three level nested design (sites within models, classrooms within sites, and children within classrooms; the PV/ comparison factor would cross sites within models). Theoretically, in this design, sites would be a random factor, and variation among sites within models would be

the appropriate error term for testing the variation among models. However, since sites were not randomly assigned within models, they cannot technically be considered a random factor. And as our analyses in Chapter III indicated, sites cannot even be argued to approximate a random factor since sites within some models are clearly different from sites within other models. Unfortunately, this argument applies to classrooms and children as well, since neither involved random assignment.*

The second problem relates specifically to sites. Two models have no replication at the site level and six models have no site replication for the Level III testing. The lack of replication of sites for some models leaves us without an estimate of the error term for those models. Although it might be argued that we should limit our analyses only to those models which have replications, we decided rather to note the problem and temper our conclusions about the effects of the models rather than to eliminate them from the analyses. Since there were no compelling

*The point is that there is no intrinsic reason in the original sampling procedure for choosing sites as the appropriate unit of analysis over classrooms or individuals. In order to make inferential assessments of model to model differences, we must make the assumption that the chosed unit analysis was a random factor--that the sample of sites, classrooms or children was randomly drawn from some larger sample. Given the sampling procedure there is no reason to select any of the possible units as more closely approximating a random factor than any other. For those readers wishing to contend that inferential statistics cannot be employed without a clear indication that our unit of analysis is a random factor, we suggest that they use the significance testing as an heuristic device.

reasons associated with the design, we therefore ruled out the selection of sites as the unit.

This leaves us with either the classroom of the individual as the appropriate unit. Two arguments convinced us that classrooms were desirable. First, one of the problems faced in any experiment of this sort is the problem of fallible data. Of particular concern here is the reliability of the various independent measures (the reliability of the dependent variables is of less concern). If we use the individual as the unit of analysis there is a considerable amount of error in the assessment of any of the background and pre-test characteristics of the child. For some measures. there is little error (sex, race and age are examples). For other measures, the reliability ranges from roughly 0.65 to 0.90 depending on the characteristic. If we move to the classroom as the unit of analysis we aggregate individual observations. Given the assumption that the errors of measurement are randomly distributed with a mean of zero, the aggregation should serve to cancel out some of the error and make our measures more reliable. By and large at the classroom level the reliability of our measures can be estimated to range from roughly 0.85 to 0.99--a substantial improvement over a range from 0.65 to 0.90*

*Take for example the PSI pre-test and the Book 3D pre-test. In Chapter II we indicated that their respective reliabilities for individuals were roughly 0.90 and 0.70. Roughly 37% of the PSI pre-test variance lies between classes and roughly 26% of the Book 3D variance lies between classes. If we

The high reliability estimates obtained by aggregating into classroom means gives us the advantage in our analysis of not having to correct our independent variables by the reliability coefficients.

The second concern that led us to classrooms as the appropriate unit stemmed from our desire to use a number of measures collected on the classroom as the unit. These included teacher and teacher aide characteristics, and estimates of the degree to which the classes were implemented. Had we used the child as the unit of analysis we would have been seriously overestimating the number of degrees of freedom available for these variables.

(3) On conceptual as well as statistical grounds it seems reasonable to select the classroom as the appropriate unit. By and large, a child's experience in Head Start is confined to one classroom, one teacher and one teacher aide. There is a great deal of variation within sites in the characteristic of the teachers and aides, and so it might be argued that children in different class-

assume the class size to be constant and roughly 12 (a bit of an underestimate) we can use Shaycroft's (1962) formula to estimate the respective classroom reliabilities as roughly 0.98 for the PSI and roughly 0.90 for the Book 3D. Shaycroft's formula is: where

$$r_{\overline{aa}} = 1 - \left(\frac{1-r_{aa}}{n}\right) \left(\frac{s_a{}^2}{s_{\overline{a}}{}^2}\right)$$

$r_{\overline{aa}}$ is the estimated group selectivity, r_{aa} is the estimated individual reliability, n = the number of children in the classes, $s_a{}^2$ = the variance for individuals, and $s_{\overline{a}}{}^2$ = the variance for the classrooms.

rooms within sites undergo different experiences. This argument is strengthened when we recall the tremendous variation within sites in the sponsor's estimates of the level of implementation. Although it might also be rightly argued that different children within the same classroom undergo different experiences and therefore should be treated separately in an analysis, we have no information on what causes these differences (as we might have if we had carried out classroom observations on individual children). Thus there seems little purpose in not aggregating the children to the classroom level where we might be able to distinguish among group experiences.

Overall, then, the decision was to choose the classroom as the unit of analysis. One byproduct of this decision over the choice of the site as the unit was to increase the number of degrees of freedom that we had to work with, thereby allowing more control variables to be entered in our analysis. In this sense the choice of the classroom represents a compromise between the site and the individual--with the individual ostensibly giving us the most degrees of freedom to play with and the site giving us the fewest.

II. Reduction of Bias

In Chapters VI and VII we are concerned with the problem of comparing groups on a number of outcome measures. For reasons discussed above most analyses use the classroom as the unit of analysis. Ideally, we would have wanted classrooms randomly assigned to groups. Randomization would have insured that the probability that groups differed initially on any variable, measured or unmeasured, was small. Thus we would have had confidence in the results of direct comparisons on outcome measures because we could assume that the groups differed on treatment only--comparisons among groups could be assumed to be unbiased.

Unfortunately, sites were not randomly assigned to models and classrooms within sites were not randomly assigned to PV and comparison status. Data presented in Chapters II and III demonstrate that the composition of sites within models and of PV and comparison classrooms within sites differ in a number of possibly important ways. In the absence of randomization, no statistical method can control for all possible variables which may influence the outcome measure. If we can isolate and measure those variables* which seem important, however, we can attempt to control for biases using a variety of analysis strategies.

Our approach to choosing an analysis strategy was agnostic. We don't know the "best way" to answer the questions addressed in this report. Thus we present data

*Such variables are called concomitant variables or covariates.

from a number of analyses which use different methods of controlling for possible biases in the data. Different analyses, however, often lead to somewhat different estimates of the effects in which we are interested. Some estimated effects are consistent across different analyses and are therefore quite compelling. Others are more sensitive to the nature of the analysis and are therefore less compelling, though often suggestive. One result of this approach is to give us rough "confidence intervals" for the sensitivity of estimates to different analytic approaches.

The data used to compare groups in this study consist of pre- and post-test scores, background characteristics, and teacher and site characteristics. In comparing groups on post-test scores we generally want to control for as many important differences among the groups as possible.* Three approaches are taken here to control for differences: cross-tabulation, covariance and matching.

*The availability of pre-test scores as concomitant variables is a great advantage, but it is not at all clear how to handle them. We mav simply treat a pre-test as any other covariate, or we may look directly at "gain" scores--differences between pre- and post-test scores. The major advantage of the latter is simplicity and ease of interpretation. On the other hand, if the relationship between pre- and post-test scores is complex, the obvious interpretation may be quite misleading. In calculating the gain score we arbitrarily fix the relationship between pre- and post-test to be 1.00--thus a difference of one point in the pre-test is fixed to be associated with a difference of one point on the post-test. This may not always be accurate. For example, suppose that children with a pre-test score of 10 end up on the average with a post-test score of 15, while children with a pre-test of

The simplest approach to comparing groups is to form sub-classes by cross-tabulating observations on several concomitant variables and calculating pre- and post-test means and standard deviations (and possibly other summary statistics) for the resulting subgroups. In Chapters III and IV, for example, we divided the data into twelve subgroups, stratifying on ethnicity, prior preschool experience, and entering grade level. Direct comparisons were made between corresponding subclasses of the different groups. Such comparisons will be unbiased with respect to the variables used in the cross-tabulation. While this approach is simple and the resulting statistics easily understandable, it generates a mass of information which may be difficult to use. Note that the more we subdivide our original groups the more control we exercise over possible biases, but the fewer observations we have per subgroups. Thus the price of greater bias control is loss of precision in our estimates.

We must always face this dilemma unless we are willing to assume more structure in the way the covariates affect the outcome measure. Our second approach does exactly this. In using the general linear model or the analysis of

20 end up on the average with 30. HEre a one point difference in the pre-test is associated with a 1.5 point difference on the post-test. Using the pre-test as a covariate where we let the data fix the relationship helps us to deal with situations of this kind.

covariance (ANCOVA), we assume that the relationship between the outcome measure and the covariate has a particular mathematical form. If this assumption is approximately correct, we can make efficient comparisons while controlling simultaneously for many variables.

The assumption is that the expected outcome measure (dependent variable) value is a linear function of a set of independent variables. These independent variables may be continuous variables, dummy variables standing for membership in various classificatory groupings (e.g., Ek/El), or variables representing interactions among measured variables or transformations of them. Thus we can use ANCOVA to express post-test score as a function of variables corresponding to membership in the groups we wish to compare as well as a variety of covariates. It is then possible to calculate the proportion of the post-test variance attributable to various independent variables. In particular, we can estimate the variance explained by group membership over and above that explained by the covariates and test its significance. We can also estimate and test the significance of differences between pairs of group means adjusted for differences in the covariates. Thus if the linear model is approximately correct, we have a powerful and flexible tool for group comparison.

One problem with ANCOVA, in addition to possible departures from the assumptions of the linear model, is that low reliability of the covariates can introduce biases into the estimates and tests of group differences. But the main difficulty with ANCOVA is the necessity to specify the form of the relationship between the outcome variable and the covariates. Our third approach avoids this problem.

This approach involves finding pairs of classrooms in different groups which are close to the same on their values of a variety of covariates. Regardless of the relationship between the covariates and the outcome measure, any difference between the outcome scores of the members of the pair cannot be attributed to differences on the covariates, if the matching procedure is exact. Thus each pair provides an unbiased comparison between two groups. Since in practice it will almost never be possible to find exact matches, the efficiency of the matching procedure will depend on our ability to find "good" matches. This can be a serious practical problem.

If the functional form underlying the ANCOVA is approximately correct, it is much more efficient than matching. Matching, on the other hand, has the advantage of robustness; that is, it requires very minimal assumptions to be valid. A minor drawback of matching is that even though it implicitly controls for any sort of complex relationship between outcome and covariates, it gives no information about the nature of these relationships.

Finally, more serious problems can arise in connection with unreliability of the covariates. Matching on fallible covariates can lead to regression artifacts which distort the observed differences between groups. In general the larger the differences among covariate means for the groups we are comparing and the lower the reliability, the more pronounced will be the effect of the regression artifact.

In summary, we will rely on three sets of analyses. The first, cross-tabulation, has the advantages of ease of interpretation and lack of assumptions about the nature of the relationships between the concomitant variables and the outcome measures. Its disadvantage stems from a lack of precision from small sample sizes created by subdividing original groups on a number of concomitant variables. The second procedure, analysis of covariance, gains its strength from a set of assumptions which specify the functional relationships between the concomitant and outcome variables. If the assumptions are reasonably accurate this method should both reduce biases and offer far greater precision than the first approach. The third approach, matching, again takes us off the hook of specifying the functional relationships between concomitant and outcome variables but leaves us without anywhere near the loss of precision of the cross-tabulation approach. The drawback of the final method is that unlike cross-tabulation we do not generate the data to describe the relationships.

III. Procedures

Our approach to presenting observed post-test scores, "gain" scores and "gain score differences" needs no explanation. Some explanation, however, is required for the other two sets of analyses.

The analyses using the general linear model may be divided into two categories. Both use the classroom as the unit of analysis. In the first category are analyses in a multiple regression format with post-test classroom aggregates as dependent variables and aggregate pre-scores, child characteristics and teacher and site characteristics as covariates. We use this approach in both the analysis of overall differences between PV and Comparison classrooms and in contrasting PV models.

Additionally, in the analysis of overall differences between PV and Comparison classrooms we allow the covariates to take on different weights for each of the PV and Comparison groups. Briefly, we enter all of the covariates with a dummy variables standing for memberships in the PV or Comparison group. The covariates are assigned observed values, unless there is missing data, in which case the subsequent correlations and regressions are calculated on the missing data matrix.* Taken alone, the resulting equation allows

*One advantage in using classroom aggreaates is that there is very little missing data. Our assumption has to be, of course, that there is no bias in the aggregates even though some data is not available for all children in the classroom.

for one set of relationships between the covariates and the outcome measures within each of the PV and Comparison groups. We then enter a new set of the same covariates, this time giving them a value of zero if the observation is in the comparison group and the observed value of the observation if it is in the PV group. This procedure allows for different relationships between the outcome measures and the covariates for the PV and Comparison groups. This may be thought of as accomodating interactions between the covariates and the PV and Comparison groups in their effects on the outcome variables.

In the regression analyses contrasting different curriculum models we take a somewhat different approach. Here we create dummy membership variables for each of the models and evaluate the magnitude of the resulting coefficients against an overall adjusted comparison group effect. In these analyses we only allow for differential relationships among the groups on two variables (the PSI pre-test score and the proportion of children with a prior preschool experience). Introduction of other interactions proved too unwieldy and not worthy of the bother.

The second set of analyses with the linear model approach used a multivariate analysis of a variance framework. This allows us to examine a number of dependent variables simultaneously. For the study of differences between PV and Comparison classrooms we used a two factor design (models by PV/

Comparison). Only classrooms in sites with both a PV and a Comparison group were included in the design. This gave us an eighteen cell design (9 models by PV/Comparison). The interpretation of the model to model differences in this design is difficult since both the PV and Comparison means are pooled to come up with a model effect. However, the PV/ Comparison contrast gives us an overall estimate of the differential effectiveness of the two groups and the interaction terms give us some idea of whether there are model to model differences in the relationship between the comparison and PV groups.

When we compare curriculum approaches within the analysis of variance framework, two multivariate analysis approaches are used. First, we directly compare the PV model groups in a one way analysis of covariance format. This is a straightforward approach but given the differences between models that we pointed out in earlier chapters, it might be misleading. Thus, we also carried out a one way design with nested PV and Comparison groups within models. This let us make one degree of freedom contrasts between PV and Comparison groups within a model. Again, only classrooms within sites with both PV and Comparison groups were used.*

*In all of the multivariate analyses we present both univeriate and multivariate tests of significance and use a variety of child aggregates, teacher characteristics, and site characteristics as covariates. The analysis of variance approach is an exact least squares solution for unbalanced designs. The particular method used calls for estimation of effects by equally weighting all appropriate cell means. Covariance adjustments are carried out around an unweighted mean of the cell means for the covariates.

Two problems should be noted with these multivariate analyses. First, although we introduce a variety of covariates, we do not test for homogeneity of the regression surfaces. Second, we do not take complete advantage of the match between PV and Comparison groups within sites. To do this we would have been required to use the site as the unit of analysis and carry out a repeated measures design -- we rejected this for reasons given above. Our only attempts to account for the match within sites was to eliminate from some analyses sites without both PV and Comparison group and to include as covariates some site level characteristics such as the variable assessing entering elementary grade level (E1/EK).

The third set of analyses used matched PV and Comparison classrooms. As we remarked earlier our purpose was to develop an analysis strategy which did not require our initially specifying the functional relationship between the covariates and the dependent variables and which did not entail the loss of precision resulting from cross-tabulation techniques. Although we matched at both the individual level and in three ways at the classroom level we present results from only two of the classroom matches. Results from the other matches were highly consistent with those reported. We first present the procedures used for matching and then consider details of the analysis. Four steps were required in creat:ng the matched samples. The steps involved solving a number of theoretical and practical problems. Since there are few precedents in the literature we go into considerable detail both to justify and to explain our admittedly ad hoc procedures.

1) The first step was to decide upon a set of variables to match with. A number of regression analyses carried out on both gain scores and post-test scores suggested that we use seven background characteristics and the pre-test scores themselves as matching variables. The seven aggregate background characteristics were mean age in the classroom, percent black, percent Mexican-American, mean income, mean household size, mean mother's education and percent with prior preschool experience. In order to have observations for all classrooms we estimated observations for the very few missing data points by assigning them the mean for the overall group. Three pre-tests were chosen--PSI, Book 3D and Book 4A. Data were present for all observations for these variables. Although the use of the pre-test scores in matching greatly increases the precision of the matching it also increases the possibility that regression artifacts will influence the estimation of effects. As a consequence we carried out matching procedures on two sets of variables--for the seven background characteristics with the three pre-test means and for the seven background characteristics alone. Due to the very high estimations of reliability for all of our aggregate variables we think that the chance

of regression artifacts seriously affecting the estimates is small and therefore we favor the ten variable match. Nonetheless, matching with both sets of variables gives the reader the opportunity to make up his own mind.

2) The second step was to develop a method for simultaneously matching on a number (either seven or ten) of variables. Two strategies came to mind. The first required ordering the variables in a particular priority and then matching classrooms in a step-wise fashion on these variables. Thus we might have first grouped PV and Comparison classrooms by categories of preschool experience and then within the categories create further subgroups on mean mother's education, etc. until all matching variables had been exhausted. We rejected this approach, however, for two reasons. We found it difficult to order the variables and we found it difficult to create meaningful categories of the variables -- which due to the aggregation were by and large, continuous. A second strategy, therefore, was adopted. In general, this approach required locating each of the PV and Comparison classrooms in multi-dimensional space defined by the matching variables. Once all of the classrooms are located in this space we can then argue that similar classrooms are close to each other while quite different classrooms are far apart from each other. Following this logic, we could then match PV and Comparison classrooms by choosing nearby pairs. Actually carrying out this procedure was

difficult, however, for the matching variables are intercorrelated. To calculate distances among points in a space defined by correlated dimensions requires working with some fairly complicated covariance terms -- something we didn't want to do. Calculating the distance, however, between points in a space defined by uncorrelated or orthogonal dimensions is quite straightforward as Pythagorus demonstrated a long time ago. We therefore solved our problem by generating a number of orthogonal variables to define a subspace within the space defined by the original matching variables. The technique used for this was principal components analysis. All 166 classrooms in the final analysis sample were observations in this analysis. Our procedure was to carry out the principal components analysis and to retain for matching purposes only components with a latent root greater than one. We then calculated scores for each of the classrooms on each of the components, retaining the differential weight of the size of the latent root. This resulted in five component scores for each classroom for the ten variable analysis and four scores for the seven variables analysis. Within the separate analyses, the sets of scores were uncorrelated. Moreover, we have some assurance that they are reasonably reliable. Aside from the fact that the original observations were classroom aggregates and therefore generally of high reliability,the component scores can be viewed as probably having greater reliability than the individual variables since they are linear composites of a number of highly correlated variables. Moreover, the elimination of some of the factors with latent roots less than 1.0 may have removed some of the random noise from the matching variables.

3. Third, after component scores were calculated for each of the classrooms, a distance matrix was constructed. The distance matrix had PV classrooms as one dimension and comparison classrooms as the other. Each cell in the matrix contained the distance between a PV classroom and a comparison classroom. The distances between classrooms were computed by taking the square root of the sum of the squared differences between the component scores of the classrooms.

4. Fourth, once we had the distance matrix, we needed to find the "best" matches. This is not a trivial problem as Rubin (1971) points out. But finding the strategy for the best fit was not the only problem. First, we wanted to match not only on the vari·bles included in the components analysis, but also on the entering grade level of the site. Second, we faced the problem of having many more PV classrooms than Comparison classrooms. If we wanted to find a match for every PV classroom we would be required to use some comparison classrooms two or more times. How were we

to deal with duplications? Third, we had to decide upon some criteria for evaluating the quality of our matches. The first problem was easily resolved -- we only matched classrooms if they were from sites with the same entering grade level; they we only matched EK PV classes with EK comparison classes. The second problem was somewhat more complicated. Our resolution of the problem of duplicate comparison classrooms was to treat PV models separately. The procedure took the entire set of PV classrooms within a model and then searched for the "best" match for each classroom from the entire set of comparison classrooms. No duplications were allowed within PV models. The idea was to not constrain the number of degrees of freedom for contrasts within models. This approach essentially created eleven separate sub-experiments, each comprised of PV classrooms within a model matched with comparison classrooms from the entire pool of comparison classrooms. Since there were at most twelve PV classrooms within a model and 65 comparison classrooms, we had a lot of leeway in our matching to accomodate extreme PV classrooms.

Third, we chose a least squares criterion for evaluating alternative matches. Our argument was based on the fact that we were matching the PV classrooms within a model altogether rather than independently -- since we did not allow duplicates within models. We therefore needed an overall measure of the average differences among different combinations of matched PV and comparison classrooms in order to get some idea of the best combination for models. We chose the criteria to be the minimum value of the sum of the squared instances between the matched PV and comparison classrooms. Another possibility was to choose the minimum sum of the distances between matched PV and comparison classrooms. In practice the two seem to result in essentially equivalent matches. With all these decisions made, we only needed to find the "best" matches. We did not solve the problem -- like Rubin, we settled on heuristic devices.* We used four general strategies.

In each of the following steps we deal with the models separately. The first step in each strategy was to select for each PV classroom in a model the 12 closest comparison classrooms. We called this a "reduced" distance matrix. If there was no overlap in the closest matches we were all set -- we simply chose the closest ones. If however, there were comparison classrooms that were closest to more than one PV classroom, we had to figure out some way of selecting the best combination of matches. One approach started by taking the shortcut distance between any of the PV and comparison classrooms and accepting that as one matched pair of classrooms. Since we did not allow duplicate comparison classrooms within a model, we then had to eliminate from the reduced distance matrix all occurences of the matched comparison classroom. After that step was carried out, we again

*For those of you who think this is a simple task, we recommend you try it with some data.

selected the closest match etc. for all of the PV classrooms. Once we had matched each of the PV classrooms with a comparison classroom, we then computed a sum of the squared distances. The second approach used was to select the PV classroom that had the worst match in terms of distance with any of the comparison classrooms. This PV classroom was matched with its nearest comparison classroom, the comparison classroom was eliminated from the reduced distance matrix and the process was repeated for the PV classroom with the next worst match. A sum of squared distances was then computed for this procedure. We might call the first procedure a heuristic maximin procedure and the second a minimax procedure.

The third approach was to select a PV classroom randomly and match it with its closest comparison classroom. Then the comparison classroom would be eliminated from the reduced distance matrix and another random PV classroom chosen, etc. A sum of squared distance was then calculated for this procedure. The fourth procedure took the best result from the other procedures and tried out a limited power approach to see whether the overall sum of squares could be reduced.

In general, the power procedure slightly improved upon other procedures. We might note that there were considerable differences in the sums of squares of the distances for the four procedures. Within each model, then, a heuristically "best" matched set of PV and Comparison classrooms were chosen.

This procedure was carried out independently twice -- for the 10 variable, five component solution and for the 7 variable, 4 component solution. Additionally, for each solution, the matching procedure was carried out separately for the sample of all Level II and III sites and for the sample of only Level III sites. We had to carry out the Level III only matches to insure that we could successfully analyze the Stanford-Binet.

To analyze the data we decided upon a one way nested analysis of variance with one covariate. Our procedure treated each of the sites as a level in a one way design using the difference between the matched PV and comparison post-test classroom means as the dependent variable and the difference between the matched PV and comparison pre-test means as the covariate. Correction for the reliability of the covariate were carried out using the Lord-Porter (see Porter, 1972) technique. Because we knew a priori that the grand mean for the covariate should be zero (since it is a difference between pre-score means for matched classrooms) we calculated the covariance adjustment around a zero mean rather than around an observed grand mean. Overall PV/Comparison contrasts and model effects were calculated by pooling unweighted adjusted means across the sites.

All of this sounds pretty complicated for a simple one way analysis of variance with covariance adjustment. Unfortunately, little theory and thought have been given to the practical problems of dealing with matching in quasi-experiments of this sort and as a consequence many of our procedures seem more than a little ad hoc. Yet for a number of reasons it seems to us that this procedure outlined above might contribute a lot of power to our analyses.

First, it allows us to deal with two very practical analysis problems. As we pointed out in Chapter II, we have no comparison classrooms for two models -- the Enablers and REC. Moreover, for two other sites we lacked either on-site or off-site comparison classrooms. Since direct comparisons among models seems to be a weak approach -- because the sites within models seem to differ on some important characteristics -- we have tended to place our reliance on an indirect comparison among models, mediated through the contrast between models and their comparison classrooms. But to carry out this procedure we need some assurance that the PV and comparison classrooms are somewhat equivalent. Pairing by location does this for those sites with both PV and comparison classrooms but it does nothing for the Enablers, REC and the two other sites without comparison classrooms. Only a matching strategy could allow us to place these problem sites in an analysis contrasting models with comparison classrooms. Second, even for sites which have both PV and comparison classrooms, certain problems exist in the analysis. As we noted, without the site as the unit of analysis, there is no natural way to use the pairing by location to reduce the error term in our analysis; classrooms are not matched within sites and often there are more PV than comparison classroom. Matching classrooms by variables rather than by location eliminates each of these problems. The matched classrooms pairs can be treated as the unit of analysis and the design becomes balanced with regard to comparisons within models.

Third, as we noted earlier, the matching procedure as a strategy for control does not require us to specify the functional relationship between the control variables and the outcome variables as other control procedures, relying completely on the general linear model. This strikes us as an extremely important argument granting, of course, that we have chosen the right variables to match with.

One final remark. Two principal problems that analysts have raised about matching stem from issues of the reliability of the covariates and the similarity of the matching covariates in the two samples in their distributed characteristics. By and large, we think the variables used for matching are extremely reliable, and by and large, the characteristics of the covariates in the samples being matched are very similar (see Chapters II and III). Yet it still seems appropriate to watch out for extreme cases in our analyses of the matched samples.

In the next two chapters we use the procedures outlined above. Chapter VI considers the question of whether there are overall differences in effects between PV and Comparison classrooms. Chapter VII focuses on the question of model to model differences in effectiveness.

Chapter VI

OVERALL DIFFERENCES IN THE EFFECTS OF PV AND COMPARISON CLASSROOMS

Introduction:

The eleven preschool models in the Planned Variation study have somewhat differing emphases on the outcomes measured in this study. We might therefore expect to find outcome differences among Planned Variation classrooms. But this does not imply that we would expect an outcome averaged across all PV classrooms to be significantly different from an average of all Comparison classroons. Because the expected differences among models are lost in looking at overall averages, it is difficult to attach much substantive meaning to a contrast of all PV classes versus all Comparison classes. If the degree of curriculum emphases in a measured domain does affect the outcome, then a prediction that the average of all PV classes would show more change than the average of all Comparison classes requires the assumption that the modal emphasis in this domain is greater for the PV classrooms. We have no way of obtaining this information and thus no way of knowing whether to expect PV classrooms on the average to "do better" or "do worse" on our measures than the Comparison classrooms.

The main reason for contrasting the overall effects of PV and Comparison classes is to determine whether the effect of the extra $350.00 per child spent on children in PV classrooms has an effect on measured outcomes. For while we cannot identify modal curriculum emphases for the two groups, we can speculate that the additional personnel and materials available to the PV classrooms might have an effect.

This chapter reports a series of analyses on the differential effects of PV and Comparison classes. The chapter has four sections. In the first section we contrast raw "gains" for the total PV and Comparison groups and for 12 subgroups within each. The child is the unit of analysis. The purpose of these contrasts is to give the reader some feel for observed differences before we carry out procedures of control and adjustment. Section II reports on a series of regression analyses which have a PV/Comparison dummy variable and a set of background and teacher characteristics (with separate slopes for PV and Comparison groups if necessary) as independent variables. In Section III, we report on a series of two-way analyses of variance. The approach used is a multivariate exact least squares solution with models as one factor and PV/Comparison as the second factor. In Section IV, we report two sets of one-way nested analyses of covariance using matched samples with adjustments for fallible covariates.

Each of the reported analyses offers some slightly different information about the effects of PV and Comparison classes. As a consequence, there are slight differences in the estimates of effects. The general conclusion that can be reached from all of these analyses is that there are no differences between the PV and Comparison groups in effects on the measured outcomes. This was not an unexpected finding and in no way implies that Planned Variations is a failure. For at no time was an objective of the Planned Variation study to demonstrate that the simple infusion of funds into preschools would have an effect. Rather the intent of the Head Start Planned Variation study has been to investigate differences in the processes and outcomes of different preschool curriculum models. To do this, the Planned Variation strategy required that preschool curricula be selected and studied for a variety of reasons -- not solely because they all intended to maximize outcomes on the variables we have measured.

I. Differences between the PV and Comparison samples -- Observed overall subgroup changes.

In Table VI-1 we present some overall descriptive statistics for the PV and Comparison groups. As discussed in earlier chapters, there is considerable similarity between the PV and Comparison groups on pre and post-test means and variances. The only test which looks very different for the two groups is the Stanford-Binet. Here we see that PV children, on the average, increase their Binet scores by

TABLE VI-1

Some selected statistics for overall PV and Comparison samples for five tests: The individual child is the unit of analysis. The sample is the total analysis sample described in chapter III.

TESTS

	Book 3D		Book 4A		PSI		Motor Inhibition		Stanford-Binet	
	PV	Comp.	PV	Comp.	PV	Comp.	PV	Comp.	PV	Comp.
Pre-Test Mean	11.851	12.103	5.548	5.889	35.498	35.835	5.047	5.116	90.511	90.042
Post-Test Mean	14.229	14.508	9.566	9.471	46.558	47.690	5.395	5.482	96.490	92.975
Pre-Test Standard Deviation	3.330	3.122	3.172	3.377	12.297	12.917	0.542	0.524	13.575	13.298
Post-Test Standard Deviation	3.132	2.942	4.719	4.749	10.326	10.300	0.594	0.528	13.308	12.683
N's	1188	805	1178	803	1197	806	465	300	389	297

roughly six points while Comparison children increase their Binet scores by only three points. This difference of three points is roughly 20-25% of the standard deviation of the Binet for these groups and roughly 18-20% of the standard deviation for the Binet for the nation as a whole. None of the other differences in overall observed gains exceeds 10-12% of the standard deviation for its test.

Table VI-2 shows tests of significance for the differences in overall mean gain (at the bottom of the Table) and for 12 subgroups of children.* Three overall differences in "observed gains" are statistically significant at the .05 level or beyond. Two of the significant differences favor the PV group (Book 4A and the Stanford-Binet) and one favors the comparison group (PSI). However, neither the Book 4A or the PSI difference is of sufficient magnitude to be of great interest -- in neither case does the difference exceed one-tenth or the post-test standard deviation.

In the body of Table VI-2 we observe nine statistically significant differences out of forty-nine possible. For three of the groups two tests show statistically significant differences. Mexican American children without preschool

*The figures in Table VI-2 are differences of difference scores. The computation of the scores had two steps. First, the pretest mean for the PV group was subtracted from the post-test mean for the PV group giving us the PV mean "observed gains". Then the pre-test mean for the comparison group was subtracted from the post-test mean for the comparison group giving us the comparison group mean "observed gain". The comparison "observed gain" was then subtracted from the PV "observed gain", giving the differences presented in Table VI-2.

TABLE VI -2

Differences between PV and comparison groups in observed "gain" scores. Values in table are [$(PV_{post} - PV_{pre}) - (comp._{post} - comp._{pre})$] means and N's for the group. Only children with valid pre and post-test scores were used in the calculations.

GROUP			TEST				
Ethnicity	Prior Preschool	Entering Grade	Book 3D	Book 4A	PSI	M-I	S-B
Mexican-American	No	El	x_{diff} = 0.705 df = 109	1.499* 109	1.455 109	0.721*** 61	-----
Mexican-American	No	Ek	-1.043 104	1.422 99	-1.758 104	-0.160 20	5.037 38
White	No	El	0.009 156	0.442 156	-2.585* 155	0.013 101	1.831 63
White	No	Ek	0.097 420	0.652 416	-0.823 424	-0.187 167	1.021 145
Black	No	El	-0.199 248	0.791 248	0.311 250	-0.245* 114	-3.039 * 127
Black	No	Ek	0.080 555	-0.068 554	-2.016** 559	-0.138 103	10.100*** 162
Mexican-American	Yes	El	-----	-----	-----	-----	-----
Mexican-American	Yes	Ek	-----	-----	-----	-----	-----
White	Yes	El	0.356 67	0.560 67	2.920 67	0.045 45	-1.729 29
White	Yes	Ek	0.966 73	2.010* 72	0.249 73	-0.255 31	0.924 25
Black	Yes	El	-0.558 109	2.413** 109	1.286 109	-0.181 48	-1.487 46
Black	Yes	Ek	-0.112 112	-0.668 111	-1.232 113	0.273 14	-1.453 27
TOTALS			-0.027 1991	0.436 * 1979	-0.795* 2001	-0.018 763	3.046*** 684

* statistically significant at the .05 level

** statistically significant at the .01 level

*** " " " " .001 "

experience in E1 sites appear to gain more in the PV classes on the Book 4A test and on the Motor Inhibition test. Black children without prior preschool in E1 sites appear to do better in Comparison classes than in PV classes on the Motor Inhibition and the Stanford-Binet tests. And black children without preschool experience in EK sites tend to gain far more on the Stanford-Binet if they are in PV classes but they tend to gain less on the PSI if they are in PV classes. The other three significant differences are scattered among the remaining seven groups.

There appears to be only one consistent pattern in this table. There is a modest tendency for PV children with prior preschool experience to do somewhat better relative to their Comparison groups than PV children without prior preschool experience do relative to their Comparison group. This holds for all tests but the Stanford-Binet. Perhaps Head Start programs with systematically planned curricula are more effective for second year preschool students relative to conventional Head Start curricula, than they are for first year preschool students.

The overall comparisons at the bottom of Table VI-2 are controlled only for the pre-test (and assume a perfect relationship between pre and post-test.) The contrasts in the body of Table VI-2 control physically for ethnicity, prior preschool experience, entering grade and their interactions as well as for the pre-test (again assuming a perfect relationship between pre and post test). When we contrasted the PV and Comparison gains controlling only for the pre-test, we found statistically significant differences on three of the five variables, two favoring the PV group. Yet when we look more closely at the data and introduce the three control variables we find only 9 of the 49 contrasts statistically significant with 6 of the 9 favoring the Comparison group. The essential message here is that the introduction of controls tends both to reduce the proportion of statistically significant findings and to cloud the question of whether the PV or the Comparison children are, on the average, gaining more. This suggests that observed differences between the PV and Comparison groups may be due more to initial and controllable differences between the composition of the two groups than to the effects of their Head Start experiences. In the following sections we pursue this issue.

II. Some regression analyses with a PV/Comparison group membership variable, and a number of covariates.

The issue addressed is whether there are statstically significant PV/Comparison group differences which express themselves in a general linear model framework with the classroom as the unit of analysis. The approach is straightforward. In multiple regression terms, we examine the coefficient for a dummy variable (indicating membership in either a PV or a Comparison class) which is entered in a regression equation with a number of control variables assessing classroom aggregate characteristics of children, teacher and site characteristics and with post-test scores as the dependent variable. The two groups are allowed to have separate coefficients for each of the dummy variables.*

Another perhaps simpler way of looking at this analysis is to think of it as a two group analysis of covariance -- in this instance the two groups are the PV and Comparison groups and the covariates are the "control" variables listed in the footnote below.

*For the PSI, Book 3D and Book 4A we present results from 2 analyses. In analysis 1 on the total sample of classrooms ve use PSI pre, Book 3D pre, Book4A pre, percent female, percent prior preschool, mean age, mean income, mean mother's education, mean household size, percent Mexican American, percent Black, years teacher experience in Head Start, teacher race, teacher aide years in Head Start, teacher certification, average staff working conditions, and whether the site is El or EK as control variables. In analysis 2 we limit the sample to the Level III sites and use the Stanford-Binet pre-score as an additional control variable. We also use the Stanford-Binet post-score as a dependent variable in this set of analyses. For the Motor Inhibition post-test we limit the sample to the classes with valid pre and post Motor Inhibition and use the PSI, Book 3D, Book 4A and Motor Inhibition pretests as well as the other child aggregate, teacher and site characteristics as control variables. In all of the analyses we allow for separate slopes for the PV and comparison groups. Following Cohen (1971) our procedure for doing this was to calculate two sets of control variables (or covariates). The first set have observed values for both the PV and comparison groups. The second set are assigned a value of zero of the classroom is a comparison classroom and the observed value if the unit is a PV classroom. The first set of covariates are forced into the equation and we then let as many of the second set (which assess differential slopes) of covariates enter as possible.

Three separate sets of analyses were carried out. In analysis set 1 the dependent variables were the PSI, Book 3D, and Book 4A. The total sample of classrooms was used for this analysis. Analysis 2 utilizes only classrooms in the Level III sites. The dependent variables were the Stanford-Binet, the PSI, Book 3D and Book 4A. Analysis 3 was conducted on the sample of classrooms with valid Motor Inhibition pre and post-test scores. The Motor Inhibition test was the only dependent variable for analysis 3.

Table VI-3 gives pre and post-test N's, means and standard deviations for the five tests used in the analyses. Data from these analyses are shown in Table VI-4. In columns 1 and 2 are zero-order correlations of the dummy PV/Comparison group membership variable with pre- and post-test scores respectively. None of the correlations is statistically significant and none of the differences between the pre and post-test correlations is significantly different from zero -- though the post-test correlation for the Stanford-Binet approaches statistical significance as does the difference between the pre and post-test correlations for the Stanford-Binet.

Column 3 contains the standardized regression coefficients for the group membership dummy variable for the total equations -- allowing for separate coefficients on the covariates for the two groups. Column 4 contains the same group membership standardized coefficients for an equation allowing for no group by covariate interactions (i.e. only one slope for each covariate is allowed.) In no instance does the group membership coefficient reach statistical significance. Clearly the PV/Comparison membership variable has little predictive power in these equations.

TABLE VI-3

Some selected statistics for Pre and post-test for both the PV and comparison groups. Classroom as the unit of analysis.

Test	PV N	Comp. N	PV Pretest Mean	Comp. Pretest Mean	PV Pretest SD	Comp. Pretest SD	PV Post-test mean	Comp. Post-tcst mean	PV Post-test SD	Comp. Post-Test SD
PSI (Analysis 1)	101	65	35.081	35.384	7.362	6.891	45.962	46.930	6.593	6.937
Book 3D (Analysis 1)	101	65	11.792	11.993	1.708	1.561	14.075	14.351	1.864	1.753
Book 4A (Analysis 1)	101	65	5.572	5.788	1.504	1.592	9.242	9.208	2.854	2.895
PSI (Analysis 2)	61	47	35.324	35.649	7.105	6.952	46.455	47.632	5.965	6.735
Book 3D (Analysis 2)	61	47	11.923	12.066	1.734	1.502	14.187	14.590	1.684	1.777
Book 4D (Analysis 2)	61	47	5.544	5.700	1.574	1.324	9.443	9.520	2.827	2.774
Motor Inhibition (Analysis 3)	87	59	5.002	5.060	0.387	0.429	5.379	5.380	0.377	0.358
Stanford-Binet (Analysis 2)	61	47	90.591	90.299	8.146	7.380	96.401	93.547	7.904	7.234

TABLE VI -4

Some statistics for a set of regression equations with post-tests as dependent variables and a large number of control variables^{T} . The independent variable of interest was a dummy (PV/comparison) group membership variable. $\mp$ Classrooms are the unit of analysis. See Table VI-3 for other related statistics.

Test	Zero-order correlations of (PV/comparison) dummy variables with test variables		Standardized Regression Coefficient		Overall %age of variance explained by total equation
	Pre-test	Post-test	All variables in equation $\mp$ (tot.eq.)	Allowing for only 1 slope (red.eq.)	
PSI (Analysis 1)	-.021	-.070	-.081	-.057	79.0
Book 3D (Analysis 1)	-.059	-.074	-.063	-.068	71.7
Book 4D (Analysis 1)	-.068	.006	.018	.009	69.6
PSI (Analysis 2)	-.023	-.092	-.034	-.046	80.3
Book 3D (Analysis 2)	-.043	-.115	-.080	-.087	77.8
Book 4A (Analysis 2)	-.052	-.014	.014	.002	70.9
Motor Inhibition (Analysis 3)	-.069	-.002	.025	.028	42.4
Stanford-Binet (Analysis 2)	.018	.183	.123	.112	57.8

Notes to Table VI-4

† PV is coded 1, comparison is coded 0.

‡ All covariates were entered as though the regression planes were entirely parallel. Then as many PV covariates as necessary were entered to adjust for differences in slopes. Stepping was terminated when the standard error of the equation reached its lowest point.

T. For analysis 1 the covariates were PSI, Book 3D and Book 4A pretests, percent female, percent prior preschool, mean age, mean income, mean mother's education, mean household size, percent Mexican American, percent black, years teacher experience in Head Start, teacher certification, average staff working conditions, teacher aide years of experience, and whether the site is an El or EK site. For analysis 2 the Stanford-Binet pre-test was an additional covariate. For analysis 3 the Motor Inhibition test was an additional covariate -- in analysis 3 the Stanford-Binet pre-test was not used.

The slight differences between the standardized coefficients for the entire equations and for the equations allowing for only one overall coefficient for each covariate reflects the fact that in some equations a few of the 17 or so covariates have a somewhat different relationship to the PV group than to the comparison group. In general, however, the provision for different slopes for the two groups adds very little explained variance over the reduced equations which provide for only one slope. Moreover, the variables which appear to operate differently in the two groups generally were not the same from equation to equation. This indicates, along with very slight change in the magnitude of the PV/Comparison coefficient, that the relationship between the covariates and an outcome measure are generally similar for the PV and Comparison groups. In other words, it appears as if there are few important interactions of these covariates with the PV/Comparison group variable.*

*We do not present an overall F for the test of the homogeneity of the regression [illegible] since in no equation did all or even most of the separate slope variables enter after the single overall covariates were entered. Specifically for analysis group 1 on the [illegible], after all seventeen of the original coefficients had entered only three variables intended to measure separate slopes came into the equation before the determinant became too small to allow numerical stability in the result.

Two of the variables were statistically significant ([illegible] and [illegible]) [illegible], indicating that at least for this sample that they had different slopes for the two groups. In analysis [illegible] for the [illegible] [illegible] three variables entered-- [illegible] only one was statistically significant ([illegible]). [illegible] coefficient variable entered significantly [illegible] analysis 2 equations.

For the [illegible] no separate slope variable entered significantly [illegible] and only one is statistically significant [illegible]. For the [illegible] [illegible] coefficient entered [illegible] the Stanford-Binet [illegible] separate slope coefficients [illegible], [illegible] race and percent female. When we are

Column 5 in Table VI-4 shows the percentages of variation explained by the total equations. The percentages range from 42.2% for the Motor Inhibition to slightly over 80% for the PSI in analysis 2. In all instances the equations are highly significant and indicate that while the simple linear model does not explain all of the variation, it does very well in most instances.

Three conclusions can be reached from this section. First, when dealing with the classroom as the unit of analysis and with the entire sample, there are no statistically significant differences between the PV and Comparison groups either without controls or after extensive linear controls for any of the five post-test scores. Second, at least for the PV/Comparison contrast it looks as if fitting an equation without consideration for separate slopes for the two groups is almost as efficient as providing for separate slope coefficients for the various covariates. Third, for at least three of the post-test variables (PSI, Book 3D and Book 4A) we find that fairly primitive equations account for a great deal of the classroom to classroom variation (over 70% for each of these variables in both analyses). And even though we explain only 42.2% and 57.8% of the classroom variation for the Motor Inhibition and the Stanford Binet tests respectively, this still indicates a reasonably good fit for a linear additive model.

dealing with 17 (analysis 1) or 18 (analysis 2 and for the MI) covariates these seem like few interactions. Though the overall fit is somewhat better with some of the separate slope coefficients added, the overall gain in precision seems slight.

IV. Results from some Exact Least Squares Solutions of Unbalanced Two Way Analyses of Covariance

This section reports statistics from three exact least squares solutions of Unbalanced Two Factor analyses of covariance. The two factors are Models (9 levels) and PV/Comparison classes. The samples include only classes from sites with both PV and Comparison classes. Classes are pooled across sites into models. Only nine models are represented--REC and the Enablers are left out of the analyses. Data from the three analyses are presented in Table VI-6.

For analysis 1 the dependent variables were the PSI, Book 3D and Book 4A. Covariates are listed in Table VI-6 and with one exception, are the same as those used in the regression analyses reported in the preceding section of this chapter. The first two columns show the overall PV and comparison group N's -- the number of classrooms used in the analysis. Columns 3 and 4 show the estimated combined means for the PV group as a whole and for the Comparison group as a whole. These means can be interpreted as the unweighted average of the nine adjusted cell means for the levels of the PV/Comparison factor. Column 5 shows the estimated effect--the difference between the two combined means. The adjustments are calculated around unweighted means of the covariates.

A comparison of the adjusted means in Table VI-6 and the raw means presented in Table VI-1 shows a strong similarity even though the samples were slightly different (the sample used in Table VI-6 is smaller due to the elimination of classrooms in sites without a comparison group), even though the means in Table VI-6 were unweighted averages and the means in Table VI-1 were weighted averages of classrooms, and finally, even though one set of means was adjusted while the others were not. The magnitude of means for the PV and comparison groups seem remarkably stable even given changes in samples, methods of estimation and methods of adjustment.

Of the three estimated differences for analysis 1 only the PV/Comparison contrast for Book 3D shows statistically significant results. The difference (-.470), favoring the comparison group, is roughly 0.15 of the standard deviation of the individual post Book 3D test and is significant at the 0.05 level. However, since the PV/Comparison effect is correlated with the model to model effects, the PV/Comparison effect does not reach significance when the model to model differences are taken out first (see the F test for PV/Comparison group differences). Moreover the overall multivariate test for differences between the PV and comparison mean vectors is not statistically significant. This indicates that the Book 3D effect is marginal at best. For neither of the other two variables does an "estimated difference" reach even 10% of the post-test standard deviation.

The last two columns of this table indicate the overall univariate F tests for the interaction term and for model to model differences. In no case was the univariate F for interactions statistically significant though the multivariate test for interactions did reach statistical significance--P < .05. This indicates first, that we are generally justified in interpreting main effects, and second, that there is a strong correspondence between the adjusted means for the PV and comparison groups within models as well as overall. The significant multivariate F for interactions, of course, tempers this final conclusion.

The last column indicates that for the PSI there are strong model to model differences in adjusted means; of course, these means are calculated by pooling both PV and comparison group classrooms and, therefore, interpretation is difficult. The univariate F's for model to model differences for the other two variables are not statistically significant. The overall multivariate F for model to model differences is highly significant.

Analysis 2 used a sample of only classrooms in the Level III tested sites, again eliminating those classrooms in sites without comparison groups. The reduction of the sample to only Level III sites allows us to include the Stanford Binet in the analyses--the post-test is included as a dependent variable and the pre-test as a covariate. Other than the sample reduction and the addition of the Stanford-Binet analysis 2 is the same as analysis 1.

Analysis 2 adds little information about the PSI, Book 3D and Book 4A except to indicate that the addition of the pre-test Binet as a covariate and the change in the sample results in a non-significant estimated difference in the PV/Comparison contrast for the Book 3D test. The magnitude of the Book 3D difference (now -0.366), however, changed only slightly from the previous analysis.

The largest change in differences can be seen for Book 4A--it goes from an estimated difference of 0.332 in analysis 1 to an estimated difference of -0.018 in analysis 2 -- neither effect is statistically significant. Once again the univariate interaction effects are all insignificant. In contrast to analysis 1, however, the Book 3D test as well as the PSI showed statistically significant differences among models. The Stanford-Binet also showed statistically significant model to model differences.

In analysis 3, four post-tests were included as dependent variables: PSI, Book 3D, Book 4A and the Motor Inhibition. Table VI-6 shows results only for the Motor Inhibition test. For the Motor Inhibition there are no significant differences for either the PV/Comparison or the Univariate Interaction contrast. The univariate test of model to model differences is statistically significant at the .01 level.

By and large these findings are consistent with the findings in earlier sections of this chapter. There is little indication of statistically significant PV/Comparison group differences. The only exception to this is the small statistically significant effect found for the Book 3D test on analysis 1.

TABLE VI-6

N's, Estimated Combined Means and Estimated Effects for a PV/Comparison 1 degree of freedom contrast. Design is a crossed two way analysis of covariance; nine models by PV/Comparison. Tests of significance for the estimated effects are shown in the Table. Tests for significance of PV/Comparison by model interaction and for overall model to model differences are also shown--note that model to model differences pool PV and Comparison groups together. Only sites with both PV and Comparison groups are included in the analyses. The classroom is the unit of analysis. See footnote for the listing of the covariates.+

Test	PV N	Comp. N	PV est. combined mean with adjusts. for covariates	Comp. estimated combined mean wth adj. for covar.	Estimated effect (Difference)	F for PV/Comp. Difference++	F for Interaction	F for model to model differences
PSI (Analysis 1)	77	65	46.12	47.00	-0.885	1.99	1.31	4.23*
Book 3D+++ (Analysis 1)	77	65	14.05	14.52	-0.470*	3.18	0.55	1.36
Book 4A (Analysis 1)	77	65	9.38	9.05	0.332	0.49	1.87	1.16
PSI (Analysis 2)	53	47	46.83	47.64	-0.812	2.01	1.69	3.86**
Book 3D (Analysis 2)	53	47	14.28	14.65	-0.366	3.24	0.49	3.19**
Book 4A (Analysis 2)	53	47	9.403	9.422	-0.018	0.67	1.96	0.94
Motor Inhib. (Analysis 3	64	59	5.337	5.415	-0.078	1.25	1.25	3.09**
Stanford-Binet (Analysis 3)	53	47	94.35	94.77	-0.422	0.01	1.01	3.81**

+The covariates for Analysis 1 are PSI Pre-test, Book 3D pre-test, Book 4A pre-test, mean age, % Black, % Mexican-American, % female, % prior preschool, mean income, mean household size, mean mother's education, teacher experience in Head Start, teacher certification, average staff working conditions, experience of teacher aide in HS and a dummy variable for EI/EK. For Analysis 2 all of the same covariates were used and the Stanford-Binet pre-test was added. For Analysis 3 the same covariates as Analysis 1 were used with the addition of the Motor Inhibition pre-test.

++Notes on Multivariate F-Tests.

1. In all instances the multivariates F for models were statistically significant beyond the .001 level.
2. In no instance was the multivariate F for the PV/ Comparison contrast statistically significant.
3. In all instances the multivariate F for interaction was statistically significant $.03 < p < .05$, though in no instance was a univariate F significant

+++In analysis 1 the estimated effect for the PV/Compa son contrast was statistically significant at the .05 level favoring the comparison group. Since, however, this effect is correlated with model effects and model effects were removed before it was the F-test for significance was less than the value required for statistical significance at the .05 level. It was significant at the $p < .08$ level.

* = statistically significant at the .05 level
** = statistically significant at the .01 level.

V. PV/Comparison group differences using Matched Samples

In Chapter V we described the procedure used to match Comparison classrooms with PV classrooms. Here we present data from two sets of analyses each on two matched samples of classrooms. In the first sample, matching was carried out on seven background and three pre-test variables. In the second sample matching was carried out using only the seven background characteristics (see Chapter V for a description of the variables).

The first set of analyses for each sample was carried out on the total final analysis sample of PV classrooms and their matched Comparison classrooms. Four dependent variables were used in this analysis; PSI post-test, Book 3D post-test, Book 4A post-test and Motor Inhibition post-test. The dependent variables are calculated by subtracting the Comparison classroom mean from its matched PV classroom mean. For the PSI, Book 3D and Book 4A, the sample is 101 matched classrooms divided among 26 sites. For the Motor Inhibition test, due to missing data, we have only 75 matched classrooms in 23 sites for the first sample and 76 classrooms in 25 sites for the second sample.* The second set of analyses were

*The matching was carried out without regard to whether the Motor Inhibition test had a valid pre- and post-score for both the PV and matched comparison classrooms. Thus the usable number of classrooms is somewhat less than the total possible number of PV classrooms with a valid pre- and post-Motor Inhibition score.

carried out only on the Level III sites. The dependent variables were the Book 3D, Book 4A and Stanford-Binet mean difference scores. For Stanford-Binet this gave us 61 classrooms in 16 sites and for the other variables we have 62 classrooms in 16 sites for this analysis.

For all analyses one covariate was used--the covariate was the pre-test score for the particular dependent variable being used. The covariate was calculated by subtracting the comparison classroom pre-test mean from its matched PV classroom mean. Furthermore, in each analysis three different levels of estimated reliability of the covariate were "corrected for" (1.00, 0.80 and 0.60). The rationale for "correcting" for the reliability of the covariate here and not in other analyses was that the procedure of taking a difference score of matched pairs of classrooms produces covariates which are substantially less reliable than the original aggregated means.*

*The procedure used to adjust the covariates for unreliability was the Lord-Porter (Porter, 1972) formula. Though this procedure produces the correct effect estimates it probably does not produce the correct standard error--it is probably a conservative estimate. By and large, however, we are less concerned with statistical significance than with the estimation of effects. We can estimate the reliability of a difference score using the following formula:

$$r_o = \frac{r_{aa} + r_{bb} - 2r_{ab}}{2(1 - r_{ab})}$$

where r_o = reliability of the difference
r_{aa} = reliability of PV scores
r_{bb} = reliability of Comp. scores
r_{ab} = correlation between matched PV and comparison classes

An example of this for the PSI in analysis 1 for the first sample we have:

$$r_o = \frac{.97 + .97 - 2(.822)}{2(1 - .822)} = \underline{\underline{0.8314}}$$

(continued on p. 153)

Table VI-7 presents data from the analyses on the two samples. Consider analysis 1 first. Here we have matched on the seven background variables and the three pre-test variables. Columns 1 and 2 show the classrooms and sites in the analyses. Columns 3 and 4 show the observed matched difference scores for the covariate (the mean difference between the comparison and matched PV classroom means ignoring the sites). By and large these differences--when compared with the standard deviations for the pre-tests taking the individual or the classroom as the unit of analysis--are small (see Tables VI-1 and VI-3). Only for the Stanford-Binet is there a difference in matched pre-test scores exceeding 0.10 standard deviations of the individual pre-test scores.

Given the correlations between the matched PV and Comparison post-tests shown in the last column of Table VI-7 and the estimated reliabilities given in Chpater II and Chapter V we estimate that all of the reliabilities of the covariates lie in the range of 0.60 to 1.0. Thus we have used three estimates (0.60, 0.80 and 1.00) of the reliability of the tests--in order to obtain some idea of the impact of the correction procedures. We should note that this approach to correcting for the reliability of the covariate (in addition to probably overestimating the standard error) ignores the critical problem of choosing an appropriate original reliability estimate--should we choose an internal reliability estimate, a test-retest estimate over what period of time, or a parallel forms reliability estimate again over what period of time? Our reason for ignoring the issue is that we have only one estimate of the reliability of the tests--an internal KR-20 estimate. Though we might have adopted Campbell and Erlenbachers approach of adjusting the reliability until the coefficient of the covariate was 1.0 this seemed inappropriate if we also present the overall "gain" scores--since this was all the procedure supplies us with.

As we point out in Chapter V the adjustment for the covariate takes place within each site around a covariate mean of zero. The estimated coefficient is taken from the pooled within regression of the dependent variable on the covariate.

Moreover, as we might have expected from previous analyses the mean differences between the PV and matched comparison post-test classroom means are not great, indicating that there is little difference in the effects of PV and comparison classrooms overall. Columns 5, 6, and 7 show the estimated differences between the PV and Comparison groups after covariance adjustment. Differences are shown for three levels of reliability (1.0, 0.80 and 0.60). The estimated mean differences were arrived at by pooling the unweighted adjusted means of the sites across all of the sites. Columns 8, 9 and 10 show the standard errors for the estimated differences.* Only one test in the sample 1 analyses, the Book 3D test in analysis 2, (Level III sites only) reaches statistical significance. The adjustment for the reliability of the covariate appears to do little to this estimate--it ranges from -0.4674 to -0.4950 favoring the Comparison group. This is a similar finding to that reported from the multivariate analysis of variance. In both instances we find one of the estimates of the Book 3D differences to be statistically significant at the 0.05 level, favoring the Comparison group, with a magnitude of roughly 0.50 points or one-sixth of the post-test standard deviation for individual children.

*The number of degrees of freedom for the estimates are equal to 1 and N-k-1 where N is equal to the number of classrooms, and K is equal to the number of sites. Thus for analysis 1 using the PSI the number of degrees of freedom are 1 and 74--to be statistically significant beyond the 0.05 level the ratio of the difference to the standard error has to be greater than 1.99.

Column 11 shows an F statistic for the test of homogeneity of regression slopes within the sites in the analysis. For one of the analyses using sample 1 the F is statistically significant at the 0.05 level indicating that the within coefficients for the separate sites are statistically different from one another, and therefore, that adjustment procedures may be inappropriate. Finally, column 12 shows the correlation between the matched classrooms on the particular pre-test--thus, for the PSI in analysis 1 the matching produced a correlation of 0.781 between the PV and Comparison classrooms.

The data for the second matched sample are presented in the second half of Table VI-7. Here, PV classrooms were matched with comparison classrooms on seven aggregate background characteristics. The format for this half of the table is the same as for the first half. There are, however, a few differences in results. First, note that the matching was much less effective here especially for the three pre-test variables that were included in the matching variables for sample 1. Thus, the two correlations for matched PSI classrooms for this sample are 0.29 and 0.48 while for the other sample they were 0.82 and 0.86. Second, note that by and large the pre-test mean differences (column 3) are very similar to the mean differences in sample 1. The chief exception to this is the PSI difference for the Level III sites (analysis 2). The difference between the PV and Comparison matched means is roughly 0.92 points for this sample and only 0.21 points for sample 1.

Looking now at the estimated differences, once again there are statistically significant differences favoring the comparison group for Book 3D in the Level III sites (analysis 2). All three of the estimated differences are statistically significant at the 0.05 level. A second set of consistently significant results occur for the Stanford-Binet, this time favoring the PV group. In contrast to the Book 3D effects, however, these estimates seem to be influenced a great deal by the the reliability "corrections" carried out on the covariate. Given the high estimated reliability of the Stanford-Binet on classroom means (roughly 0.94) and the low intercorrelation between the two groups of classrooms for the pre-test the internal reliability of the covariate is probably close to 0.85 (see formula in preceding footnote). This indicates that the best estimate of the differences is probably around 2.0 points or roughly one-sixth of the standard deviation of Stanford-Binet individual post-test scores. This estimate is roughly double the estimated difference found for the Stanford-Binet in the first sample analysis.

There are two other statistically significant differences in this Table. For both analyses on sample 2 the PSI difference, with 0.60 as the reliability estimate for the covariate, is statistically significant favoring the comparison group. These differences (of roughly 1.0 and 1.5 points) are between 0.10 and 0.20 standard deviations of the PSI post-test for individuals. The high reliability

TABLE VI-7

Selected statistics from a one-way nested analysis of covariance with matched PV and Comparison classrooms. Classrooms are the Unit of Analysis. The dependent variables are calculated by taking the differences between the means of PV and Comparison matched classrooms for the pre-tests. For each analysis the covariate pre-test difference corresponds to the post-test which is used to form the dependent variable. (Thus, for the PSI analysis the PSI pre-test difference between PV and Comparison matched classrooms is the covariate.) Three levels of reliability are assumed for the covariate. For each level the Lord-Porter correction is used.

Analyses for Sample 1 Using Matching on 10 Variables

Test	N Classrms.	N Sites	Pre-test Mean Difference	Post-test Mean Difference	PV/Comparison Differences Reliability Levels 1.0	0.80	0.60	Standard Error Reliability Levels 1.0	0.80	0.60	F-test for homogeneity	r bet. PV & Comp. Pre-tests
Analysis 1 (Levels II, III)												
Book 3D	101	26	-0.1468	-.096	-.1144	-.1184	-.1251	.1672	.1665	.1703	3.07***	.7610
Book 4A	101	26	.0104	.4591	.4639	.4640	.4641	.2515	.2515	.2514	0.78	.5065
PSI	101	26	.9462	.5782	.0553	-.0675	-.2721	.5268	.5362	.5561	0.75	.822
Motor Inhib.	75	23	- .0923	.0112	.0610	.0710	.0876	.0514	.0519	.0530	1.52	.274
Analysis 2 (Level III)												
Book 3D	62	16	-.1187	-.5335	-.4950*	-.4846*	-.4674*	.1986	.1995	.2015	1.08	.7917
Book 4A	62	16	-.1176	-.0094	-.00175	-.0147	-.0196	.3603	.3610	.3676	0.86	.4901
PSI	62	16	.2124	-.7903	-.9312	-.9537	-.9911	.7132	.7140	.7157	1.33	.8551
Stanford-Binet	61	16	-1.8492	.3039	.7085	.8267	1.0236	.8471	.8626	.8952	0.81	.4475

Analysis for Sample 2 Using Matching on 7 Variables

Test	N Classrms.	N Sites	Pre-test Mean Difference	Post-test Mean Difference	PV/Comparison Difference Reliability Levels			Standard Errors Reliability Levels			F-test for Homogeneity	r bet. PV & Comp. Pre-tests
					1.0	0.80	0.60	1.0	0.80	0.60		
Analysis 1 (Levels II, III)												
Book 3D	101	26	-.0884	-.1976	-.1382	-.1250	-.1029	.1630	.1631	.1634	1.76*	.1522
Book 4A	101	26	-.0698	.2559	.2736	.2822	.2965	.2857	.2857	.2858	1.71*	-.1458
PSI	101	26	1.1130	.1873	-.6462	-.8429	-1.1707*	.4944	.4977	.5040	2.08*	.2900
Motor Inhib.	76	25	.0135	.0610	.0590	.0568	.0533	.0510	.0511	.0511	0.84	.0762
Analysis 2 (Level III)												
Book 3D	62	16	.0913	-.3919	-.4486*	-.4688*	-.5024*	.2021	.2025	.2033	1.44	.3822
Book 4A	62	16	-.1149	-.0280	.0331	.0484	.0739	.3548	.3550	.3554	0.55	.0101
PSI	62	16	.9229	-.3566	-1.0614	1.2478	-1.5886*	.6689	.6736	.6836	1.48	.4813
Stanford-Binet	61	16	-1.5063	.9053	1.8638*	2.0985*	2.4898*	.9022	.9092	.9242	2.19*	.2930

* statistically significant at the .05 level
** statistically significant at the .01 level
*** statistically significant at the .001 level

of the PSI aggregate classroom scores and the relatively low correlation between matched PV and Comparison pre-test classroom means probably indicates that the "best" estimate of the reliability of the covariate is 0.80--a value which does not produce statistically significant differences between the PV and Comparison means.

Conclusions:

Our conclusions from this chapter are quite simple. By and large, we find no important differences between the PV and Comparison groups in their overall effects on the measured outcomes. We reach this conclusion in spite of the fact that in one analysis or another there are statistically significant differences between the two groups for each of the outcome measures. Consider the measures one at a time, for analyses of overall differences among the PV and Comparison groups.

(1) PSI: Though there were statistically significant results favoring the Comparison group in a few of the contrasts made in this chapter, the vast majority of the contrasts were not statistically significant. Nonetheless, many of the estimated differences between the groups were in the general area of 0.75 points favoring the comparison group. If we were to make a best bet on some real difference between the PV and comparison sample we would guess that the comparison children, on the average, outperformed the PV children by roughly 0.75 points, give or take a point. At the low end of this interval the difference is negligible--

at the high end it is roughly 20% of the standard deviation of the PSI post-test.

(2) Book 3D: This is the only test where a modest case can be made for a consistent difference favoring the Comparison groups. Although no differences occurred in the contrast of raw "gains" or in the regression analyses, statistically significant differences occurred in three of the six contrasts in the other two sets of analyses. All of the significant differences favored the comparison group with the modal difference being roughly 0.40 points or 14% of the post-test standard deviation for Book 3D. The largest estimated difference for this test was roughly 0.50 points favoring the comparison group and the smallest roughly a zero difference.

(3) Book 4A: Estimated differences between the PV and Comparison groups for this test range from 0.027 points favoring the comparison group to 0.44 points--a statistically significant difference favoring the PV group. In almost all instances the differences were very small and insignificant though they generally favor the PV group.

(4) Motor Inhibition: In no instance in an analysis of overall differences between the PV and Comparison group was a contrast statistically significant for this measure. By and large the majority of differences favor the PV group with the average difference being roughly 0.05 points or 8-10% of the post-test standard deviation.

(5) Stanford-Binet: This is the only test for which an argument can be made that the PV group outperformed the Comparison group. The maximum difference between the two groups occurred in the analysis of differences between observed "gains"--a difference of three points which is statistically significant beyond the 0.001 level. The bulk of this difference is accounted for by one sub-group--blacks with no prior preschool experience who will enter kindergarten "gained" ten points more in PV than in Comparison classes. No statistically significant differences occurred in the regression analyses, the multivariate analyses of variance or in the analyses of sample 1 of the matched data. For the analyses in sample 2 of the matched data statistically significant differences of roughly two points occurred favoring the PV group. With the exception of the multivariate analyses of variance where there was a difference of 0.422 points favoring the Comparison group all of the other differences favored the PV group with the average difference being roughly 1.0 points.

Chapter VII

SHORT-TERM EFFECTS OF ELEVEN HEAD START PROGRAM MODELS

Introduction:

This chapter contrasts the impacts of eleven Head Start curriculum models on five measured child outcome variables. Two broad analysis strategies are used: model effects are directly contrasted; and differential model effects are inferred by contrasting each model with a matching Comparison sample. The results of the analyses are presented in sections four through nine of this chapter: sections four through eight consider the five outcome measures separately and section nine summarizes the results by model. Before we present the results, however, we will consider some expectations we bring to this study and their implications for our analysis and interpretation of the data. Section one presents our broad expectations or hypotheses about the data. Section two discusses the issues of Type I and Type II error--of finding differences when, in fact, there are no differences, and of finding no differences when, in fact, there are differences. Section three outlines the procedures used for analysis and interpretation in the chapter.

1. Expectations about the Data:

During the past decade a massive amount of survey evidence has accumulated suggesting that existing variations in elementary and secondary school resources (including curriculum) bear little relationship to variations among children in their scores on standardized achievement and IQ tests. (See, for example, Jencks et al., 1972; Mosteller and Moynihan, 1972; Coleman et al., 1966; ISR study, in press; Children and Their Primary Schools, 1967; Racial Isolation in the Public Schools, 1967; Averch et al., 1972.) These works corroborate fifty years of experimental research which indicate that there are few differences among curricula in effectiveness. Reports from Follow Through Planned Variations also support the thesis that experimental manipulation of elementary school curricula produces roughly uniform effects on children's standardized achievement test scores (see SRI, 1972).

Work with preschool curricula has not been as extensive though the trend is the same. Weikart, 1970, for example, found that three different preschool curricula produced roughly equivalent short-term effects on children's test scores.* DeLorenzo, 1969, in a study of different preschool curricula, found few important short-term differences in effects. Other investigators (Karnes, 1968, for example) have found some evidence of differential impacts but her samples were small and her

*While Weikart found equivalent "gains", it should be noted that they were very large, supporting our argument of the overall short-term effect of preschool experience.

results may have reflected sample biases. Finally, the first year results from Head Start Planned Variations (SRI, 1971) seem to indicate that there are differential effects of types of preschool programs on children's test scores though the differences found were small and the investigators indicate that they may reflect uncontrolled biases in the samples.

On the basis of this past research, then, we did not initially expect to find many instances of differential curriculum effects on standardized tests. Two interpretations of previous findings are relevant to this study and supported our expectations of few differences. The first involves the degree to which different curricula actually alter the experiences of school children. The second stems from the limitations of standardized tests.

With regard to the first we stress that the finding that school or preschool variation in curriculum bears little relation to variations in outcomes does not mean that schools have no effect.

We find the data presented in Chapter IV about the effects of preschool versus no preschool to be convincing--as we find the argument that without school few children would learn long division. However, we also find compelling the argument that to a great extent preschools, no matter what their curricula, are strikingly homogeneous relative to the condition of no preschool. For almost all Head Start children the school year is roughly 150 days long, the "school" day is roughly four hours long, the adult child ratio is roughly eight to one, the environment is safe and pleasant and rich with opportunity. Moreover, most preschool teachers are warm, love children, and have a sense that they are important to the general well-being of all children. These facts suggest that the gross similarities among preschools <u>may</u> greatly overshadow their differences.

The relative homogeneity of preschool environments can help to explain the insensitivity of standardized tests to existing variations in curricula. Generally, standardized tests are required to have adequate psychometric properties for an entire tested sample. Thus, they must be appropriate to a wide range of individual differences among children and consequently be somewhat insensitive to subtle variations in experience. Moreover, standardized IQ tests like the Stanford-Binet are designed to measure stable traits which by definition only change under relatively extreme differences in conditions.

Overall, then, we should not be surprised that different preschools have similar effects on children's test scores. We suspect that, by and large, most standardized tests are affected by the gross experiences of children, and are little affected by relatively minor variations in style and strategy of teaching. This relative homogeneity does not mean that preschools are not important or that they do not have an effect on children. Rather, as assessed by standardized achievement and IQ tests, it suggests that they will have roughly equal impacts on children.

The authors of this report have been involved in evaluating the effects of schooling for the past five years. It would be misleading to assume that this experience has not influenced our expectations. In short, when we began the study we were very skeptical about whether different preschool curricula have different short-term or long-term impacts on standardized test scores. In part this skepticism has stayed with us. This led us to an overall initial expectation:

(1) We expected to find no differences among the curricula in their impact on standardized test scores. Put another way, we anticipated that the data would permit rejection of the null hypothesis of no differences among models.

Yet set against this skepticism are the experiences of trips by some of us to different sites, of discussions with sponsors, and the firm belief that variations in enivronment have an impact on children and adults. Our trips and our discussions have convinced many of us that there are relatively major differences in inputs among the models -- inputs as gross as materials and as subtle as different ways that adults relate to children. Much of the data for this conclusions is presented in the report on Implementation of Planned Variations, 1970-71. While that report points out that models are not as systematically different as we might have expected from talking with sponsors, it also gives evidence that the choice of model substantially influences the every day activities of children in Head Start centers.

If we acknowledge that there are systematic differences among models in their inputs and that differences among environments do affect child outcomes, we are but a short way from rejecting our initial expectation. The next step would be to argue that the environmental differences fostered by the variation in inputs among models bear a relation to standardized test scores.

Here our task is somewhat more difficult. For we have little idea how much variation in inputs is necessary to create differences in measured outputs. The only data we can draw on has been summarized in White et al., 1972. These data tentatively suggest that highly structured school programs using reinforcement principles might have some noticeable inpact on test scores over and above the gross inpact of school experiences. Moreover, White et al. found that no other differences in curriculum make a noticeable impact.

As we indicated in Chapter II, three models fit White's description of a potentially "effective" program (University of Oregon, University of Kansas and Pittsburgh). Other models, while varying in their emphasis on academic teaching, do not come close to placing the importance on structure and academic drill that these models do.

If we accept our general conclusion that most variations in curricula will have little differential effect on test scores and adopt as a possibility White's argument that emphasis on academic structure and drill might have an effect, we can formulate a second more tentative expectation:

(2) We expect to find no differences between the eight less academically-oriented Head Start models on short-term academic measures of output. Those three models, however, which stress academic drill and reinforcement principles, might appear more "effective" than the others on standardized measures of achievement.

Of perhaps more importance, however, is something implied by both expectations. We think it extremely unlikely that any of the models will be less effective than the comparison Head Start programs. By and large conventional Head Start programs do not have a structured academic emphasis. Their goals are broad and as Boyd (1966) noted, Head Start directors "reveal a preference for a supportive, unstructured, socialization program rather than a structured, informational program." In this sense they are similar to the eight PV models which do not place heavy emphases on academic drill. Following our earlier argument, therefore, we would expect to find no differences between their effects on achievement and the effects of the eight models which do not stress academic drill and reinforcement principles.

II. Type I and Type II Error

Given our general expectation that there will be few differences among models in effects on cognitive outcomes, we are inclined to be skeptical about rejecting a null hypothesis of no differences. This is particularly true if the data suggest that some models are "less effective" than their comparison classes. Our skepticism suggests a conservative strategy. It suggests that we should minimize the chance of Type I error, of finding differences when no differences exist. This strategy, however, has the disadvantage, given the fixed sizes of the samples, of maximizing the possibility of Type II error -- or not detecting differences when, in fact, they do occur.

While a conservative strategy is suggested by our expectations, it might not be fully justified. A convincing argument may be made that previous research into the effects of curriculum variation should not determine the strategy for analysis and interpretation in this study. It can be argued, for example, that no other study of preschool curricula has ever had as great a diversity of "treatments" as Planned Variation. And therefore, even though differences in "effectiveness" did not occur in other studies, there is no compelling reason to believe that they would not appear in this study. Moreover, perhaps researchers have an obligation to attempt to tease differences out of data rather than cover them over in the guise of a conservative strategy. It may be worth the chance of some Type I error to uncover potentially important patterns of results.

We are sympathetic to these arguments, and by and large they determined the approach taken in the companion report, "Cognitive Effects of Preschool Models on Different Types of Children." In this report, however, two factors other than prior research convinced us of the necessity for a conservative strategy. The first factor is the existence of unknown biases in the data. The second was the thought that it may be premature to take risks on labeling some models as "failures" or "successes" on the basis of a short-term evaluation.

The notion that there are uncontrolled differences among the samples which might lead to biased estimates of effects is not new to readers of this report. While biases can lead to either Type I or Type II error we felt that less damage of either a political or a substantive nature would result from a mistaken finding of no differences than from a mistaken finding of differences.

With regard to the prematurity of evaluation of certain models we are acutely sensitive that the theoretical bases of some of the approaches can lead to hypotheses of long-term "effects" on cognitive outcomes while not leading to especially overwhelming short-term outcomes. Models designed to improve the self-concept and increase the sense of control over environment of a child and thereby influence his later achievement may show dramatic long-range effects without showing short-term ones. The same holds for models that attempt to influence the way that children structure and attain information. While we do not expect these models to be less "effective" than most other models or than conventional Head Start classes, in the short run we feel that such a finding could lead to premature conclusions about their eventual "effectiveness".*

We may, therefore, be acting irresponsibly by including these models in our analyses. We can justify our actions only by the argument that we should have information about short-term as well as long-term effectiveness for all of the models. But although we think this justification compelling enough to permit our contrasting models, it is not compelling enough to permit a liberal approach to Type I error.

Two final points should be made. Both have to do with undiscovered differences among curricula. First, when we reject a null hypothesis of no differences among models, it means we think we have sufficient evidence to demonstrate that models do, in fact, differ. However, our acceptance of a null hypothesis does not mean that we have demonstrated that there are no differences (on the measured trait) among

*One assumption of "production function" research of the type being carried out here is that all of the different models are attempting to maximize the same measured outcomes. Clearly, the fact that some models are not attempting to maximize short term achievement while others are suggests that our data do not meet this assumption.

the models. It simply means that we do not have sufficient evidence to solidly demonstrate that there are differences. Second, as we have stressed throughout this report, our range of outcome measures is severely limited. Different model programs may well have substantially different impacts on outcomes unmeasured in this study.

III. Procedures for Analysis and Interpretation:

We use two strategies for minimizing Type I error.* First, data from a variety of analyses are displayed. In this way we hope to be able to determine which estimates are sensitive to different analytic approaches and which estimates are robust. Second, we focus attention on results which suggest that there are differences rather than on results which indicate no differences. For example, we will spend much more time trying to interpret why model A is significantly different from its comparison group than trying to interpret why model B is not significantly different.

The strategies will be used simultaneously. We want to be able to make interpretations of the following sort: "Although model A appears more effective than its on-site comparison group in the multivariate covariance analysis.

* The conventional way to minimize Type I error is to set the significance level for rejection of the model hypothesis at a very low value: e.g., .001. Since we are not secure about the use of tests of statistical significance with this data we have chosen other ways of minimizing Type I error.

this may reflect unmeasured differences in the samples indicated by the higher prescoresof children in the PV model group. This interpretation is strengtnened when we look at the results of the matched classroom analyses and see that model A is not significantly more effective than its matched comparison sample." Our conclusion from this would be that model A is probably not "more effective" than conventional Head Start on the particular measured outcome. If, however, we found that model A was significantly "more effective" in both analyses, our interpretation that unmeasured differences in samples account for the effects of the original analysis would be weakened and our conclusion might be that model A is indeed "more effective" than conventional Head Start.

We use estimates of statistical significance in the presentation of results in primarily an heuristic fashion -- as signals that something is going on in the data. We are more concerned with the magnitude of effect estimates and their consistency than with trying to indicate precise probability levels for differences, with data which probably do not conform to the assumptions for significance testing.

We present the data in the following manner. In sections four through eight of this chapter we consider the five outcome measures separately. Four tables are presented for each test. The first table shows mean prescores and gain scores for each of the PV and comparison sites in the final analysis sample. The means are unweighted averages of the classroom means. Tables 2, 3 and 4 for each outcome measure contain effect estimates. Four sets of analyses are represented. The first uses the child as the unit of analysis--two analyses are presented both contrasting the overall PV group within a model with its comparison group. One contrasts the PV and comparison groups on simple "gain" scores while the other makes use of the "expected" pre- and post-test scores described in Chapter IV. The "effect estimate" for the first is calculated by the following formula: [(PV post mean - PV pre-test mean) - (Comparison post-test mean - Comparison pre-test mean)]. This is a simple difference of differences. The second "effect estimate" is calculated in a more complex manner but the idea is similar. The estimate equals: ([(PV post-test observed mean - PV post-test expected mean) - (PV pre-test observed mean - PV pre-test expected mean)] - [(Comparison post-test observed mean - Comparison post-test expected mean) - (Comparison pre-test observed mean - Comparison pre-test expected mean)]). Another way of expressing this is to say that we are contrasting the differences between the PV observed and expected gains, and the Comparison observed and expected gains. The validity of each of these analyses rests on the comparability of the PV and Comparison groups within a model. Since models differ in this regard our confidence in the estimates will differ by model--for example, the Far West model's Comparison group might be thought of as quite dissimilar to the PV group since there is only one Comparison site for the two PV sites and it is "off-site". The "effect estimates" for these analyses may, therefore, not be valid for the Far West model. These data are presented in the second table for each outcome measure.

The second set of "effect estimates" stems from a direct comparison of PV classrooms within models in a multivariate analysis of covariance framework. Comparison sites are not included in this analysis.

The third set of analyses contrasts PV and Comparison classrooms. Two approaches are used. One treats all of the Comparison classes together " s a model" and tests for differences between the PV models and the overall "conventional" Head Start model in a linear regression framework. The other takes PV and Comparison groups by models using only those sites which contain both PV and Comparison classes. A multivariate analysis of covariance is used and one degree of freedom contrasts between PV and Comparison classes are carried out for each model. Data from the second and third sets of analyses are in the third table of each section.

The fourth table displays data from the analyses of "matched" PV and Comparison classes. We present observed and covariance adjusted differences by model for both the five and four factor analyses.

Our approach to interpretation of the data is straightforward.

First, we consider site to site differences in pretest means and gains for both PV and Comparison groups. The focus is on the distribution of the data--extreme cases and the range of scores between the 25th and 75th percentiles are indicated. Second, we consider model to model differences in PV and Comparison groups for both "observed gains" and for "observed-expected gains".

Third we consider the "adjusted" differences between groups from a variety of analyses. The approach is to point out patterns of results and to explain inconsistencies in the data. Finally we summarize the data for the test outcome

IV. Book 3D

A. Differences Among Sites:

Table VII-1 displays Book 3D pre-test scores and gains by site for both PV and Comparison groups. The range of PV pre-test means is from 8.8 to 14.8--roughly two individual level standard deviations (see Chapter III). The range of Comparison pre-test means is somewhat smaller--from 10.0 to 13.8. The middle fifty percent of the PV sites have a range of less than two points--from 10.94 to 12.75. The middle fifty percent of the Comparison sites also differ by less than two points ranging from means of 11.06 to 12.73.

PV site "gains" range from -0.24 to 3.86, roughly 1.4 individual standard deviations. The range for Comparison site "gains" is smaller, going from 0.54 to 3.44 points, or about one individual standard deviation. When we look at the middle fifty percent of the PV "gains" the range is only one point, from 1.94 to 2.95 points. The Comparison middle fifty percent range is roughly three-quarters of a point (2.09 - 2.86).

Since we aggregate sites within models and introduce variables as covariates, both techniques which generally lead to a reduction in differences, the site to site range is likely to be as large as we will see for Book 3D. For benchmark purposes, then, the largest difference in "gains" we can expect to find among groups

TABLE VII- 1

Book 3D

Pre-test means and mean "gains" (post-test mean - pre-test mean) by site for PV and Comparison groups. Site means are unweighted averages of classroom means.

Sponsor	Code	Community	Testing Level	PV Pre-test mean	Comp. Pre-test mean	PV "Gain"	Comp. "Gain"	PV classrooms (N)	Comp. classrooms (N)
Nimnicht	02.04	Duluth	III	12.90		1.94		4	
	02.04	St. Cloud	III		12.56		3.25		2
	02.13	Tacoma	II	11.91		2.97		4	
Tucson	03.08	LaFayette	III	13.81		2.60		4	
	03.08	Albany	III		13.65		3.10		4
	03.16	Lincoln	III	11.03		3.07		4	
Bank St.	05.01	Boulder	III	13.14	13.60	2.19	1.60	4	4
	05.11	Wilmington	II	9.91		0.65		4	
	05.11	DeLaWar	II		11.76		0.54		4
	05.12	Elmira	III	11.22	11.02	1.73	2.96	3	3
Becker & Engle-mann	07.03	E. St. Louis	III	12.33	11.89	2.26	3.44	4	4
	07.11	Tupelo	III	13.60	13.70	2.28	2.24	4	4
	07.14	E. Las Vegas	II	11.66		3.11		4	
	07.14	W. Las Vegas	II		11.57		2.20		4
Bushell	08.04	Portageville	III	10.31	10.01	2.54	2.07	4	4
	08.08	Mounds, Ill.	II	11.28	12.72	2.48	1.86	4	2
Weikart	09.02	Ft. Walton B.	III	8.85		3.61		4	
	09.02	Pensacola	III		11.06		2.19		3
	09.06	Greeley	III	12.75	12.73	1.03	2.11	4	3
	09.10	Seattle	II	11.51	13.51	3.86	2.55	4	3
Gordon	10.02	Jonesboro	III	12.89	13.83	2.17	2.09	3	3
	10.07	Chattanooga	III	11.29	12.27	3.70	2.86	4	4
	10.10	Houston	II	9.76	10.15	1.52	2.58	4	4
EDC	11.05	Washington	III	12.22	10.83	-0.24	1.10	4	4
	11.06	Paterson	II	10.56	12.48	2.10	2.08	3	4
	11.08	Johnston Co.	III	13.49	12.50	2.48	3.36	4	4
Pitts-burgh	12.03	Lock Haven	III	10.40		2.95		4	
	12.03	Mifflenburg	III		10.78		2.43		4
NWL	20.01	Kansas City	III	10.94		1.84		4	
Enablers	27.04	Billings	II	11.83		2.12		4	
	27.05	Colorado Sp.	II	11.91		2.23		4	
	27.03	Bellows Falls	II	11.91		1.95		4	

in our analyses is roughly 1.4 individual standard deviations. When PV sites are looked at within models the effect of aggregating becomes apparent. The model with the site "gaining" the least is EDC. Yet both of the other EDC sites yield "gains" in the middle fifty percent range. And the model containing the site with the largest "gain" (High/Scope) also includes a site whose "gain" falls in the bottom twenty-five percent. No model has more than one site in the top twenty-five percent and only one model (Bank Street) has two sites in the bottom twenty-five percent. The variation in "gains" within models, therefore, makes model to model differences considerably smaller than site to site differences.

B. Model to Model Differences:

Columns 1 and 2 in Table VII-2 clearly makes this point.* PV model mean "gains" in this table vary from 1.49 points to 2.93 points--a spread of roughly 0.5 individual standard deviations. Comparison "model" mean "gains" also have a range of roughly 0.5 individual standard deviations, from 1.62 to 3.09 points.

When differences between PV and Comparison groups within model are considered the paucity of large differences

*In Table VII-1 classroom means were aggregated to yield site means. In Table VII-2 individual scores were aggregated to yield model means. Thus, there may be some small seemingly inconsistent results if the reader compares the two tables. The inconsistencies result from unequal numbers of children in classrooms and sites.

TABLE VII- 2

Model Statistics for the Book 3D Test

Column 1 shows the mean gain for PV children in the model.
Column 2 shows the mean gain for Comparison children in model location.
Column 3 shows the difference between Column 1 and Column 2. (A positive score indicates that PV children gained more than Comparison children).
Column 4 shows the difference between PV and Comparison children in observed-expected gains.
The individual is the unit of analysis.[1]

Model	PV "Gains"	Comparison "Gains"	PV "Gains"-Comparison "Gains"	PV (observed-expected) "gains"-comparison (observed-expected) "gains"

Far West Laboratory	SD=2.94
	2.46
	N=69
		[illegible]
		3.09
		46
			-0.62	-.73

Arizona	2.41 2.80 132	2.44 2.89 61	-0.08	-.91*
Bank St.	3.13 1.49 121	3.28 1.62 96	-0.13	-0.05
U. of Oregon	2.90 2.49 180	2.74 2.66 168	-0.16	-0.00
U. of Kansas	3.20 2.51 106	2.36 1.93 61	0.58	0.56
High Scope	2.91 2.93 122	2.66 2.20 96	0.73	0.92
U. of Florida	3.31 2.68 111	2.53 2.54 123	0.14	0.37
EDC	2.96 1.95 138	2.82 2.42 123	-0.47	-0.54
U. of Pittsburgh	3.63 2.93 42	3.15 2.52 31	0.41	0.62
REC	3.35 1.88 49			
Enablers	2.29 2.14 118			

* Statistically significant at the .05 level
** Statistically significant at the .01 level
1. All children in the basic analysis sample were used (see Chapter III)

becomes even clearer. Column 3 shows differences between PV and Comparison group "gains" within models. The differences in "gains" range from -0.62 point avoring a Comparison group) to 0.73 points (favoring a PV model). None of the differences are statistically significant. Moreover the largest difference favoring the Comparison group may be due to an inappropriate match between PV and Comparison group since it arises in a model where there were two PV sites and only one Comparison site (an "off-site" Comparison at that).

Column 4 shows the differences between "Observed and Expected" mean gains for the PV and Comparison groups within models. The range of differences is from -0.91 favoring the Comparison group to 0.92 favoring the PV group. In this instance, the introduction of control variables increased the spread of differences. One of the differences is statistically significant (for the Arizona model favoring the Comparison group). Interpretation of this significant finding is difficult, however, since it may be due to inappropriate Comparison groups. For only one of the two PV sites in the Arizona model is there a Comparison site and it is an off-site Comparison. Moreover, one of Arizona's PV sites had "gains" in the top twenty-five percent of the site gain distribution while the other had "gains" in the middle fifty percent. The "off-site" Comparison group had the fourth largest "gain" of the Comparison groups ending up with an "observed" post-test mean of 16.75, roughly three points above its "expected" post-test score. This was the largest difference between "observed" and "expected" post-test site means. It seems, then, that in this analysis the Arizona PV sites had the misfortune of being contrasted with one particularly effective Comparison site. To conclude from this data that the Arizona model is not as "effective" as other models or as most conventional Head Start programs would be a mistake.

Another way of looking at these data suggests that there might be larger differences among the models than we have indicated. If, for example, we contrasted the differences between the most "effective" PV model (judged by contrasting it with its Comparison group) with the least "effective" PV model, we obtain a difference of 1.83 points (0.92 - -.91). But even this difference is only roughly .60 standard deviations--not trivial but certainly not overwhelming given the methodological problems associated with the analyses.

C. "Adjusted" Differences among Groups*

Tables VII-3 and VII-4 show data from a variety of

*We will use the terms "adjusted differences", "estimated effects", and "effects" interchangeably in this chapter. The context of the discussion should make our meaning clear. We will discuss the results of the Book 3D analyses in a more comprehensive fashion than they deserve--by doing so we hope to be able to familiarize the reader with some of the analyses used for each of the five outcome measures.

analyses of the Book 3D data. The results of 108 contrasts are presented. Five contrasts yield statistically significant results. Four different types of analyses are represented in the 108 contrasts.* A complete summary of the data would produce a litany of "small and insignificant differences". Instead of going that route we will briefly interpret the four sets of analyses and then point out a few patterns in the data.

(1) Column 1 of Table VIII-3 shows the "adjusted" differences between PV model means and an unweighted overall PV grand mean. The technique used was a one-way multivariate analysis of covariance. The overall univariate F for Book 3D is 0.8744, indicating that the differences among the models could easily have occurred by chance, if there are no "true" differences. None of the "effect estimates" are statistically significant.

(2) Columns 2 and 3 show "adjusted differences" between model means and an overall pooled Comparison group mean. The technique used was regression analysis. In column 3 a single set of covariates was used. None of the models showed statistically significant different means from the pooled Comparison mean. In column 4 separate slope coefficients for two covariates (PSI pre-score and Percent Prior Preschool Experience) were allowed for the different models. Three separate slope coefficients entered with statistically significant coefficients indicating

*We do not want to give the impression that the 108 contrasts are all independent. Clearly they are not. Our purpose in reporting the number of statistically significant contrasts is simply to give the reader an overall first impression of the data.

TABLE VII - 3
Book 3D

Model "effect" estimates for the test. Columns 1-4 show differences between "adjusted" PV model means and some standard. Column 1 shows the simple contrasts between the PV model "adjusted" means and an unweighted grand mean of the model means for an exact least squares one way ANCOVA. Columns 2 and 3 show regression coefficients for each model in an analysis where all of the comparison classes are pooled together to form a comparison "model". The regression coefficients can be thought of as representing the difference between the "adjusted" PV model means and the "adjusted" Comparison "model" means. Column 2 shows the coefficients for a regression analysis not allowing for separate slope coefficients for the covariates for the different models. Column 3 shows the coefficients allowing for separate model coefficients for the PSI pre-test and for percent prior preschool. Column 4 shows the difference between PV and Comparison group "adjusted" means within models for sites with both a PV and a Comparison group. The estimates are 1 degree of freedom contrasts in the framework of a one way ANCOVA design. Column 5 shows the PV and Comparison n's for column 4 analysis. A note following the Table lists the covariates used in the analysis. In all analyses the classroom is the unit of analysis. See text (Chapters V and VII) for further discussion of the approaches.

Model	Estim.effects around PV unweighted mean	Estimated effects of PV models against pooled compar. classes[2]		DF contrast PV v. site comp.pooled by models[3]	PV N	Comp. N
		analysis 1	analysis 2			
Far West Laboratory	0.46 n=8	0.09	0.12	-1.0	4	2
Arizona	0.12 8	0.06	0.05	-1.53	4	4
Bank St.	-0.41 11	-0.48	0.34	-0.01	11	8
U. of Oregon	0.77 12	-0.03	0.38	0.01	12	12
U. of Kansas	-0.03 8	-0.62	-0.70	-0.27	3	6
High Scope	-0.03 12	0.17	0.10	-0.10	12	9
U. of Florida	-0.33 11	-0.68	-0.79*	-0.71	11	11
EDC	-0.63 11	-0.71	-0.77*	-0.73	11	9
U. of Pittsburgh	0.44 4	0.25	0.42	0.10	4	4
REC	-0.41 4	0.05	0.28			
Enablers	0.04 12	-0.23	0.05			
Grand Mean	13.97	14.18	14.18	14.29		

TABLE VII-3

(Page 2)

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

1. Only PV classrooms are included in this analysis. The multivariate F with the PSI, Book 3D, and Book 4A in the analysis is 2.36; significant at the .001 level. The overall multivariate F for Book 3D is 0.87, which is not statistically significant.

2. Both analyses were in the regression framework with the pooled Comparison classrooms as the "dummy variable" left out of the regression. Analysis 1 did not contain separate slope coefficients for the various models. Analysis 2 allowed for separate slope coefficients for PSI pre-score and prior preschool experience. Analysis 1 explained 71.3% of the total variation; analysis 2 explained 76.0% of the total variation.

3. Only sites with both PV and Comparison classrooms (on or off-site) were included in this analysis.

Note: All analyses included the following covariables: PSI pre-test mean, Book 3D pre-test mean, Book 4A pre-test mean, mean age, percent black, percent Mexican-American, percent female, mean income, mean household size, teacher experience in Head Start, teacher certification, mean mother's education, percent prior preschool, average staff working conditions, whether the site is El or Ek. In the analyses in column 1 the variable "site administered by CAP or by Public School was also included. In the regression analyses in columns 2 and 3 teacher race was included. In analyses of the Stanford-Binet, the Stanford-Binet pre-test was also included as a covariate--these analyses used only Level III sites. In analyses of the Motor Inhibition only classrooms with valid Motor Inhibition scores for both fall and spring were included.

that for some models the relationship between the covariates and Book 3D differed from the overall pooled relationship. The effect of the separate slope coefficients was to change some of the "adjusted differences" between the model means and the pooled Comparison mean. Specifically separate slope coefficients for PSI pre-score entered for Bank Street, University of Oregon and for the Enablers model--in each of these cases the "adjusted difference" for the model changed rather substantially though in none of the three cases did it reach statistical significance. For two other models, however, the effect of the new covariates slightly changed their relationship to the Comparison group, shifting them from a non-significant status in the column 3 regression to a statistically significant magnitude in the column 4 analysis. The University of Florida and EDC models have "adjusted means" significantly smaller than the "adjusted Comparison" mean at the .05 level.

(3) Column 4 of Table VII-3 shows one degree of freedom contrasts between PV and Comparison model means adjusted for covariates. The sample represents only those PV and Comparison classes in sites with both PV and Comparison classes.* None of the contrasts are statistically significant.

*Thus, the contrast between the Arizona model mean and its Comparison mean uses only the PV classes in Lafayette and the Comparison classes in Lafayette's off-site Comparison, Albany.

(4) Tables VII-4A and 4B show data from matched PV and Comparison classroom analyses. In Table VII-4A none of the contrasts between PV model means and their matched Comparison class means is statistically significant (see columns 4-6). In Table VII-4B one set of contrasts shows significant results. In these analyses it appears as if the High/Scope model is significantly more effective than its Comparison classes. No other contrasts are statistically significant.

The very small percentage of statistically significant contrasts in these tables, the lack of any robust "effects" (models that show significant results in a variety of analyses), and the overall similarity of the observed gains for the different PV models and Comparison groups, suggest that there are few important differences among the Head Start curricula in their effects on the Book 3D test.

There are, however, a number of patterns in the data which can be reported.

(1) First, we find no data to support our "tentative" expectation about the special effectiveness of highly structured, academically oriented models. No contrast involving these models showed statistically significant results. Although in all three of the models (University of Kansas, University of Oregon and University of Pittsburgh), all PV sites fell in the middle fifty percent or upper

TABLE VII-4A

Selected Statistics for Matched Classroom Analysis of Book 3D for the 5 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

				"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.60
Far West Laboratory	8	-0.25	-0.09	0.06	0.05	0.04
Arizona	8	-0.48	-0.08	-0.13	-0.15	-0.17
Bank St.	11	0.22	-0.47	-0.44	-0.44	-0.43
Univ. of Oregon	12	0.13	0.34	0.36	0.36	0.37
Univ. of Kansas	8	-0.61	0.01	-0.05	-0.07	-0.10
High Scope	12	-0.65	0.33	0.25	0.23	0.20
Univ. of Florida	11	0.04	-0.51	-0.51	-0.51	-0.50
EDC	11	-0.05	-0.47	-0.48	-0.48	-0.48
Univ. of Pittsburgh	4	-0.83	-0.21	-0.30	-0.33	-0.36
REC	4	-0.28	-1.15	-1.19	-1.19	-1.21
Enablers	12	0.34	0.38	0.41	0.42	0.44

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measured = .781. The overall F for the test of homogeneity of the covariate regression coefficient = 3.07.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = -0.11; with r_{tt} estimated as 0.80 the coefficient = -.14; for r_{tt} = 0.60, the coefficient = -.18.

TABLE VII-4B

Selected Statistics for Matched Classroom Analysis of Book 3D for the 4 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Tes.) (Adjusted for Pre-Test Covariance) Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.6
Far West Laboratory	8	0.65	1.28	0.91	0.82	0.67
Arizona	8	0.01	-0.13	-0.13	-0.13	-0.13
Bank St.	11	-0.62	-1.07	-0.72	-0.63	-0.48
Univ. of Oregon	12	0.58	0.47	0.14	0.06	-0.08
Univ. of Kansas	8	-0.35	0.23	0.39	0.44	0.51
High Scope	12	0.10	1.26	1.20**	1.18**	1.16**
Univ. of Florida	11	-1.06	-1.44	-0.81	-0.66	-0.41
EDC	11	-0.30	-0.94	-0.77	-0.73	-0.66
Univ. of Pittsburgh	4	-1.67	-1.38	-0.43	-0.20	0.20
REC	4	-1.82	-2.26	-1.57	-1.40	-1.11
Enablers	12	1.20	0.33	-0.36	-0.53	-0.81

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.15. The overall F for the test of homogeneity of the covariate regression coefficient = 1.76.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.57; with r_{tt} estimated as 0.80 the coefficient = 0.71; for r_{tt} = 0.60, the coefficient = .94.

twenty-five percent in terms of gains, their Comparison groups did almost as well--only one Comparison site of six for this group of models fell in the bottom twenty-five percent in terms of raw "gains". When we look at model level data we find that, on the average, the Kansas and Pittsburgh PV classes gained slightly more than their Comparison groups in terms of both "observed" gains and "observed-expected" gains, while the University of Oregon model only held its own with its Comparisons. This pattern, however, does not hold for the different analyses. Oregon, for example, appears above average in the analysis directly comparing models, equal to or slightly above average in the comparisons with the pooled Comparison classes, almost exactly average in the one degree of freedom contrasts with Comparison classes in the same sites, slightly above average in the five factor matching analyses and just about average in the four factor matching analyses. The estimates of the effectiveness of the Kansas model vary somewhat more. In the direct contrasts among models Kansas appears average, in the analyses with the pooled Comparison classes it appears below average, in the contrast with Comparison groups in the same sites it is slightly below average, in the five factor match it is roughly average and in the four factor matched analyses it is slightly above average. Pittsburgh appears slightly above average in the direct contrasts among PV classrooms and in the contrasts against the pooled Comparison classrooms, about average in the one degree of freedom contrast with its Comparison site, and slightly below average in the analyses of the two matched sample analyses. Certainly these data offer no support for our tentative expectation.

(2) Four other models, (Far West Laboratory, Arizona, Bank Street and the Enablers) show mixed patterns of results. With the exception of the Arizona model there are no statistically significant contrasts for any of these models--the one significant result for Arizona was discussed earlier. On the basis of these data we see no reason to argue that any of these models differ from the average in effectiveness on the Book 3D test.

(3) Two other models (High/Scope and REC) also show mixed patterns of results but in each instance some set of contrasts is of sufficient magnitude to deserve attention. The High/Scope model is particularly interesting. Three PV sites are included in the analyses. On the average, the pre-test means for the High/Scope PV sites are below the overall Comparison mean, the mean for their own on location Comparison sites,and the overall PV mean. Two of the sites have the largest observed gains in the sample while the third site ranks in the bottom twenty-five percent of observed gains. All three Comparison sites for this model have average "observed" gains.

In the comparison of "observed" gains the High/Scope

PV sites come out looking somewhat better than their Comparison sites even though there is one "weak" PV site. However, when "adjusted post-test" means for the PV sites are contrasted to either (a) their own Comparison site "adjusted post-test means", (b) to the overall Comparison group "adjusted post-test mean" or (c) to the overall PV "adjusted post-test mean", the model appears only average. In short, the "adjustments" do not compensate entirely for the initially low scores. This may be appropriate and the estimated effects may be unbiased. If we had complete faith in our "adjustments" we would judge the High/Scope model to be of only average effectiveness. The results of the matched analyses, however, suggest that we may be "underadjusting". In both of the matched analyses the High/Scope PV classes look somewhat better than their matched Comparison classes. In the five factor sample analyses the differences are small but in the four factor match they are large and statistically significant. Our inclination in this case is to equivocate--the High/Scope model may be more effective than average but our data is not strong enough to be convincing.

The situation for the REC model is also ambiguous. Here our basic problem is that there is only one PV site and no Comparison site. The one PV site, however, scores in the bottom twenty-five percent of the sites in terms of observed "gains" giving the model the second lowest model "observed" gains. When contrasted with the overall

PV "adjusted post-test mean" and the overall pooled Comparison "adjusted post-test mean" the REC site looks about average. In the matched analyses, however, the REC site comes off looking very badly--showing differences favoring the Comparison group of roughly 0.33 to 0.50 individual standard deviations. Since there is only one site none of the contrasts are statistically significant although they are clearly out of the ordinary.

(4) The final two models (EDC and Florida) each show consistent patterns of results. All estimates for EDC suggest that it is somewhat less effective than the other models and the Comparison Head Start groups--the differences are all within the range of -.47 to -.77 points. As we noted earlier this entire observed effect seems to be due to one outlying EDC PV site. In this site the children actually appear to have "lost" information (their average "gain" was -0.24). Two things should be noted about this site. First, of an original 85 children in the site only twenty were included in the final analysis sample--for whatever reason the remainder were excluded (see Chapter III for possible reasons). This gives us only an average of four children per class. Second, according to the OCD consultant this site underwent great turmoil during the school year. The turmoil was perceived as having a substantial effect on the teachers, advisory staff and on the children. Taken together these two factors suggest to us that the data from this site should

TABLE VII- 5

Book 4A

Pre-test means and mean "gains" (post-test mean - pre-test mean) by site for PV and Comparison groups. Site means are unweighted averages of classroom means.

Sponsor	Code	Community	Testing Level	PV Pre-test mean	Comp. Pre-test mean	PV "Gain"	Comp. "Gain"	PV classrooms (#)	Comp. classrooms (#)
Nimnicht	02.04	Duluth	III	6.64		3.10		4	
	02.04	St. Cloud	III		6.86		2.83		2
	02.13	Tacoma	II	6.13		3.87		4	
Tucson	03.08	LaFayette	III	5.18		6.10		4	
	03.08	Albany	III		6.29		3.69		4
	03.16	Lincoln	III	5.39		4.63		4	
Bank St.	05.01	Boulder	III	7.40	5.80	2.60	1.60	4	1
	05.11	Wilmington	II	4.04		0.86		4	
	05.11	DeLaWar	II		5.14		1.13		4
	05.12	Elmira	III	6.03	5.39	2.35	5.08	3	3
Becker & Engle-mann	07.03	E. St. Louis	III	5.89	5.36	3.99	5.98	4	4
	07.11	Tupelo	III	5.88	6.96	6.35	2.78	4	4
	07.14	E. Las Vegas	II	5.55		5.08		4	
	07.14	W. Las Vegas	II		5.85		3.32		4
Bushell	08.04	Portageville	III	4.77	4.21	6.17	1.49	4	4
	08.08	Mounds, Ill.	II	5.46	5.78	5.99	3.19	4	2
Weikart	09.02	Ft. Walton B.	III	3.68		0.52		4	
	09.02	Pensacola	III		5.68		2.50		3
	09.06	Greeley	III	6.80	6.76	-2.96	3.25	4	3
	09.10	Seattle	II	6.38	9.44	4.02	3.52	4	3
Gordon	10.02	Jonesboro	III	6.05	6.72	3.55	4.20	3	3
	10.07	Chattanooga	III	3.17	6.18	7.85	5.94	4	4
	10.10	Houston	II	4.78	4.39	1.75	2.03	4	4
EDC	11.05	Washington	III	5.27	5.22	-0.14	0.62	4	4
	11.06	Paterson	II	5.60	6.92	2.30	-0.16	3	1
	11.08	Johnston Co.	III	5.79	5.33	7.00	7.94	4	4
Pitts-burgh	12.03	Lock Haven	III	5.61		3.43		4	
	12.03	Mifflenburg	III		4.14		2.59		4
RLC	20.01	Kansas City	III	4.89		2.68		4	
Enablers	27.04	Billings	II	4.63		1.78		4	
	27.05	Colorado Sp.	II	6.09		1.39		4	
	27.03	Bellows Falls	II	8.01		4.58		4	

not be considered as representative of the effectiveness of the EDC model.

The Florida model also shows a fairly regular pattern of negative results with the exception of the "observed gains" where it appears to do as well or slightly better than its Comparison group. In all of the other analyses, however, the Florida PV model seems to be somewhat less effective than the other PV models and than Comparison classes. In only one instance, however, does a contrast reach statistical significance and the contrast differences while consistent, never exceed roughly thirty percent of an individual standard deviation.

In summary, two models (Florida and REC) may be very slightly less effective than the other PV models on the Book 3D test, though the data are not strong enough to be convincing. One model (High/Scope) may be more effective though again we are unconvinced. We detect no differences among the other eight models.

V. Book 4A

A. Site to Site Differences:

Planned Variation sites range in pre-score means from 3:17 to 8.01 points on the Book 4A test, roughly 1.6 individual standard deviations. The middle fifty percent ranged from pre-test means of 5.18 to 6.05 or less than thirty-three percent of an individual standard deviation. Comparison sites had pre-test means ranging from 4.14 to

9.44, roughly 1.75 standard deviations. The middle 50% of the comparison pre-test means ranged from 5.39 to 6.18 or less than thirty-three percent of a standard deviation. Thus, though it appears as if there is a wide range of variation in pre-test means the bulk of the sites fall in a very narrow range.

"Observed gains" behave in a somewhat different way. The range of PV gains is from -0.14 to 7.85 while the range of Comparison site "observed gains" is from -0.16 to 7.94 points--each range representing roughly 2.4 individual standard deviations. The middle fifty percent of PV "gains" ranges from 2.30 to 5.08, roughly 0.90 standard deviations while the middle fifty percent range of Comparison "gains" is from 2.03 to 3.69, roughly 0.6 standard deviations. On the surface, these differences suggest that some PV and Comparison sites may differ greatly in their effectiveness in imparting knowledge of letters, numbers and shape names with the average PV sites appearing slightly more effective.

When sites are examined within models the differences attenuate as they do for the Book 3D test. The model with the lowest "gaining" site is again EDC (the same site which showed the negative "gains" on the Book 3D test). The results from this site will be discounted for the reasons given earlier. EDC also has the site with the second largest "gains" though it should be noted that the on-site Comparison classes for this site had the largest average gains of any of the PV or Comparison sites. Only one PV model (the Enablers) has two sites in the bottom twenty-five percent of average "gains." Finally, one model (Kansas) has each of its two sites in the top twenty-five percent of average "gains."

B. Model to Model Differences

Table VII-6 shows model to model differences in "gains" for the PV models and their Comparison sites. The range in "observed gains" for models for Book 4A is considerably larger than for Book 3D. The PV model showing the smallest "gains" is Bank Street (1.88 points) while the largest "gains" are made by the Kansas model (6.06 points). This is a difference of roughly 1.4 individual standard deviations. A similar range exists when PV model "observed gains" are contrasted to their Comparison groups' "observed gains." Here the range is five points, from -0.85 points, favoring the Comparison group from Bank Street, to 4.19 points, favoring the Kansas PV model. Three of the contrasts show statistically significant results favoring the PV groups (the Arizona, U. of Oregon and U. of Kansas models). Similar results occur in the contrasts between PV and Comparison "observed-expected gains." Again the range is roughly five points and the same three PV models show favorable statistically significant results. In neither set of contrasts in Table VII-6 is a Comparison set of classrooms significantly more effective than the PV model classrooms.

TABLE VII- 6

Model Statistics for the Book 4A Test

Column 1 shows the mean gain for PV children in the model.
Column 2 shows the mean gain for Comparison children in model location.
Column 3 shows the difference between Column 1 and Column 2. (A positive score indicates that PV children gained more than Comparison children).
Column 4 shows the difference between PV and Comparison children in observed-expected gains.
The individual is the unit of analysis.[1]

Model	PV "Gains"	Comparison "Gains"	PV "Gains"-Comparison "Gains"	PV (observed-expected) "gains"-comparison (observed-expected) "gains"

Far West Laboratory	SD=3.91
	3.60
	N=67
		3.15
		2.91
		46
			0.68	0.86

Arizona	4.16 5.33 132	3.71 3.56 61	1.78**	1.43*
Bank St.	3.72 1.86 117	3.70 2.71 94	-0.85	-0.71
U. of Oregon	4.06 5.40 182	3.62 4.08 168	1.32**	1.39***
U. of Kansas	4.01 6.06 105	3.39 1.87 61	4.19***	4.20***
High Scope	3.90 2.39 121	4.12 3.15 96	-0.76	-0.65
U. of Florida	4.66 4.64 110	4.36 4.72 123	-0.08	-0.03
EDC	4.58- 4.25 138	5.19 4.19 123	0.06	0.09
U. of Pittsburgh	3.47 3.48 42	3.42 2.39 31	1.09	1.23
REC	3.60 2.71 49			
Enablers	3.19 2.50 115			

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

[1] All children in the basic analysis sample were used (see Chapter III)

C. "Adjusted Differences Between Groups"

The patterns of results for the Book 4A test are considerably clearer than for the Book 3D test. Of the 108 contrasts made in Tables VII-7 and VII-8, twenty are statistically significant. Ten of the twenty statistically significant differences occur for one model (U. of Kansas); the other ten statistically significant differences are scattered over four models. Four patterns stand out in the data.

1) The University of Kansas model appears to be considerably more effective than the Comparison classes and than the other models in imparting information tested by the Book 4A test. The Kansas model, in every analysis, has the highest estimated "effect." Its average "observed gain" is roughly 0.75 points higher than the next nearest model. It exceeds its Comparison groups by over four points in both "observed" and "observed-expected" gains. In the direct contrast between models its "adjusted effect" exceeds the overall PV mean by 2.71 points. The smallest estimated effects for this model occur in the regression analyses contrasting PV model "adjusted means" with the overall pooled Comparison mean--these "estimated effects" are roughly 2.23 points. In the other analyses the range of "estimated effects" is from 3.11 to 3.88 points. There seems to be little question but that the Kansas model was more effective than the average of the other models and

than the Comparison classes in 1970-71 for the Book 4A test. The effect seems to be on the order of 0.70 to 1.3 individual level standard deviations.

2) Both the University of Oregon and the University of Pittsburgh models show positive "estimated effects" in all statistical contrasts. Though statistically significant in only a few instances, the pattern of effects, together with the University of Kansas finding, strongly suggest that the highly structured and academically oriented models are somewhat more successful than the Comparison classes and than most of the other models in imparting to children knowledge of letters, numbers, and shape names.

3) None of the other models consistently have positive "estimated effects," though Far West and Arizona each exceed their comparison groups in the matched classroom analyses by a substantial margin. The analyses in Table VII-7 indicate, however, that Far West and Arizona have only average effectiveness.

4) Two models show moderately consistent patterns of negative "effects" (REC and Enablers). Two of the Enabler sites had "observed gains" in the bottom quartile of site "gains." When contrasted with the overall PV mean, the "adjusted mean" for the Enabler model was roughly one point lower. In contrast to the overall Comparison group they were significantly different, with an effect of roughly -1.4 points

TABLE VII-7
Book 4A

Model "effect" estimates for the test. Columns 1-4 show differences between "adjusted" PV model means and some standard. Column 1 shows the simple contrasts between the PV model "adjusted" means and an unweighted grand mean of the model means for an exact least squares one way ANCOVA. Columns 2 and 3 show regression coefficients for each model in an analysis where all of the comparison classes are pooled together to form a comparison "model". The regression coefficients can be thought of as representing the difference between the "adjusted" PV model means and the "adjusted" Comparison "model" means. Column 2 shows the coefficients for a regression analysis not allowing for separate slope coefficients for the covariates for the different models. Column 3 shows the coefficients allowing for separate model coefficients for the PSI pre-test and for percent prior preschool. Column 4 shows the difference between PV and Comparison group "adjusted" means within models for sites with both a PV and a Comparison group. The estimates are 1 degree of freedom contrasts in the framework of a one way ANCOVA design. Column 5 shows the PV and Comparison n's for column 4 analysis. A note following the Table lists the covariates used in the analysis. In all analyses the classroom is the unit of analysis. See text (Chapters V and VII) for further discussion of the approaches.

Model	Estim.effects around PV unweighted mean	Estimated effects of PV models against pooled compar. classes[2]		DF contrast PV v. site comp.pooled by models[3]	PV N	Comp. N
		analysis 1	analysis 2			
Far West Laboratory	-0.54 n=8	-0.19	-0.51	.02	4	2
Arizona	-0.85 8	0.13	0.03	-.28	4	4
Bank St.	0.24 11	0.09	1.10	-.06	11	8
U. of Oregon	1.76* 12	0.59	1.07	0.95	12	12
U. of Kansas	2.71*** 8	2.22**	2.24**	3.11**	8	6
High Scope	-2.40*** 12	-0.92	-0.84	-0.47	12	9
U. of Florida	-0.67 11	-0.73	-1.26	-0.63	11	11
EDC	-0.17 11	-0.00	-0.08	-1.22	11	9
U. of Pittsburgh	2.21 4	1.75	1.77	1.45	4	4
REC	-1.37 4	1.10	1.42			
Enablers	-0.92 12	-1.69**	-1.37*			
Grand Mean	9.23	9.23	9.23	9.25		

TABLE VII-7

(Page 2)

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

1. Only PV classrooms are included in this analysis. The multivariate F with the PSI, Book 3D, and Book 4A in the analysis is 2.36; significant at the .001 level. The overall multivariate F for Book 4A is 3.70, significant at the .001 level.

2. Both analyses were in the regression framework with the pooled comparison classrooms as the "dummy variable" left out of the regression. Analysis 1 did not contain separate slope coefficients for the various models. Analysis 2 allowed for separate slope coefficients for PSI pre-score and prior preschool experience. Analysis 1 explained 70.6% of the total variation; analysis 2 explained 73.4% of the total variation.

3. Only sites with both PV and Comparison classrooms (on or off-site) were included in this analysis.

Note: All analyses included the following covariables: PSI pre-test mean, Book 3D pre-test mean, Book 4A pre-test mean, mean age, percent black, percent Mexican-American, percent female, mean income, mean household size, teacher experience in Head Start, teacher certification, mean mother's education, percent prior preschool, average staff working conditions, whether the site is El or Ek. In the analyses in column 1 the variable "site administered by CAP or by Public School" was also included. In the regression analyses in columns 2 and 3 teacher reace was included. In analyses of the Stanford-Binet, the Stanford-Binet pre-test was also included as a covariate--these analyses used only Level III sites. In analyses of the Motor Inhibition only classrooms with valid Motor Inhibition scores for both fall and spring were included.

TABLE VII-8A

Selected Statistics for Matched Classroom Analysis of Book 4A for the 5 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.6
Far West Laboratory	8	0.54	0.81	0.82	0.82	0.82
Arizona	8	-0.98	0.44	0.44	0.44	0.43
Bank St.	11	0.83	-0.10	0.10	0.11	0.11
Univ. of Oregon	12	-0.19	0.97	0.97	0.97	0.97
Univ. of Kansas	8	-0.53	1.54	3.23***	3.23***	3.23***
High Scope	12	-0.06	0.17	0.17	0.17	0.17
Univ. of Florida	11	-0.32	-0.24	-0.24	-0.24	-0.24
EDC	11	0.13	0.57	0.58	0.58	0.58
Univ. of Pittsburgh	4	1.31	2.76	2.77*	2.77*	2.78*
REC	4	-0.45	-2.58	-2.58*	-2.58*	-2.59*
Enablers	12	0.24	-0.62	-0.62	-0.61	-0.61

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = .51. The overall F for the test of homogeneity of the covariate regression coefficient = 0.78.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = -0.00; with r_{tt} estimated as 0.80 the coefficient = -.01; for r_{tt} = 0.60, the coefficient = -.01.

TABLE VII-8B

Selected Statistics for Matched Classroom Analysis of Book 4A for the 4 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.60
Far West Laboratory	8	1.22	2.01	1.12	0.90	0.53
Arizona	8	-1.00	0.75	1.48	1.66	1.97
Bank St.	11	0.26	-1.05	-1.25	-1.29	-1.37
Univ. of Oregon	12	0.24	0.56	0.38	0.34	0.27
Univ. of Kansas	8	-0.06	3.80	3.85***	3.86***	3.88***
High Scope	12	0.37	0.88	0.60	0.54	0.42
Univ. of Florida	11	-1.26	-1.15	-0.23	-0.01	0.38
EDC	11	-0.50	-0.84	-0.47	-0.38	-0.22
Univ. of Pittsburgh	4	-0.19	1.09	1.23	1.27	1.32
REC	4	-1.56	-2.07	-0.94	-0.65	-0.18
Enablers	12	0.95	-0.38	-1.07	-1.24	-1.53

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = -0.15. The overall F for the test of homogeneity of the covariate regression coefficient = 0.91.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.73 with r_{tt} estimated as 0.80 the coefficient = .91; for r_{tt} = 0.60, the coefficient = 1.21.

or roughly 0.50 standard deviations. In contrast with matched Comparison samples their "effects" ranged from -0.60 to -1.53. Though only two of the contrasts were significant, the overall pattern does suggest that the Enabler "model" was not as effective as the other Head Start programs as assessed by the Book 4A test.

The results for the REC model are less clear. The lack of replication for this model again makes its effects difficult to assess. Children in this model "gained" 2.71 points on the average on the Book 4A test, placing them third from the bottom in terms of mean model gains. Compared to the overall PV mean, the adjusted REC "effect" is -1.37 points while contrasted to the overall Comparison adjusted mean, the effect was positive (roughly 1.4 points). In the two matched classroom analyses it had negative "estimated differences." In the five factor sample analysis the negative differences are considerable (about 2.58 points or roughly 0.80 standard deviations). The four factor sample analyses yield much smaller negative differences. As in the case of the Book 3D test, we are ambiv.lent about reaching conclusions about REC, since it has only one site.

In summary, there is a clearly exemplary model with respect to the Book 4A test. The University of Kansas model exceeds all other models and the Comparison classes in all analyses by a substantial margin. Moreover, there appears to be some evidence that the other two highly structured and academically oriented models (University of Oregon and University of Pittsburgh) are also especially effective on this outcome measure. There is some evidence that the Enabler model and the REC model are not as effective as the other Head Start programs, but the evidence is not at all conclusive.

VI. The Preschool Inventory

A. Site to Site Differences (See Table VII-9)

Differences among PV sites in PSI pre-score means range from 23.71 to 49.05 points, roughly 2.5 standard deviations. The middle fifty percent, however, range from only 29.98 to 37.77 points or about 0.80 individual standard deviations. Comparison site pre-test means are somewhat more closely bunched, ranging from 26.49 to 45.95 points, about 1.9 standard deviations, while the middle fifty percent range from 30.80 to 39.60 points, or 0.90 standard deviations.

These initial differences among the sites are comparable to the size of the differences found for the Book 4A test. The differences in relative "gains," however, are considerably smaller. For the PV sites the range of "observed" gains is from 2.08 to 17.0 points, roughly 1.4 standard deviations of the individual test scores. This contrasts with a difference of 2.4 standard deviations for the Book 4A test. Comparison site "gains" have a similar range--from 6.72 to 19.07 points, 1.3 individual standard deviations. The middle fifty percent

TABLE VII-9

Preschool Inventory

Pre-test means and mean "gains" (post-test mean - pre-test mean) by site for PV and Comparison groups. Site means are unweighted averages of classroom means.

Sponsor	Code	Community	Testing Level	PV Pre-test Mean	Comp. Pre-test Mean	PV "Gain"	Comp. "Gain"	PV classrooms (#)	Comp. classrooms (#)
Nimnicht	02.04	Duluth	III	37.56		10.56		4	
	02.04	St. Cloud	III		37.84		14.98		2
	02.13	Tacoma	II	36.02		8.45		4	
Tucson	03.08	LaFayette	III	41.86		12.56		4	
	03.08	Albany	III		40.33		11.47		4
	03.16	Lincoln	III	34.07		11.84		4	
Bank St.	05.01	Boulder	III	36.16	35.18	13.26	15.18	4	1
	05.11	Wilmington	II	25.74		9.72		4	
	05.11	DeLaWar	II		28.18		9.15		4
	05.12	Elmira	III	33.23	30.80	7.79	15.01	3	3
Becker & Engle-mann	07.03	E. St. Louis	III	35.67	37.58	13.67	19.07	4	4
	07.11	Tupelo	III	45.24	45. 5	9.19	9.10	4	4
	07.14	E. Las Vegas	II	36.29		12.33		4	
	07.14	W. Las Vegas	II		37.57		12.49		4
Bushell	08.04	Portagoville	III	32.09	29.09	13.06	11.21	4	4
	08.08	Mounds, Ill.	II	37.77	37.81	10.13	11.41	4	2
Weikart	09.02	Ft. Walton B.	III	23.72		17.0		4	
	09.02	Pensacola	III		31.38		12.91		3
	09.06	Greeley	III	40.45	41.50	2.08	6.72	4	3
	09.10	Seattle	II	40.43	43.24	12.69	10.28	4	3
Gordon	10.02	Jonesboro	III	36.41	39.88	11.99	12.0	3	3
	10.07	Chattanooga	III	35.82	37.61	11.48	10.71	4	4
	10.10	Houston	II	28.54	31.67	7.42	8.00	4	4
EDC	11.05	Washington	III	29.35	26.49	10.32	11.45	4	4
	11.06	Paterson	II	27.98	29.44	9.96	15.12	3	1
	11.08	Johnston Co.	III	45.12	39.66	8.98	10.46	4	4
Pitts-burgh	12.03	Lock Haven	III	29.07		13.06		4	
	12.03	Mifflenburg	III		26.81		11.04		4
RFC	20.01	Kansas City	III	29.97		11.27		4	
Enablers	27.04	Billings	II	27.76		16.58		4	
	27.05	Colorado Sp.	II	34.82		9.15		4	
	27.03	Bellows Falls	II	49.05		7.71		4	

of the PV site gains range from 9.15 to 12.56 points--a difference of under 0.40 standard deviations. The middle fifty percent of Comparison site "gains" range from 10.28 to 12.49 points, under 0.30 standard deviations. Clearly the majority of sites in this study are closely bunched in terms of "gains" on the PSI test.

None of the PV models has more than one site in the bottom twenty-five percent of the distribution of PSI "gains" though the Enabler model has one site at the 25th percentile and one below. The third Enabler site however, has a "gain" of 16.58 points, placing it second in "observed gains" among sites. Moreover, the model containing the site with the smallest "gain" is also the model with the site having the largest "gain". This model (High/Scope) has two sites above the 75th percentile in "gains". This happens for no other model.

II. Model to Model Differences

Table VII-10 displays model to model differences in PV and Comparison "gains" on the PSI. For the PV models the "observed gains" range from 9.19 to 13.10 points, roughly 0.40 standard deviations.

When PV and Comparison classes within models are compared on observed PSI gains the differences range from 1.64 points favoring the University of Pittsburgh model to -5.24 points favoring the Comparison group for the Far

TABLE VII-10

Model Statistics for the PSI Test

Column 1 shows the mean gain for PV children in the model.
Column 2 shows the mean gain for Comparison children in model location.
Column 3 shows the difference between Column 1 and Column 2. (A positive score indicates that PV children gained more than Comparison children).
Column 4 shows the difference between PV and Comparison children in observed-expected gains.
The individual is the unit of analysis.[1]

Model	PV "Gains"	Comparison "Gains"	PV "Gains"-Comparison "Gains"	PV (observed-expected) "gains"-comparison (observed-expected) "gains"

Far West Laboratory	SD=7.78
	9.19
	N=69
		6.71
		14.43
		44
			-5.24***	-5.33***

Arizona	7.90 12.03 132	6.55 11.16 61	0.87	-2.09
Bank St.	7.87 10.52 124	8.47 12.20 98	-1.69	-1.15
U. of Oregon	8.77 11.22 183	9.72 13.76 167	-2.53*	-1.60
U. of Kansas	6.96 11.91 106	7.45 11.18 62	0.73	0.35
High Scope	8.86 11.63 123	7.42 10.23 96	1.40	1.86
U. of Florida	7.09 10.80 109	6.89 10.36 124	0.44	0.40
EDC	6.26 9.44 146	8.15 11.65 123	-2.20*	-2.30**
U. of Pittsburgh	9.42 13.10 32	9.11 11.45 31	1.64	2.48
RFC	7.35 11.52 19			
Enablers	7.27 11.25 120			

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

[1] All children in the basic analysis sample were used (see Chapter III)

West model. If we eliminate the Far West model from our consideration because of inappropriate on-site comparisons the range if halved, going from 1.64 to -2.35 (the latter being for the Oregon model). The "observed-expected" differences are essentially similar to the "observed" differences.

There is some indication from Table VII-10 that the Far West Laboratory, the University of Oregon and the EDC models are somewhat less than average in their effectiveness as assessed by the PSI. The Far West and the EDC models have the smallest model "gains", both being smaller than any of the Comparison group"gains." While the University of Oregon PV model has a "gain" in the average range, its Comparison group has the second largest average "gain" among the Comparison groups.

C. "Adjusted Differences Among Groups"

Of the 108 contrasts in Tables VII-11 and VII-12 only nine reach statistical significance. Moreover, there are no clear patterns of results as there were for the Book 4A test. The lack of interesting results can most clearly be seen by looking at the matched classroom results in Table 12. Of the 66 contrasts in this Table only one reaches statistical significance and that only when the reliability of the covariate is assumed to be only 0.60--undoubtedly an underestimate (see Chapter VI).

TABLE VII - 11

PSI

Model "effect" estimates for the test. Columns 1-4 show differences between "adjusted" PV model means and some standard. Column 1 shows the simple contrasts between the PV model "adjusted" means and an unweighted grand mean of the model means for an exact least squares one way ANCOVA. Columns 2 and 3 show regression coefficients for each model in an analysis where all of the comparison classes are pooled together to form a comparison "model". The regression coefficients can be thought of as representing the difference between the "adjusted" PV model means and the "adjusted" Comparison "model" means. Column 2 shows the coefficients for a regression analysis not allowing for separate slope coefficients for the covariates for the different models. Column 3 shows the coefficients allowing for separate model coefficients for the PSI pre-test and for percent prior preschool. Column 4 shows the difference between PV and Comparison group "adjusted" means within models for sites with both a PV and a Comparison group. The estimates are 1 degree of freedom contrasts in the framework of a one way ANCOVA design. Column 5 shows the PV and Comparison n's for column 4 analysis. A note following the Table lists the covariates used in the analysis. In all analyses the classroom is the unit of analysis. See text (Chapters V and VII) for further discussion of the approaches.

Model	Estim.effects around PV unweighted mean	Estimated effects of PV models against pooled compar. classes[2]		DF contrast PV v. site comp.pooled by models[3]	PV N	Comp. N
		analysis 1	analysis 2			
Far West Laboratory	-0.46 N=8	-1.68	-2.01	-5.00	4	2
Arizona	0.20 8	0.47	0.09	-2.42	4	4
Bank St.	-0.73 11	-0.77	-0.72	1.41	11	8
U. of Oregon	3.05 12	0.19	5.44**	-3.34*	12	12
U. of Kansas	2.25 8	0.26	0.25	0.91	8	6
High Scope	-0.04 12	0.49	-0.30	0.36	12	9
U. of Florida	-3.17* 11	-2.96*	-2.85*	-2.14	11	11
EDC	-3.04* 11	-2.54*	-2.49*	-0.96	11	9
U. of Pittsburgh	0.46 4	0.23	0.93	3.22	4	4
REC	1.51 4	1.15	1.93			
Enablers	-0.94 12	-0.87	-1.26			
Grand Mean	45.54	46.34	46.34	46.56		

TABLE VII-11

(Page 2)

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

1. Only PV classrooms are included in this analysis. The multivariate F with the PSI, Book 3D and Book 4A in the analysis is 2.36; significant at the .001 level. The overall univariate F for the PSI is 2.27, significant at the .05 level.

2. Both analyses were in the regression framework with the pooled Comparison classrooms as the "dummy variable" left out of the regression. Analysis 1 did not contain separate slope coefficients for the various models. Analysis 2 allowed for separate slope coefficients for PSI pre-score and Prior Preschool Experience. Analysis 1 explained 78.1% of the total variation; analysis 2 explained 81.4% of the total variation.

3. Only sites with both PV and Comparison classrooms (on or off-site) were included in this analysis.

Note: All analyses included the following covariables: PSI pre-test mean, Book 3D pre-test mean, Book 4A pre-test mean, mean age, percent black, percent Mexican-American, percent female, mean income, mean household size, teacher experience in Head Start, teacher certification, mean mother's education, percent prior preschool, average staff working conditions, whether the site is El or Ek. In the analyses in column 1 the variable "site administered by CAP or by Public School" was also included. In the regression analyses in columns 2 and 3 teacher race was included. In analyses of the Stanford-Binet, the Stanford-Binet pre-test was also included as a covariate--these analyses used only Level III sites. In analyses of the Motor Inhibition only classrooms with valid Motor Inhibition scores for both fall and spring were included.

TABLE VII-12A

Selected Statistics for Matched Classroom Analysis of the PSI for the 5 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Col.mn 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variable: for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.6C
Far West Laboratory	8	1.09	-1.18	-1.76	-1.90	-2.15
Arizona	8	-0.02	2.57	2.58	2.58	2.58
Bank St.	11	-0.75	-1.41	-1.02	-0.92	-0.75
Univ. of Oregon	12	1.55	0.86	0.04	-0.17	-0.51
Univ. of Kansas	8	4.71	3.71	1.20	0.57	-0.47
High Scope	12	2.04	2.12	1.03	0.76	0.31
Univ. of Florida	11	0.20	-0.35	-0.46	-0.48	-0.53
EDC	11	-0.64	-0.54	-0.20	-0.11	0.03
Univ. of Pittsburgh	4	0.92	1.39	0.90	0.78	0.57
REC	4	-0.58	-2.82	-2.51	-2.43	-2.30
Enablers	12	1.62	1.14	0.28	0.06	-0.30

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.82 The overall F for the test of homogeneity of the covariate regression coefficient = 0.75.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.53; with r_{tt} estimated as 0.80 the coefficient = .67; for r_{tt} = 0.60, the coefficient = .89.

TABLE VII-12B

Selected Statistics for Matched Classroom Analysis of the PSI for the 4 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

		Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
	N's			Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.6C
Far West Laboratory	8	5.03	2.67	-1.09	-2.03	-3.60
Arizona	8	1.39	0.75	-0.28	-0.54	0.98
Bank St.	[illegible]	-2.39	-4.11	-2.31	-1.87	-1.12
Univ. of Oregon	12	1.59	0.72	-0.48	-0.77	-1.27
Univ. of Kansas	8	6.12	4.75	0.16	-0.98	-2.90
High Scope	12	4.43	5.78	2.46	1.63	0.25
Univ. of Florida	11	-2.51	-3.67	-1.78	-1.31	-0.52
EDC	11	-2.25	-1.77	-0.08	0.34	1.04
Univ. of Pittsburgh	4	-4.14	-5.71	-2.60	-1.83	-0.54
REC	4	-4.45	-4.56	-1.22	-0.39	1.01
Enablers	12	4.74	2.23	-1.33	-2.21	-3.69*

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.29. The overall F for the test of homogeneity of the covariate regression coefficient = 2.08.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.75; with r_{tt} estimated as 0.80 the coefficient = 0.94; for r_{tt} = 0.60, the coefficient = 1.25

Three rough patterns can be suggested, however.

(1) First, there is no indication of special effectiveness for the academically oriented, highly structured models on PSI gains. Although Pittsburgh does have the highest model "gain" its contrasts with other models and Comparison classrooms suggests that, as a model, it is of only average effectiveness. Similarly both the Kansas and Oregon models show only average effects.

(2) The Arizona, High/Scope, REC and Enablers models also show inconsistent and generally average results. While these models appear similar when their sites are aggregated they are quite different when individual sites are examined. For example, both Arizona sites have observed gains in the middle range of site gains while the High/Scope sites show great variance in their observed gains, as do the Enabler sites.

(3) Far West Laboratories, Bank Street, the University of Florida, and EDC show generally negative estimates of effects though few of the contrasts are significant and occasionally even of a positive sign. Both EDC and the University of Florida show statistically significant negative estimates when contrasted to the other PV models and to the overall pooled Comparison classes. Far West and Bank Street do not show statistically significant results though, with one exception for each model, all of their effects are in a negative direction. The effects,

however, are very small, never exceeding 0.30 standard deviations when the Comparison group is appropriate. Our general conclusions, therefore, is to assume that the models are all of roughly equal effectiveness.

The lack of clear differences among models in their effectiveness as assessed by the PSI may be due to the nature of the test. It, more than any of the other tests examined here, was designed to tap the general dimensions of a preschool experience. As such it should be less sensitive than other tests like Book 4A to specific differences in curricula. One indication of this is the relatively small range of differences in observed gains among the sites. It could well be that this test is inappropriate for an analysis of differences among curricula. Although it may serve a general purpose in pointing out particularly weak or strong sites (note the High/Scope site differences) it perhaps is better suited for analyses of individual differences among types of children. The report "Cognitive Effects of Preschool Programs on Different types of Children" explores this issue in detail.

VII. The Stanford-Binet

A. Site to Site Differences:

Table VII-13 shows site pre-test means and observed gains for both PV and Comparison groups. Only the sixteen Level III sites are included in the analyses of the Stanford-Binet. This excludes the Enabler model and reduces the maximum number of sites per model to two. PV pre-test

TABLE VII- 13

Stanford-Binet

Pre-test means and mean "gains" (post-test mean - pre-test mean) by site for PV and Comparison groups. Site means are unweighted averages of classroom means. . .

Sponsor	Code	Community	Testing Level	PV Pre-test mean	Comp. Pre-test mean	PV "Gain"	Comp. "Gain"	PV classrooms (#)	Comp. classrooms (#)
Nimnicht	02.04	Duluth	III	90.16		4.50		4	
	02.04	St. Cloud	III		98.61		4.44		2
	02.13	Tacoma	II						
Tucson	03.08	LaFayette	III	90.33		5.72		3	
	03.08	Albany	III		88.75		2.30		4
	03.16	Lincoln	III	94.88		3.07		4	
Bank St.	05.01	Boulder	III	99.76	101.80	-1.53	-0.22	4	1
	05.11	Wilmington	II						
	05.11	DeLaWar	II						
	05.12	Elmira	III	96.15	98.63	-1.80	2.39	3	3
Becker & Englemann	07.03	E. St. Louis	III	92.00	99.56	5.81	0.13	4	4
	07.11	Tupelo	III	94.95	91.77	-0.57	0.59	4	4
	07.14	E. Las Vegas	II						
	07.14	W. Las Vegas	II						
Bushell	08.04	Portageville	III	91.70	87.51	2.70	1.20	4	4
	08.08	Mounds, Ill.	II						
Weikart	09.02	Ft. Walton B.	III	77.40		30.59		4	
	09.02	Pensacola	III		82.52		7.39		3
	09.06	Greeley	III	87.62	96.52	12.05	7.36	4	3
	09.10	Seattle	II						
Gordon	10.02	Jonesboro	III	80.04	85.44	9.62	7.73	3	3
	10.07	Chattanooga	III	78.55	86.77	1.61	2.94	4	4
	10.10	Houston	II						
EDC	11.05	Washington	III	93.83	86.64	-0.37	-0.47	4	4
	11.06	Paterson	II						
	11.08	Johnston Co.	III	86.95	86.14	5.08	4.97	4	4
Pittsburgh	12.03	Lock Haven	III	98.40		8.16		4	
	12.03	Mifflenburg	III		86.80		5.66		4
REC	20.01	Kansas City	III	95.44		7.33		4	
Enablers	27.04	Billings	II						
	27.05	Colorado Sp.	II						
	27.03	Bellows Falls	II						

means range from 77.4 to 99.76 points--a range of roughly 1.7 individual standard deviations in this sample or from a very low "normal" level to the national average in terms of national norms. The Comparison site range is almost as large--from 82.52 to 101.80 points. The middle fifty percent of the PV sites range from 87.62 to 94.95 points, a difference of roughly 0.5 standard deviations, while the middle fifty percent for the Comparison group range from 86.80 to 96.80 points. Thus although there are a few sites with very low pre-test means, by and large, the sites cluster between one-third and one standard deviation below the national mean.

Observed gains for the PV sites range from -1.53 to 30.59 points. The latter, however, is an extreme outlyer--without it the range is reduced to -1.53 to 12.05 points, a gap of about one standard deviation. The Comparison site range of observed gains is not even as large as this reduced range, going from -0.47 to 7.73 points. When we look at the middle fifty percent range of gains the PV spread becomes only about 4.6 points, from 2.70 to 7.33 points while the Comparison site spread is also about 4.6 points, ranging from 1.20 to 5.66 points--roughly 0.35 standard deviations.

When we look at sites within models the spread does not reduce quite so much as it does for the other tests. One model (Bank Street) has two PV sites in the bottom quartile while another (High/Scope) has the two sites making the greatest gains. Of note, however, is the fact that the Comparison sites for Bank Street also gain very little (relative to the other sites) while the Comparison sites for High/Scope are both in the top quartile of Comparison site gains.

B. Model to Model Differences:

The spread in model to model"gains"is shown clearly in Table VII-14. The High/Scope PV model far outgains any of the other PV models, averaging 23.4 points in "gains" while Bank Street lags behind with an average "gain" of -1.73 points.* Two other PV models show higher than average gains--both the University of Pittsburgh and REC show gains of slightly over eight points. All of the other PV models gain between 2.5 and 5.24 points, a difference of less than 1/4 of a standard deviation of individual test scores. The Comparison groups show less variation with the Bank Street Comparison group having the smallest "gains" (-0.65 points) and the High/Scope Comparison group the largest (7.18 points).

In the contrasts between the observed and the "observed-expected" gains for the PV and Comparison groups the High/

*When interpreting these gains it is important to remember that we expect some deterioration in Stanford-Binet over the seven months a child is in preschool. Thus, all of the models are producing slight positive effects (see Chapter IV).

TABLE VII-14

Model Statistics for the Stanford-Binet

Column 1 shows the mean gain for PV children in the model.
Column 2 shows the mean gain for Comparison children in model location.
Column 3 shows the difference between Column 1 and Column 2. (A positive score indicates that PV children gained more than Comparison children).
Column 4 shows the difference between PV and Comparison children in observed-expected gains.
The individual is the unit of analysis.[1]

Model	PV "Gains"	Comparison "Gains"	PV "Gains"-Comparison "Gains"	PV (observed-expected) "gains"-comparison (observed-expected) "gains"

Far West Laboratory	SD=11.71
	3.32
	N=13
		8.02
		3.71
		19
			-0.39	-0.10

Arizona	8.77 4.14 55	9.10 2.30 25	1.84	0.58
Bank St.	9.10 -1.73 36	7.71 -0.65 15	-1.08	-0.37
U. of Oregon	9.52 2.49 77	9.41 0.25 55	2.25	1.72
U. of Kansas	9.34 2.72 27	9.93 0.92 25	1.80	1.22
High Scope	12.37 23.54 47	7.78 7.18 40	16.37***	16.58***
U. of Florida	9.47 5.24 43	9.79 4.24 51	1.00	-0.53
EDC	8.41 3.43 47	10.21 2.73 52	0.70	0.67
U. of Pittsburgh	12.17 8.23 21	7.30 4.74 15	3.49	1.75
REC	9.10 8.09 23			
Enablers				

***Statistically significant at the .001 level

[1] All children in the basic analysis sample were used (see Chapter III)

Scope PV model stands out as clearly different from all of the others with an advantage favoring the PV group of roughly 16.5 points. None of the other measured differences exceeds 3.5 points. Thus, in terms of simple gains and differences between PV and Comparison groups there is only one main finding in this data-- the High/Scope model appears to be extraordinarily effective in raising Stanford-Binet scores at least in the short run. Other than that there are no differences of note in the data shown in Table VII-14.

C. "Adjusted Differences Between Groups":

When the data in Tables VII-15 and VII-16 are examined, the picture becomes only slightly more complex. Six PV models require little attention. Far West Laboratories, Arizona, Oregon, Kansas, Florida and EDC all show small and inconsistent effects,. Note that this group includes two of the highly structured academically oriented models, an indication that this approach is not necessarily more effective than other approaches when the Stanford-Binet is the outcome measure. In the following we consider the four remaining models.

(1) The University of Pittsburgh model had observed gains averaging over eight points. Although it appears less effective than the average PV model, (see Column 1, Table VII-15) it shows positive effects in all of the other contrasts, in Tables VII-15 and VII-16. Of the eight other contrasts five show

TABLE VII - 15
Stanford-Binet

Model "effect" estimates for the test. Columns 1-4 show differences between "adjusted" PV model means and some standard. Column 1 shows the simple contrasts between the PV model "adjusted" means and an unweighted grand mean of the model means for an exact least squares one way ANCOVA. Columns 2 and 3 show regression coefficients for each model in an analysis where all of the comparison classes are pooled together to form a comparison "model". The regression coefficients can be thought of as representing the difference between the "adjusted" PV model means and the "adjusted" Comparison "model" means. Column 2 shows the coefficients for a regression analysis not allowing for separate slope coefficients for the covariates for the different models. Column 3 shows the coefficients allowing for separate model coefficients for the PSI pre-test and for percent prior preschool. Column 4 shows the difference between PV and Comparison group "adjusted" means within models for sites with both a PV and a Comparison group. The estimates are 1 degree of freedom contrasts in the framework of a one way ANCOVA design. Column 5 shows the PV and Comparison n's for column 4 analysis. A note following the Table lists the covariates used in the analysis. In all analyses the classroom is the unit of analysis. See text (Chapters V and VII) for further discussion of the approaches.

Model	Estim.effects around PV unweighted mean	Estimated effects of PV models against pooled compar. classes[2]		DF contrast PV v. site comp.pooled by models[3]	PV N	Comp. N
		analysis 1	analysis 2			
Far West Laboratory	-0.03 N=4	-5.74		-2.04	4	2
Arizona	1.63 7	-1.30		0.43	3	4
Bank St.	-11.33* 7	-3.34		-3.13	7	4
U. of Oregon	-6.64 8	0.96		0.11	8	8
U. of Kansas	-6.78 4	-2.86		-1.28	4	4
High Scope	15.46*** 8	15.31***		10.06**	8	6
U. of Florida	-0.23 7	-3.47		-1.37	7	7
EDC	7.93 8	-0.00		0.34	8	8
U. of Pittsburgh	-1.30 4	6.66*		9.08*	4	4
REC	17.20*** 4	10.97***				
Enablers						
Grand Mean	96.90	95.16	-	95.24		

TABLE VII-15

(Page 2)

* Statistically significant at the .01 level
** Statistically significant at the .05 level
*** Statistically significant at the .001 level

1. Only PV classrooms are included in this analysis. The multivariate F with the PSI, Book 3D, Book 4A and Stanford-Binet in the regression is 2.90; significant at the .001 level. The overall F for the Stanford-Binet is 7.80, significant at the .001 level.

2. Both analyses were in the regression framework with the pooled Comparison classrooms as the "dummy variable" left out of the regression. Analysis 1 did not contain separate slope coefficients for the various models. Analysis 2 allowed for separate slope coefficients for PSI pre-score and Prior Preschool Experience. Analysis 1 explained 70.7% of the total variation.

3. Only sites with both PV and Comparison classrooms (on or off-site) were included in this analysis.

Note: All analyses included the following covariables: PSI pre-test mean, Book 3D pre-test mean, Book 4A pre-test mean, mean age, percent black, percent Mexican-American, percent female, mean income, mean household size, teacher experience in Head Start, teacher certification, mean mother's education, percent prior preschool, average staff working conditions, whether the site is El or Ex. In the analyses in column 1 the variable "site administered by CAP or by Public School" was also included. In the regression analyses in columns 2 and 3 teacher race was included. In analyses of the Stanford-Binet, the Stanford-Binet pre-test was also included as a covariate--these analyses used only Level III sites. In analyses of the Motor Inhibition only classrooms with valid Motor Inhibition scores for both fall and spring were included.

TABLE VII-16A

Selected Statistics for Matched Classroom Analysis for the Stanford-Binet for the 5 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.60
Far West Laboratory	4	-9.09	-10.11	-7.80*	-7.22*	-6.26
Arizona	7	-2.27	-1.27	-0.70	-0.56	-0.32
Bank St.	7	0.19	-6.02	-5.43*	-5.44*	-5.46*
Univ. of Oregon	8	0.18	1.70	1.66	1.64	1.62
Univ. of Kansas	4	-0.47	-2.49	-2.37	-2.35	-2.30
High Scope	8	-6.80	8.99	10.73***	11.16***	11.88***
Univ. of Florida	7	-5.01	-5.34	-4.17	-3.85	-3.32
EDC	8	-0.20	1.73	1.78	1.79	1.81
Univ. of Pittsburgh	4	8.15	13.15	11.08***	10.57***	9.70**
RFC	4	-0.53	2.55	2.69	2.76	2.78
Enablers						

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = .41. The overall F for the test of homogeneity of the covariate regression coefficient = 0.81.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.25; with r_{tt} estimated as 0.80 the coefficient = 0.32; for r_{tt} = 0.60, the coefficient = 0.43.

TABLE VII-16B

Selected Statistics for Matched Classroom Analysis of the Stanford-Binet for the 4 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

		Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
	N's			Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.60
Far West Laboratory	4	-7.64	-6.47	-1.40	-0.13	1.99
Arizona	7	-3.53	-1.10	1.24	1.83	2.81
Bank St.	7	6.48	-0.14	-4.44	-5.52	-7.31**
Univ. of Oregon	8	2.83	4.24	2.35	1.88	1.10
Univ. of Kansas	4	0.51	-0.03	-0.37	-0.45	-0.59
High Scope	8	-5.29	10.59	14.10***	14.98***	16.44***
Univ. of Florida	7	-5.80	-6.39	-2.79	-1.83	-0.22
EDC	8	-2.00	-0.92	0.41	0.74	1.29
Univ. of Pittsburgh	4	2.28	8.09	6.57	6.19	5.56
REC	4	-3.14	1.18	3.26	3.78	4.65
Enablers						

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.29. The overall F for the test of homogeneity of the covariate regression coefficient = 2.19.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.66; with r_{tt} estimated as 0.80 the coefficient = 0.83; for r_{tt} = 0.60, the coefficient = 1.11.

statistically significant results--no difference is less than 5.6 points or roughly one-half a standard deviation. Although it is difficult to draw a conclusion about Pittsburgh since it has only one site, there is a strong indication that it is more effective than the Comparison Head Start programs in imparting gains on the Stanford-Binet.

(2) The REC model showed an average gain of slightly over seven points. When placed in an analysis directly contrasting PV models the REC model shows a highly significant effect of 17.20 points. In contrast with the overall Comparison group it also shows a statistically significant effect of 10.97 points. In the matched sample analyses, however, the REC model does not show a large effect (it never exceeds 4.65 points) although in all instances the direction of the effect is positive. We are unclear about the cause of this rather dramatic set of differences between estimated effects from different analyses. Presumably, it has something to do with the form and nature of the covariates used in the analyses. Whatever the reason, however, there is clearly an indication that the REC model may be more effective than most PV and Comparison Head Start programs. It is important, though, to remember that there is only one REC site so it is impossible for us to reach a firm conclusion about the model.

(3) The Bank Street model had the smallest "gains" on the Stanford-Binet of any of the PV models. Both Bank Street sites had similar pre-test and gain scores. In the analyses described in Tables VII-15 and VII-16 Bank Street shows a consistently negative effect, ranging from -11.33 to -3.13 points. Of the nine contrasts five are statistically significant. If we disregard the largest negative effect because it occurs in the PV model to model contrasts,(the analysis we have least confidence in) the range is from roughly -7.5 to -3.13 points. The only reason we have not to conclude that the model is less effective than the other PV models stems from the extraordinarily high pre-test means in both of its sites for PV and Comparison groups. Although these sites show moderately high pre-test means for the other outcome measures, they do not come close to approximating the relative magnitudes of the Stanford-Binet pre-scores. This suggests that the Bank Street site pre-scores for both PV and Comparison children may be over-inflated--perhaps because of an over-zealous tester. This would give the children in these sites little opportunity to show impressive gains on the Stanford-Binet. A somewhat more conventional interpretation might be that a regression artifact was working on the Bank Street scores--the high initial scores would have to be due to substantial positive errors in both sites for this explanation to work. Since the average classroom N for the Stanford-Binet analyses

is only about five this explanation may be plausible. Our inclination, then, is not to reach a firm conclusion about the Bank Street model's effectiveness for the Stanford-Binet as the outcome measure.

(4) The High/Scope model shows dramatic positive effects in all analyses of the Stanford-Binet. Estimates of adjusted differences range from 10 to 17 points--from 0.7 to 1.3 standard deviations. This effect is comparable in magnitude to the effect found for the University of Kansas model on the Book 4A test but it potentially is far more important. Its special importance stems from the characteristics of the outcome measure. The Stanford-Binet was developed to tap general intelligence--a trait that by definition is not sensitive to slight changes in environment. Moreover, in practice, Stanford-Binet scores are generally difficult to change very substantially. Yet here we see an estimated change of almost a standard deviation in magnitude effected over a seven month preschool program. What accounts for this effect?

Three issues are important. The first two have to do with the data and the third has to do with the nature of the High/Scope program. First, although the High/Scope PV sites ranked first and second in observed gains there was a dramatic difference between the two sites. In one site the average gain was roughly 30 points--of some importance is the fact that the four classes in this site had gains of almost equal magnitude. The second site averaged gains of only twelve points. Although both gains are impressive in magnitude the difference between themsuggests that the effects of the High/Scope program may be sensitive to differences among sites in such things as pupil composition or location. In this instance the site with the thirty point gains is located in the rural South, has a racial composition of roughly 70% black and 30% white with none of the children having previously attended preschool. The other High/Scope site is located in a small urban northern city and has a racial composition of about three-quarters Mexican-American, about one-sixth of whom had previously attended preschool.

Second, the pre-score mean for the children in the southern rural site was the lowest of the site pre-score means. On the one hand, this suggests that a regression artifact might account for some of the thirty-point gain. Even supposing, however, that the Binet had a reliability of only 0.70 (undoubtedly an underestimate for even the individual test administration much less classroom aggregated means) and assuming that the "true" population pre-test mean was 95 (probably an overestimate since it exceeds the overall pre-test mean for the entire PV sample), then the regression effect would account for a little over five points of the thirty point gain.*

On the other hand, however, there is some independent information suggesting that the low pre-test scores are valid. For each of the other academic outcome measures discussed in this chapter the southern rural High/Scope site had either the lowest or second lowest pre-test site mean. Since these tests were given by other testers than those who gave the Stanford-Binet, there is good reason to believe that the pre-scores on the Binet are roughly accurate. In the other High/Scope site, the pre-score mean for the Binet is close to the overall PV sample mean and is consistent with the other outcome measures. This suggests there is little chance of a regression effect for this site.

*Knowing the observed mean, reliability and population mean, we can estimate the magnitude of the regression effect. In this instance, we have an observed mean of about 77, we have taken a lower bound for a possible reliability (0.70) and we have taken a high estimate of a population mean (95). Our approach, therefore, will overestimate the regression effect. The general formulation is that the regression effect is equal to 1.0 reliability times the difference between the population mean and the observed sample mean. In this case we have:

$$\text{regression effect} = (1.0 - 0.70) \times (95 - 77) = 5.4 \text{ points}$$

This argument also deals in part with some observations made by SRI personnel at the southern rural site. Apparently, the fall Stanford-Binet tester was very efficient about his work, spent little time with the children and as a consequence, apparently had little rapport with them. The spring tester, however, was loved by the children and spent a much longer period of time administering the test. The difference in style might account for some of the gain. It is possible that the fall tester was obtaining underestimates of the "true" scores while the spring tester was obtaining overestimates. The very low pre-test scores for the other outcome measures, however, suggest that the fall tester was probably not particularly biased. A possible bias on the part of the spring tester cannot be so easily dealt with. We note that for both the Book 3D and the PSI outcomes, the gain scores for this site were either near the largest or the largest. This, however, does not account for a thirty point Binet gain. Our best guess is that roughly ten of the thirty points are probably due to a combination of tester and regression effects. There do not seem to be any peculiarities about the testers in the other High/Scope site. Thus, we estimate that the true gain for the southern rural site is roughly 20 points while the "true gain" for the northern urban site is roughly 12 points.

Third, there is some preliminary indication that the children in the High/Scope sites are getting certain items correct on the Stanford-Binet post-test that children in other programs are not getting right. These items have to do with differences and similarities -- concepts that are an integral part of the High/Scope curriculum. An analysis of this issue as well as of possible tester bias is included in an appendix to this report.

In summary, we conclude that the High/Scope model is particularly effective in producing gains on the Stanford-Binet. We estimate the "true observed gain" to be in the range of 15 to 20 points while the differences between High/Scope and conventional Head Start gains range from 10 to 17 points with a "true" effect probably closer to the bottom end of that range. We reached no firm conclusions indicating positive or negative effects for any of the other models though there is some indication that Bank Street may be less effective than other models and that the University of Pittsburgh and the REC models may be slightly more effective.

VIII. Motor Inhibition

A. Site to Site Differences

Table VII-17 shows site pre-test means and observed gains on the Motor Inhibition test.* Pre-test means for the PV sites range from 4.48 to 5.71, roughly 1.5 individual level standard deviations. Comparison site pre-test means have a range of similar size, from 4.25 to 5.63. Although the comparison site distribution of means is slightly lower overall than the PV distribution the middle 50% of the means of the two groups overlap almost perfectly. The middle range of the PV means goes from 4.79 to 5.19 while the Comparison site middle range is from 4.77 to 5.23--roughly 0.9 standard deviations in both instances.

In terms of "gains" the PV site distribution is considerably tighter than the Comparison site distribution.

*As described in Chapter III a child's score on this test was calculated in a somewhat complicated way. First, in order for a test score to be included in the analysis the child had to answer correctly two or more out of four questions developed to assess whether he understood the words "slow" and "fast". The sample used here contains only children who met this criteria in both the Fall and Spring testings. Additionally, the tester had to certify both test administrations as valid. The test is comprised of three sections: "draw a line", "walk slowly", and "truck pull". In each section the child is asked to complete the task at normal speed and "slowly". We eliminated the "truck pull" task from the analysis for psychometric reasons. A child's score was calculated by taking the log of the sum of the "slow" times (in tenths of a second) for the other two tasks.

TABLE VII- 17

Motor Inhibition

Pre-test means and mean "gains" (post-test mean - pre-test mean) by site for PV and Comparison groups. Site means are unweighted averages of classroom means.

Sponsor	Code	Community	Testing Level	PV Pre-test mean	Comp. Pre-test mean	PV "Gain"	Comp. "Gain"	PV classrooms (#)	Comp. classrooms (#)
Nimnicht	02.04	Duluth	III	5.21		0.54		4	
	02.04	St. Cloud	III		5.11		0.68		2
	02.13	Tacoma	II	4.92		0.32		4	
Tucson	03.08	LaFayette	III	5.19		0.17		4	
	03.08	Albany	III		5.38		0.12		4
	03.16	Lincoln	III	4.73		0.52		4	
Bank St.	05.01	Boulder	III	5.71		-0.20		3	
	05.11	Wilmington	II	4.79		0.66		3	
	05.11	DeLaWar	II		5.14		0.28		3
	05.12	Elmira	III	5.17	4.77	0.49	0.39	3	3
Becker & Engle-mann	07.03	E. St. Louis	III	4.75	5.36	0.49	0.09	3	4
	07.11	Tupelo	III	5.22	5.08	0.01	0.13	4	4
	07.14	E. Las Vegas	II	4.96		0.51		4	
	07.14	W. Las Vegas	II		5.15		0.13		4
Bushell	08.04	Portageville	III	4.70	5.63	0.48	-0.59	3	2
	08.08	Mounds, Ill.	II	4.75	4.72	0.73	0.80	4	2
Weikart	09.02	Ft. Walton B.	III	4.57		0.15		2	
	09.02	Pensacola	III		4.87		-0.08		2
	09.06	Greeley	III	4.90	4.99	0.34	0.49	4	3
	09.10	Seattle	II	4.96	4.86	0.34	0.50	4	3
Gordon	10.02	Jonesboro	III	5.06	5.23	0.45	0.16	3	3
	10.07	Chattanooga	III	4.48	5.05	0.70	0.55	4	4
	10.10	Houston	II	4.64	5.54	0.44	0.13	2	4
[illegible]C	11.05	Washington	III	5.15	4.25	0.15	0.75	2	3
	11.06	Paterson	II	4.89	4.74	0.42	0.37	3	1
	11.08	Johnston Co.	III	5.50	5.22	0.11	0.60	4	4
Pitts-burgh	12.03	Lock Haven	III	4.75		0.26		4	
	12.03	Mifflenburg	III		4.57		0.58		4
REC	20.01	Kansas City	III	5.35		-0.16		3	
Enablers	27.04	Billings	II	4.70		0.72		4	
	27.05	Colorado Sp.	II	5.21		0.62		4	
	27.03	Bellows Falls	II	5.37		0.50		1	

The overall range of PV gains is from -.20 to 0.72, or roughly one standard deviation. The Comparison site range of gains is from -.59 to 0.80, about 1.4 standard deviations. The middle 50% of the PV distribution is only slightly more tightly bunched than the middle 50% of the Comparison group. The PV range is from 0.15 to 0.52 points (.7 standard deviations) while the Comparison site gains range from 0.13 to 0.58 points (0.90 standard deviations).

Relative to the other tests, the variations of mean site gains for the Motor Inhibition test is larger than for the PSI and somewhat smaller than for the Book 4A test. Since the degree of variation of gains appears to be related to the occurrence of clear "effects" in the data this indicates that there may be some effects for the Motor Inhibition test.

When sites within models are examined two models stand out as having a clear pattern of large observed gains. The University of Kansas model has the site with the largest average gains of all the PV sites and a second site with a gain just below the 75th percentile level. In the Enabler model two of the three sites show gains well above the 75th percentile of PV gains and the third site is only very slightly below the top 25 percent. On the low end of the scale the EDC model has two sites slightly below the 25th percentile. Note also that the only site in the REC model shows a loss of -0.16 points, placing it at the very low end of the distribution of site gains. Sites in the other models seem to show little pattern with most models having both relatively high and low scoring sites.

B. Model to Model Differences

The same four models stand out in Table VII-18.* The University of Kansas and the Enabler models have the two largest mean gain scores while EDC and REC show the smallest "gains". The overall range of gains for the PV models is roughly 1.4 standard deviations, from -0.06 to 0.64 points. Since only four children in the REC model received valid scores we will eliminate this model from future discussion of this test. The range without REC is from 0.21 to 0.64, about 80% of an individual standard deviation.

A contrast of the PV model gain means with the means of their Comparison groups shows two statistically significant differences each favoring the Comparison group. The mean gain for the Far West PV group is 0.36 while its

*Recall that the means in Table VII-18 are calculated by pooling all children in all of the sites of a model while the site means in Table VII-17 are means of classroom means. Since there are different numbers of children in different classrooms the two ways of aggregating scores occasionally produce somewhat different results. Thus, the average classroom mean gain for the REC site is -0.16 while the average individual gain is -0.06.

TABLE VII-18

Model Statistics for the Motor Inhibition Test

Column 1 shows the mean gain for FT children in the model.
Column 2 shows the mean gain for Comparison children in model locations.
Column 3 shows the difference between Column 1 and Column 2. (A positive score indicates that FT children gained more than Comparison children).
Column 4 shows the difference between FT and Comparison children in observed-expected gains.
The individual is the unit of analysis.[1]

Model	FT "Gains"	Comparison "Gains"	FT "Gains"-Comparison "Gains"	FT (observed-expected) "gains"-comparison (observed-expected) "gains"

Far West Laboratory	SD=0.46
	0.36
	N=32
		0.50
		0.64
		32
			-0.28*	-0.26

Arizona	0.72 0.34 81	0.44 0.16 36	0.18	0.18
Bank St.	0.69 0.35 21	0.50 0.39 22	-0.04	-0.01
U. of Oregon	0.80 0.31 86	0.66 0.12 68	0.19	0.30*
U. of Kansas	0.52 0.64 27	0.70 0.57 11	0.07	0.05
High Scope	0.57 0.26 34	0.59 0.41 22	-0.15	-0.17
U. of Florida	0.67 0.46 35	0.60 0.37 44	0.09	0.12
EDC	0.58 0.21 76	0.53 0.54 56	-0.33**	-0.32**
U. of Pittsburgh	0.46 0.25 18	0.47 0.59 9	-0.34	-0.35
UNC	0.90 0.[illegible]6 4			
Boulder	0.47 0.51 51			

* Statistically significant at the .05 level
**Statistically significant at the .01 level
[1] All children in the basic analysis sample were used (see Chapter III)

Comparison group has a mean gain of 0.64 yielding a difference of 0.28 points statistically significant at the 0.05 level. Since only one of the two Far West sites has a Comparison group (an off-site comparison) the difference may well reflect uncontrolled sampling bias. Indeed the means for the Far West PV site which has a Comparison group is 0.54, only 0.14 points below its Comparison group mean. Our inclination is to attribute this difference to chance. The second model showing a significant difference is EDC--the difference is 0.33 points favoring the Comparison group. Since all three of the EDC sites have on-site Comparisons and in each instance the PV children "gain" less than the Comparison children there seems good reason to think that this effect may be valid.

When "Observed-Expected" gains are contrasted for the PV and Comparison groups the Far West model does not show a significant difference while the EDC model continues to gain significantly less than its Comparisons. Another contrast also shows significant results in this column. The children in the University of Oregon model appear to gain significantly more than their Comparisons. Inspection of Table VII-17 reveals that this difference of roughly 0.30 points may be due more to the poor showing of the Comparison children than to a strong showing for the Univ. of Oregon PV children. Each of the three

Oregon Comparison sites gains fall below or at the 25th percentile.

Since the Enabler group does not have Comparison sites there is no way of knowing from this table whether its effectiveness is due to the model or to the samples of children in the Enabler sites. The Univ. of Kansas PV model, which has the largest observed "gains", does only slightly better than its Comparison group in the contrasts in Table VII-18. It must be noted, however, that only 11 of the Comparison children in the Univ. of Kansas sites had valid pre and post Motor Inhibition scores.

C. "Adjusted Differences Among Groups"

Tables VII-18, 19 and 20 contain 97 contrasts. Thirteen are statistically significant. The results present a very mixed picture. No model stands out as clearly more effective than others. The results, however, seem to follow three general patterns.

1). Six models (Far West, Univ. of Arizona, Univ. of Oregon, High Scope, Univ. of Florida and EDC) show genuinely mixed results. In some instances the "effect estimates" for these models are positive, in other instances negative. Only one of the 54 estimates for these models is statistically significant. The generally small estimates and the mixed pattern of results indicate to us that there are no compelling differences among these models. With regard to ree of the six models this conclusion should not be surprising. The gain score data in Tables VII-17 and VII-18 indicated that the Univ. of Arizona, High Scope and the University of Florida PV programs were only of average effectiveness. There were, however, indications that the other models might be somewhat different. In particular we pointed out that Far West did not seem to do quite so well as its Comparison group. Our explanation for this rested upon potential differences between the PV and Comparison groups. Based upon the data in Tables VII-19 and VII-20 this explanation appears valid. A second model (EDC) also did not seem as effective as its Comparison group. For EDC we had no ready explanation for the difference. And when EDC is contrasted in the Multivariate Analysis of Variance with is Comparison group (see column 4, Table VII-19) the PV group still appears somewhat less effective, though the difference is not statistically significant. Yet when the EDC PV model is compared with other PV models, with the Comparison classes in general, or with matched Comparison classes there do not appear to be any differences. The third model (Univ. of Oregon) appeared somewhat more effective than its Comparison classes in the gain score analyses. However, when the Univ. of Oregon is contrasted with other groups its effects seem to disappear.

2). Two models (Univ. of Pittsburgh and REC) seem to be systematically less effective than the other models.

TABLE VII - 19

Motor Inhibition

Model "effect" estimates for the test. Columns 1-4 show differences between "adjusted" PV model means and some standard. Column 1 shows the simple contrasts between the PV model "adjusted" means and an unweighted grand mean of the model means for an exact least squares one way ANCOVA. Columns 2 and 3 show regression coefficients for each model in an analysis where all of the comparison classes are pooled together to form a comparison "model". The regression coefficients can be thought of as representing the difference between the "adjusted" PV model means and the "adjusted" Comparison "model" means. Column 2 shows the coefficients for a regression analysis not allowing for separate slope coefficients for the covariates for the different models. Column 3 shows the coefficients allowing for separate model coefficients for the PSI pre-test and for percent prior preschool. Column 4 shows the difference between PV and Comparison group "adjusted" means within models for sites with both a PV and a Comparison group. The estimates are 1 degree of freedom contrasts in the framework of a one way ANCOVA design. Column 5 shows the PV and Comparison n's for column 4 analysis. A note following the Table lists the covariates used in the analysis. In all analyses the classroom is the unit of analysis. See text (Chapters V and VII) for further discussion of the approaches.

Model	Estim. effects around PV un-weighted mean		Estimated effects of PV models against pooled compar. classes[2]		DF contrast PV v. site comp. pooled by models[3]	PV N	Comp. N
			analysis 1	analysis 2			
Far West Laboratory	-0.35	N=8	0.07		-0.18	4	2
Arizona	-0.30	8	-0.12		-0.09	4	4
Bank St.	0.47***	9	0.30*		0.40*	6	6
U. of Oregon	-0.02	11	-0.14		0.36	11	12
U. of Kansas	-0.07	9	-0.06		-0.04	7	4
High Scope	-0.21	10	-0.27*		-0.19	10	8
U. of Florida	-0.12	9	-0.07		-0.17	9	11
EDC	0.22	9	-0.05		-0.23	9	8
U. of Pittsburgh	0.13	4	-0.23		-0.25	4	4
PLC	-0.31	3	0.00				
Enablers	0.56***	9	0.24				
Grand Mean	5.35		5.38		5.38		

TABLE VII-19

(Page 2)

* Statistically significant at the .05 level
** Statistically significant at the .01 level
*** Statistically significant at the .001 level

1. Only PV classrooms are included in this analysis. The multivariate F with the PSI, Book 3D, Book 4A and Motor Inhibition in the analysis is 2.43; significant at the .001 level. The overall univariate F for the Motor Inhibition is 2.62, significant at the .001 level.

2. Both analyses were in the regression framework with the pooled Comparison classrooms as the "dummy variable" left out of the regression. Analysis 1 did not contain separate slope coefficients for the various models. Analysis 2 allowed for separate slope coefficients for PSI pre-score and Prior Preschool Experience. Analysis 1 explained 47.2% of the total variation.

3. Only sites with both PV and Comparison classrooms (on or off-site) were included in this analysis.

Note: All analyses included the following covariables: PSI pre-test mean, Book 3D pre-test mean, Book 4A pre-test mean, mean age, percent black, percent Mexican-American, percent female, mean income, mean household size, teacher experience in Head Start, teacher certification, mean mother's education, percent prior preschool, average staff working conditions, whether the site is El or Ek. In the analyses in column 1 the variable "site administered by CAP or by Public School" was also included. In analyses of the Stanford-Binet, the Stanford-Binet pre-test was also included as a covariate--these analyses used only Level III sites. In analyses of the Motor Inhibition only classrooms with valid Motor Inhibition scores for both fall and spring were included.

TABLE VII-20A

Selected Statistics for Matched Classroom Analysis of the Motor Inhibition for the 5 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance)		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.60
Far West Laboratory	7	0.04	0.03	0.01	0.01	-0.00
Arizona	8	-0.22	-0.04	0.07	0.10	0.15
Bank St.	8	0.08	0.33	0.29	0.28	0.26
Univ. of Oregon	11	-0.27	-0.12	0.02	0.05	0.11
Univ. of Kansas	6	-0.46	0.11	0.34	0.40*	0.50**
High Scope	9	-0.04	-0.09	-0.08	-0.07	-0.06
Univ. of Florida	7	-0.38	-0.18	0.02	0.07	0.15
EDC	7	-0.01	-0.01	0.00	0.00	0.01
Univ. of Pittsburgh	3	0.20	-0.28	-0.39	-0.42	-0.46
REC	3	0.57	0.03	-0.27	-0.34	-0.47
Enablers	6	0.11	0.31	0.25	0.24	0.21

* Statistically significant at the .05 level

** Statistically significant at the .01 level

***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.27. The overall F for the test of homogeneity of the covariate regression coefficient = 2.04.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.52; with r_{tt} estimated as 0.80 the coefficient = 0.65; for r_{tt} = 0.60, the coefficient = 0.86.

TABLE VII-20B

Selected Statistics for Matched Classroom Analysis of the Motor Inhibition for the 4 Factor Match

(See Chapter V for description of matching procedures.) Column 1 shows the number of matched pairs of classrooms for the model. Column 2 shows the covariate means for each model (PV pre-test - Matched Comparison pre-test). Column 3 shows the unadjusted dependent variable means for each model (PV post-test - Matched Comparison post-test). Columns 4, 5 and 6 show adjusted dependent variables for each model (the DV adjusted for the covariate) under three conditions of estimates of the reliability of the covariate (column 3 estimates the reliability as 1.00, column 4 as 0.80 and column 5 as 0.60). The Lord-Porter correction is used to "correct" the covariate for its reliability.

	N's	Covariate Mean PV Pre-Test - Comp. Pre-Test	Unadjusted Difference PV Post-Test - Comp. Post-Test	"Adjusted Differences" (PV Post-Test - Comp. Post-Test) (Adjusted for Pre-Test Covariance		
				Covariate Rel. = 1.00	Covariate Rel. = 0.80	Covariate Rel. = 0.6C
Far West Laboratory	5	0.15	0.41	0.34	0.32	0.29
Arizona	7	0.04	-0.01	-0.03	-0.03	-0.04
Bank St.	8	0.17	0.25	0.17	0.16	0.12
Univ. of Oregon	10	-0.20	-0.24	-0.14	-0.12	-0.08
Univ. of Kansas	6	-0.25	0.24	0.36*	0.39*	0.44*
High Scope	9	0.00	0.16	0.16	0.16	0.16
Univ. of Florida	9	-0.26	-0.07	0.06	0.09	0.14
EDC	9	0.11	-0.08	-0.13	-0.14	-0.17
Univ. of Pittsburgh	3	-0.11	-0.55	-0.50*	-0.49	-0.47
REC	3	0.45	-0.21	-0.42	-0.47	-0.56*
Enablers	8	0.39	0.54	0.35*	0.30	0.23

* Statistically significant at the .05 level
** Statistically significant at the .01 level
***Statistically significant at the .001 level

[1] The overall correlation between PV pre- and Comparison pre-test matched classroom measures = 0.08. The overall F for the test of homogeneity of the covariate regression coefficient = 0.84.

[2] The regression coefficient for the covariate for the analysis with reliability (r_{tt}) estimated as 1.00 = 0.47; with r_{tt} estimated as 0.80 the coefficient = 0.59; for r_{tt} = 0.60, the coefficient = 0.79.

REC will not be considered since the sample size is so small. The University of Pittsburgh model, like REC, has only one site so we cannot make strong claims about its effects. Yet for all but one of the contrasts in Tables VII-18 through VII-20 the estimated effect for this model is negative. Of the negative estimates the range is from -.23 to -.50 or from one-half to one individual level standard deviation. Due to the relatively small number of children (18) and the small number of classes (4) only one of the effects is statistically significant. Our conclusion is to suspend judgement about the effectiveness of the Pittsburgh model for this outcome measure.

3). Three models (Bank Street, University of Kansas and the Enablers) appear to be of above average effectiveness in teaching motor control. Although Bank Street appeared only to be as equally effective as its Comparison group in Tables VII-17 and VII-18 it has a consistently positive pattern of effects in the contrasts in the other tables. All three of the contrasts in Table VII-19 are statistically significant indicating that the Bank St. PV classes generally have children exhibiting greater motor control than the other PV model classes, than the Comparison classes in general and than its Comparison classes located in the same sites (the elimination of one of the PV sites because it lacked a Comparison group of classes accounts for the difference between the effects in column 4 of Table

VII-19 and the effects in Table VII-18). In the matched classroom analyses none of the effects for the Bank St. PV group are statistically significant although they are all positive. The overall range of effects for Tables VII-19 and VII-20 are from 0.12 to 0.47. The University of Kansas also seems to show a generally positive set of effect estimates. We noted earlier that both of the Kansas PV sites had relatively large observed gains. We also noted that the Kansas Comparison sample of children was particularly small. This suggests that we should disregard the contrasts in Tables VII-17 and VII-18 and in column 4 of Table VII-19. If we do this we find positive contrasts for six of the eight other instances with significant results in five of the six positive cases. All six of the positive contrasts are for the matched classroom analyses where the range of effect estimates is from 0.34 to 0.50 points--from two-thirds to a full standard deviation. When, however, the Kansas PV classrooms are contrasted directly with the other PV model classrooms or with the Comparison classes overall the estimated effect for the model is essentially zero. This contradiction in results may stem from the very low pre-score means for the Kansas PV sites.

As we noted earlier the Enabler sites all seem to provide greater overall gains than average on the Motor Inhibition. Moreover, when the Enabler model is contrasted with the other models its effect is the largest. Finally in the matched analyses the effect for the Enabler model is always positive and while significant in only one of the contrasts never has an effect of less than 0.40 standard deviations.

We tend to be optimistic about positive effects for both the Bank Street and University of Kansas PV models though we cannot reach a firm conclusion. Our optimism stems in part from the results presented here and in part from the fact that it makes sense for both of these models to have an effect on a child's motor control. Although a Bank Street classroom is not structured in the same sense as a University of Kansas classroom (with academic drill) it generally has a quite formalized set of conventions regarding the nature of adult-child and child-child interactions. Children are taught to have respect for others and to be self-conscious about their aggressive behavior. Such instruction should bear a relation to motor control and the inhibition of impulsive behavior. The Motor Inhibition test should tap this dimension. Similarly, the reinforcement principles effected by the University of Kansas model might tend to encourage children to increase their motor control. We have no explanation for the apparent success of the Enabler model on the Motor Inhibition.

In summary, there do not seem to be any models which are definitively more or less effective in aiding in the development of motor control. There is some indication, however, that the University of Pittsburgh model may be relatively less effective and that Bank Street, the University of Kansas and the Enabler models may be relatively more effective than the other models.

IX. Summary of the Effectiveness of Different Planned Variation Models

Table VII-21 crudely summarizes our findings regarding differential model effectiveness. The eleven PV models are the rows of the table while the five outcome measures are each represented by a column of the table. The cell entries indicate effectiveness relative to the other PV models and to appropriate conventional Head Start classrooms. Four categories are used to indicate whether the model is: a). Probably less effective than average; b). Of average effectiveness; c). Probably more effective than average; and d). Almost certainly more effective than average. Six general conclusions may be reached after inspection of this table.

1). We began this chapter with a major expectation: that there will be few strong differences among the models in effectiveness as assessed by our five outcome measures. By and large this expectation was realized. Table VII-21 clearly indicates that for each of the outcome measures we have classified the majority of the models as having average effectiveness. Moreover, no model stands out as either more or less effective than the others on more than two of the five outcomes. In the crudest terms there are no overall winners or losers.

TABLE VII-21

Summary of Planned Variation Model Effectiveness on Five Outcome Measures

Zero (0) indicates model is of average effectiveness on outcome measure.

Minus (-) indicates model may be of below average effectiveness.

Plus (+) indicates model may be of above average effectiveness.

Double plus (++) indicates model is probably highly effective.

Model	Book 3D	Book 4A	PSI	Stanford Binet	Motor Inhibition
Far West Laboratory	0	0	0	0	0
Arizona	0	0	0	0	0
Bank St.	0	0	0	-	+
Univ. of Oregon	0	+	0	0	0
Univ. of Kansas	0	++	0	0	+
High Scope	+	0	0	++	0
Univ. of Florida	-	0	0	0	0
EDC	0	0	0	0	0
Univ. of Pittsburgh	0	+	0	+	-
REC	-	-	0	+	0
Enablers	0	-	0		+

2). A second more tentative expectation suggested early in the chapter was that models which emphasized academic drill combined with systematic reinforcement would be more effective than other models on the four cognitive outcome measures. This expectation was realized only for one of the four cognitive measures. Only for the Book4A measure--a test assessing knowledge of letters, numerals, and shape names--is there evidence of greater effectiveness for the models emphasizing drill and reinforcement. The University of Kansas model is the clearest example of this finding. We found it to be clearly superior to all of the other models and to the Comparison classes in its effectiveness in raising Book4A test scores. The two other models we rated as emphasizing academic drill (University of Oregon and University of Pittsburgh) both appear to be above average in their impact on this test. No other model has an above average effect for this test.

On the other cognitive tests there is no indication of special effectiveness of these three models. Only the University of Pittsburgh model on the Stanford Binet shows an other than average effect. These findings are at some variance with the findings of other researchers in the preschool area (see Bissell, 1970 and White, et al, 1972). These researchers indicated that there may be a general positive effect of structured academic emphasis and drill on cognitive tests. Our data, however, indicate that the effect is specific rather than general. In particular it appears as if this approach may be more effective for imparting information that is easily taught through systematic drill while it is only of average effectiveness in other cognitive areas. Of the four cognitive tests the Book4A test most clearly assesses specific skills. The other tests, particularly the PSI and the Stanford Binet, assess general information and cognitive functioning.

3). *One model clearly stands out as more effective than the others in raising Stanford Binet test scores. The High Scope PV model appears to increase Stanford Binet scores by an estimated twelve to fifteen points, roughly 0.9 individual level standard deviations.* The average effect of other PV and Comparison models is on the order of two to three points or roughly 0.2 standard deviations. The effect of the High Scope model is particularly strong in one Southern rural site where the measured average gain is slightly over thirty points. Although we can probably attribute some of the measured gain to tester and regression effects the "corrected" gain is still on the order of a very substantial twenty points. Preliminary analyses of the item profiles of children in the High Scope sites indicates that the gains may partly be attributed to the emphasis of the High

Scope model on the concepts of similarities and differences. (See Butler, in preparation as a separately bound appendix to this report.)

The particular effectiveness of the High Scope model on the Stanford Binet does not appear to generalize to the other outcome measures used here. For three of the four remaining tests the model appears to be of only average effectiveness. On the fourth test, Book3D, there is some indication that the High Scope model may be of above average effectiveness but no firm conclusion may be reached from the data.

4). Two of the eleven models (University of Pittsburgh and REC) account for 40% of the 15 cells in Table VII-21 where there is an indication that a model has other than average effectiveness on an outcome measure. Pittsburgh appears above average on the Book4A and Stanford Binet tests and below average on the Motor Inhibition test. REC appears below average on the Book3D and Book4A tests and above average on the Stanford Binet. No other model is rated as other than average on more than two of the measures. Three things are common to REC and Pittsburgh. Each uses some form of programmed instruction, each was a first year model in 1970-71, and each has only one site in this study. Although the first two common elements may be important our inclination is to view the fact that each model has only one site as the principal reason that these models have more than their share of "other than average" effects. As we note throughout the chapter it is common for models with two or more sites to show considerable site to site variation in effects. This may be due to differential effectiveness of the models in different sites or to uncontrolled biases in our data. Whatever the reason our inclination is to be very skeptical about attributing clear effects to any model with only one site.

5). All models are rated as showing average effectiveness on the PSI test. We had not expected this result since our preliminary analyses of the PSI indicated that it is probably our most reliable measure. In retrospect, however, we suspect that the reason for the lack of clear differences among models on the PSI is due to the nature of the test itself. The PSI was developed as a general test to assess the overall impact of preschools on children. As such it attempts to measure a wide range of skills probably rendering it relatively insensitive to particular differences among curricula. Thus it is probably more appropriate to the tasks of assessing the overall average impact of preschools (see Chapter IV) and of individual differences among children (see "Cognitive Effects of Preschool Models on Different Types of Children").

6). Three models (Bank Street, the University of Kansas and the Enabler models) appear to be above average in effectiveness as assessed by the Motor Inhibition test. We argue in section VIII of this chapter that there are substantive reasons for the result relating to the curricula of Bank Street and the University of Kansas. We do not know why the Enabler model appeared more effective than most other models.

Chapter VIII

MAJOR CONCLUSIONS

This chapter briefly summarizes major conclusions of the report. An extensive summary of this report and the other three preliminary reports on Head Start Planned Variation, 1970-71 is being prepared by the Huron Institute.*

Three main questions were addressed in this report:

1. What are the short term effects of a Head Start experience on children?

2. Are there discernable differences between the effects on children of a Head Start Planned Variations experience and a conventional Head Start experience?

3. Do Planned Variation models differ in their effects on Head Start Children?

Five measured outcomes were used to assess each question. The PSI, is a general standardized achievement test for preschool children. The NYU Book 3D and NYU Book 4A are tests of specific achievement areas. The Stanford-Binet is a well known test of general "intelligence". The Motor Inhibi-

*The other three reports in this series are concerned with the quality of the data, the issue of implementation and interactions between program and child characteristics which affect cognition outcomes.

tion test assesses a child's ability to control motor behavior.

With regard to the question of short term effects of Head Start we reach four conclusions. (See Chapter IV for details)

1. The Head Start experience substantially increased children's test scores on all five outcome measures. On four of the five outcome measures children's scores were estimated to increase "naturally" over the seven or eight months of the Head Start program. Thus, even had the children not been exposed to Head Start, their scores would have risen. For two of these measures (PSI and Book 3D) the Head Start experience was estimated to double the "natural" rate of growth. For two other measures (Book 4A and the Motor Inhibition tests) the Head Start experience was estimated to better than triple the "natural" rate of growth. Increments attributable to Head Start ranged from 0.26 standard deviations (for the Motor Inhibition test) to 0.85 standard deviations (for the Book 4A test). On the fifth measure, the Stanford-Binet, our estimates indicate that the scores of children in this sample would have "naturally" decreased by about 0.20 standard deviations had they not attended Head Start. The Head Start experience arrested this apparent decrease and further increased

<u>Head Start participants' Stanford-Binet scores by roughly 0.40 standard deviations.</u>

2. Children who had a prior preschool experience gained less overall ("natural" + Head Start related growth) than children for whom 1970-71 Head Start was their first year of preschool. This effect held for all outcome measures and for most of the subgroups studied in Chapter IV. If, however, we allocate the total gains for the two groups of children between "natural growth" and the Head Start experience, we find that the effects attributable to Head Start are roughly equal for children with and without prior preschool experience. This indicates that the expected "natural growth" for children with prior preschool experience is less than for children without prior preschool. The prior preschool experience appeared to reduce differences in test scores between children of different ages. In other words, a common preschool experience partially overcomes the effect of age differences among children on the five outcome measures. Some support for this notion comes from the fact that variances on four of the five outcome measures are somewhat smaller at post-test time than at pre-test time. This indicates that differences

among children are less at the end of the preschool program than they are at the beginning of the program. Preschools may have a "fan-close" rather than a "fan-spread" effect on children.

3. Children who would enter first grade (E1) directly from Head Start tend to gain more than children who would enter kindergarten (Ek) directly from Head Start on the Book 4A, Book 3D, PSI and Stanford-Binet tests. On the Motor Inhibition test the Ek children gained more. (The average age of E1 children when they entered Head Start was 65 months -- Ek children were roughly one year younger.) The greater gain for E1 children was most pronounced for the Book 4A test and least for the Stanford-Binet. When the gains attributable to Head Start were examined, the effect appears to strengthen, though they are still small for the Stanford-Binet. These effects are probably due to a combination of two things. First, the larger gains attributable to Head Start for E1 children on the cognitive measures and particularly the Book 4A test (a measure of letters, numerals, and shape names) may be due to older children's advanced academic readiness. Second, there may be a greater interest by Head Start teachers in E1 sites in preparing children for reading and arithmetic.

4. There seem to be no consistent differences among Mexican American, black and white children in their Head Start gains on the five outcome measures.

In Chapter IV we discuss the methodological procedures used to arrive at these conclusions. Since we did not have a group of "control" children (children who did not have the benefit of an Head Start experience) our estimation procedures relied on natural variations in proscores for children of different ages. The reader, therefore, is warned to treat these data as rough estimates and to evaluate for himself the assumptions of the procedures.

The second major question regards overall differences in effects for Planned Variation and conventional Head Start programs. This question is addressed in Chapter VI. At the beginning of that chapter we argue that the question has very little importance. For while we might expect there to be differences among PV programs in their effects on the five outcome measures, we have little reason to suspect that there should be systematic differences between an overall PV effect and an overall effect of conventional Head Start programs. This question, like most total program impact questions, totally obscures systematic differences among treatments.

The sole rationale for studying the question was to determine whether the extra funds allocated to PV Head Start programs had a consistent effect on the measured outcomes. Our conclusion supports the findings of a large number of recent research efforts which have failed to detect any systematic relationship of gross expenditures to variations in outcomes. We conclude there are no differences in effects between the PV programs (taken together) and the Comparison Head Start programs on any of the five outcome measures.

The third question addresses differences among PV programs in their effects on Head Start children. We reach four major conclusions in this area. (See Chapter VII for details).

1. There are a relatively small number of differences in effects among PV programs that are of sufficient stability and size for us to reject a null hypothesis of no differences. This is a conservative statement. We recognize that there may be many more "true" differences among the models on the five outcome measures than we report. We also recognize that there are undoubtably outcome differences among the models in domains where we lacked measures.

   The few differences we found are scattered among different models and different outcome measures. No model stands out as being overall more or less effective than the other models.

2. One tentative expectation in our analysis was that models which used systematic reinforcement procedures and which emphasized academic drill would have a greater effect than the other models on cognitive outcomes. On three of the cognitive tests (PSI, Book 3D and the Stanford-Binet) this expectation was not confirmed. On the fourth cognitive test (Book 4A) there is a strong indication that the expectation is valid. Of the three models which fit this criterion, one (University of Kansas) stands out as being more effective than all other models in imparting knowledge of letters, numerals, and shape names as measured by Book 4A. The effect of the Kansas model was on the order of 0.75 to 1.0 standard deviations. The two other models also fitting the criterion (University of Oregon and University of Pittsburgh) were clearly above the averages of the other models in effectiveness on the Book 4A outcome measure.

3. One model (High Scope) was clearly more effective than other models in producing gains on the Stanford-Binet. We estimate that "true" gains for children in the High Scope model averaged roughly 12 to 15 points while "true" gains for the other models averaged 2 to 4 points. An Appendix to this report attempts to pinpoint reasons for the success of the

High Scope model on the Stanford-Binet.

4. Other findings in the data were less dramatic than the University of Kansas' model effect on Book 4A and the High Scope model's effect on the Stanford-Binet. On one outcome measure (the PSI), we found no model which departed significantly from the others. For the other outcome measures we found indications that two or three models showed either above or below average effectiveness. In most instances, the "effects" which differed from the average made sense. The "effects" appear to be related to the structure and content of the models. <u>One tentative conclusion from this is that differential model effects are more easily discerned if the outcome measures tap specific rather than general cognitive growth.</u> The lack of differential "effects" for the PSI indicates that tests designed to assess the overall impact of a preschool experience may be insensitive to variations in curricula. It appears that we need more highly specific outcome measures like the Book 4A test to obtain a reasonable assessment of model to model differences in outcome effectiveness.

REFERENCES

Averch, H. A., et al. How effective is schooling? A critical review and synthesis of research findings. Santa Monica, California: The Rand Corporation, 1972.

Bachman, J. & Johnston, L. Understanding Adolescence. Boston: Allyn and Bacon, 1972.

Bissell, J. The cognitive effects of preschool programs for disadvantaged children. Unpublished doctoral dissertation, Harvard University Graduate School of Education, 1970.

Boyd, J. Project Head Start, Summer 1966: Facilities - Resources of Head Start Centers. Princeton, New Jersey: Educational Testing Service, 1966.

Cicerelli, G., et al. The impact of Head Start: An evaluation of the effects of Head Start on children's cognitive and affective development. (Westinghouse Learning Corporation and Ohio University. Contract b89-4536 with the Office of Economic Opportunity) Washington, D.C.: Office of Economic Opportunity, 1969.

Coleman, James, et al. Equality of educational opportunity. U. S. Department of Health, Education, and Welfare, Office of Education, OE-38001, National Center for Educational Statistics. Washington, D.C.: U. S. Government Printing Office, 1966.

Coller, A. & Victor, J. Early Childhood Inventories Project. New York City: Institute for Developmental Studies, New York University School of Education, 1971

Datta, L. A report on evaluation studies of Project Head Start. Paper presented at the 1969 American Psychological Association Convention, Washington, D.C.

Di Lorenzo, L., et al. Prekindergarten programs for educationally disadvantaged children. Final report Washington, D. C.: U. S. Office of Education, 1969.

Dittman & Kyle Head Start Planned Variation Case Studies. University of Maryland, 1970-71.

Jencks, C., et al. Inequality: A Reassessment of the Effect of Family and Schooling in America. New York: Basic Books, 1972.

Karnes, Merle. Research and development program on preschool disadvantaged children. Final Report, U. S. Department of Health, Education and Welfare, 1969.

Levine, et al. California Preschool Competency Scale Manual. Palo Alto, California: Consulting Psychologists Press, Inc., 1969.

Maccoby, E. E., et al. Activity level and intellectual functioning in normal preschool children. Child Development, 1965, 36, p. 761-770.

Maccoby, E. E. & Zellner, M. Experiments in primary education: Aspects of Project Follow-Through. New York: Harcourt Brace, 1970.

McMeekin, R. Costs Analysis of Planned Variation Head Start Models. The Huron Institute, in preparation.

Mosteller, F. & Moynihan, D.P., eds. On Equality of Educational Opportunity. New York: Random House, 1970.

Plowden, Children and Their Primary Schools. Central Advisory Council for Education, London: Her Majesty's Stationery Office, 1967.

Porter, A. G. How Errors of Measurement affect ANOVA, ANCOVA and regression analyses. Paper presented at the 1971 AERA convention.

Rainbow Series. U. S. Department of Health, Education and Welfare, Office of Child Development, 1972.

Rubin, D. Matching to remove bias in observational studies. Technical Report no. 33, December 14, 1970, Department of Statistics, Harvard University.

Shaycroft, Marion. The statistical characteristics of school means. In Flanagan, et al. Studies of the American High School. University of Pittsburgh, 1962.

Stanford Research Institute. Evaluation of the National Follow-Through Program 1967-71. , Menlo Park, California: draft

Stanford Research Institute. Implementation of Planned Variation in Head Start. Menlo Park, California: 1972.

Stearns, M. S. Report on preschool programs: The effect of preschool programs on disadvantaged children and their families. Washington, D. C.: Office of Child Development, 1971.

Terman, L. M. & Merrill, M. A. Stanford-Binet Intelligence Scale: Manual for the Third Revision Form L-M. Boston: Houghton-Mifflin, 1960.

The Study of Natural Variations in Head Start 1969. Office of Child Development, 1971, draft.

Tukey, J. W. Discussion on Temporal changes in treatment - effect correlation. In Glass (Ed.) Proceedings of the 1971 Invitational Conference on Testing Problems, Princeton, New Jersey, in press.

United States Civil Rights Commission. Racial Isolation in Public Schools. Washington, D.C.: 1967

Weikart, D. P. Preschool intervention: A preliminary report of the Perry Preschool Project. Ann Arbor, Michigan: Campus Publications, 1967.

Weikart, D. P. Relationship of curriculum, teaching and learning in preschool education. In J. C. Stanley Preschool Programs for the Disadvantaged. Baltimore: Johns Hopkins University Press, 1972.

White, S., et al. Federal Programs for Young Children: Review and Recommendations. The Huron Institute, 1972.

Appendix A

DESCRIPTION OF VARIABLES

A. Outcome Measures

All outcome measures used in this report exist both as pre and post-tests. The measures are briefly described in Chapter II and fully described in "The Quality of the Head Start Planned Variation Data". In order to be included in an analysis, a classroom (or a child depending on the unit of analysis) must have had a valid pre and post-test score on the particular outcome measure. Validity was assessed by the tester. The outcome measures are:

1. NYU Book 3D -- an achievement test assessing knowledge of pre-math, and pre-science concepts and of prepositions.

2. NYU Book 4A -- an achievement test assessing knowledge of letters, numerals and shape names.

3. Preschool Inventory -- a general achievement test designed to assess the overall impact of a preschool experience.

4. Stanford-Binet -- a generalized measure of "intelligence".

5. Motor Inhibition -- a measure assessing a child's ability to control his motor behavior.

B. Child Characteristics

All measures of child characteristics are taken from the Classroom Information form (see Chapter II). Child characteristics were calculated by on the individual level and on the classroom aggregate level. The description below is for the individual level. At the classroom level, a mean of the characteristics for the children in the classroom was computed. In instances where the characteristic was a binary variable (e.g. sex), the classroom mean can also be thought of as a proportion or percentage.

1. Sex -- Females were coded 1 and males were coded 0.

2. Race -- Generally two dummy variables were used indicating whether the child was: a) Black or not, and b) Mexican American or not.

3. Mother's Education -- A variable assessing the number of years of schooling a child's mother has completed. The range is from 0-20 indicating number of years of school.

4. Family Income -- A variable assessing income coded into units of $100.00. The maximum value for the variable is 99, standing for an income equal to

or greater than $9900.00.

5. Family Size -- A variable indicating the number of persons living in the child's household.

6. Prior Preschool Experience -- A variable indicating whether or not a child had any preschool experience prior to entering Head Start in 1970.

7. Age -- A variable indicating a child's age in months on October 1, 1970.

C. Teacher Characteristics

All Teacher Characteristics variables were taken from the teacher information form (see Chapter II).

1. Teacher Experience -- A variable indicating the number of years of experience that the teacher has in Head Start prior to 1970.

2. Teacher Certification -- A variable indicating whether or not a teacher was certified by the city or state to teach in a preschool or public school.

3. Teacher Race -- A variable indicating whether the teacher was white.

4. Average Staff Working Conditions -- A summary measure of teachers' evaluations of their working conditions.

D. Experience of Teacher Aide

This variable was taken from the Teacher Aide questionnaire and assesses the number of years of experience the Teacher Aide has in Head Start.

E. Site Characteristics

1. "Site administered by a CAP or Public School". This variable was taken from an Head Start Director's questionnaire. It assesses the administrative structure of the Head Start Center -- whether it is administered by a community action program or by the public schools.

2. "Site is either an entering first or an entering kindergarten site". This variable was taken from the Head Start Director's questionnaire. It indicates whether a majority of children in a site will attend first grade or kindergarten directly after Head Start.