DOCUMENT RESUME

ED 093 497 PS 007 382

AUTHOR Weisberg, Herbert I.
TITLE Short Term Cognitive Effects of Head Start Programs: A Report on the Third Year of Planned Variation--1971-72.
INSTITUTION Huron Inst., Cambridge, Mass.
SPONS AGENCY Office of Child Development (DHEW), Washington, D.C.
REPORT NO OCD-H-1926
PUB DATE Jun 74
NOTE 509p.; For related document see ED 082 834

EDRS PRICE MF-$0.90 HC-$24.60 PLUS POSTAGE
DESCRIPTORS Academic Achievement; Age Differences; Analysis of Covariance; *Cognitive Development; *Compensatory Education Programs; Data Analysis; Evaluation Criteria; Factor Analysis; Intervention; *Methodology; *Preschool Programs; *Program Evaluation; Racial Differences; Social Differences; Standardized Tests
IDENTIFIERS Planned Variation; *Project Head Start

ABSTRACT

This report focuses on three main questions: (1) To what extent does a Head Start experience accelerate the rate at which disadvantaged preschoolers acquire cognitive skills? (2) Are the Planned Variation models, simply by virtue of sponsorship more effective than ordinary nonsponsored Head Start programs? and (3) Are some Planned Variation models particularly effective at imparting certain skills? The first chapter gives an overall picture of the Head Start Planned Variation study, while the second chapter summarizes data concerning background characteristics and distribution of test scores. Chapter 3 provides a general discussion of methodological issues and some of the major difficulties resulting from the study design. Chapters 4-7 attempt to present a picture of the pattern of overall effects of various programs through ranking analysis, residual analysis, analysis of covariance, and resistant analysis. The final chapters explore the question of whether the relative effectiveness of various programs is related to certain child background characteristics, such as sex, ethnicity, age, prior school experience, and mother's education. One major conclusion drawn as a result of the intermodel comparisons was that Head Start programs are quite homogeneous in their ability to promote general cognitive development. (CS)

SHORT TERM COGNITIVE EFFECTS OF HEAD START PROGRAMS: A REPORT ON THE THIRD YEAR OF PLANNED VARIATION - 1971-72

HERBERT I. WEISBERG

JUNE, 1974

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
Richard Elmore
Helen Featherstone
Nathan Fox
David Gordon
Deborah Gordon
Sharon Hauck
Gregory Jackson
Carol Lukas
Robert McMeekin
Anne Monaghan
David Napior
Ann Taylor
Deborah Walker
Jack Wiggins
Cicero Wilson
Cynthia Wohlleb
Joy Wolfe
Stanley Yutkins
Diane Zipperman

Acknowledgments

As author of this report, I accept full responsibility for all conclusions and the correctness of the statistical analyses on which they are based. A study of this magnitude would, however, have been impossible to carry out without the support of many others. I feel it would be remiss of me not to mention those whose help was most valuable.

I wish to thank the project director, Marshall Smith, for general guidance on substantive matters and valuable discussion of methodological issues. I wish to thank Anthony Bryk for helpful discussions of methodological issues and for his willingness and ability to translate abstract ideas into concrete computer programs. Gregory Jackson provided valuable technical assistance, particularly in implementing the ranking analysis described in Chapter IV. Sharon Hauck helped develop and implement the resistant analysis described in Chapter VII, and wrote Appendix E. Thomas Cerva provided general technical assistance. I wish to thank Jane David for valuable comments on earlier versions of this report which significantly improved the presentation. Finally, thanks go to the Stanford Research Institute for carefully collecting and transmitting to us the data on which this study is based.

Herbert I. Weisberg
Cambridge, Massachusetts
August, 1973

TABLE OF CONTENTS

Chapter I

INTRODUCTION

Background

In 1965 Project Head Start was initiated with great fanfare and optimism. It was thought that since the "disadvantaged" child arrives at school handicapped by an educationally impoverished home environment, he starts out behind the middle class child in terms of basic cognitive and socio-emotional development. This initial gap is then propagated throughout the child's school career, leading ultimately to large deficits in educational attainment and career success. It was hoped that a summer-long or year-long preschool compensatory program would give disadvantaged children the "head start" they need to start off school on an equal footing with middle class children and progress from there on at a comparable rate.

The basic assumption justifying Head Start, then, is that a limited intervention which alters the child's environment at some point can permanently influence his potential for future educational achievement. If we accept this premise, we must still ask certain key questions. How extensive an intervention is necessary to effect permanent change? When

in the child's life should the intervention begin, and how long must it persist? Around the time when Head Start was conceived, there was considerable optimism that basic intelligence was malleable even into the early elementary years. This belief, coupled with several encouraging reports on preschool compensatory programs (e.g., Weikart, et al., 1964; Gray and Klaus, 1963; Bereiter, et al., 1965), fed the hope that a relatively inexpensive solution to the problem of educational disadvantage might be feasible.

Head Start is a program lasting a few hours per day over the course of a few months which attempts to rectify the cumulative effects of four or five years of deprivation. Early evaluations, most notably the Westinghouse-Ohio national evaluation (1969) showed modest positive results. Moreover, there was evidence (e.g., Wolff and Stein, 1965; Holmes and Holmes, 1966) that Head Start effects were disappearing in the early elementary years. Head Start itself does not appear to be the solution to educational disadvantage. It may, however, be valuable as part of a more comprehensive approach involving a more extensive intervention in children's lives.

As a step in this direction, the Follow Through program was started in 1967.

By 1969 there were over 170 school districts with Follow Through programs. Follow Through attempted to enrich the curricula of early elementary (grades K to 3) programs, particularly for children with Head Start experience. By consolidating and building upon their preschool experiences, the program hoped to be able to influence permanently children's chances for success in school.

According to Smith and Bissell (1970), Head Start centers were practically autonomous and programs varied greatly. "Although lists of goals and objectives were developed by OEO, a laissez-faire attitude predominated." When Follow Through was originated, it was felt that carefully designed and implemented programs based to some extent on theories of child learning offered greater hope of success. A number of well-defined curricular programs, or "models" were developed by sponsors, who were individuals or organizations with expertise in early childhood education. By fall 1969, most Follow Through schools had adopted one of these models. Thus, there was deliberately "planned variation" in the models implemented. By comparing the effects of these models, information could hopefully be obtained on what kinds of curricula produce what kinds of

results. Children entering selected Follow Through schools during the years 1969-72 were to be tested at entrance and followed through grade 3.

In 1969 the planned variation approach was adopted for an experiment in Head Start. The Head Start Planned Variation (HSPV) study was to involve 3 cohorts of children in programs during the school years 1969-70, 1970-71, and 1971-72. Several of the models used in Follow Through were to be implemented at 2 or more sites throughout the country. There were three restrictions on the way this was to be done:

1. Sites were to contain pre-existing Head Start programs.
2. Children in a program must live in an area served by a school with a Follow Through program.
3. The Head Start model must be the same as the Follow Through model in that area.

Children in Planned Variation (PV) programs were to be tested at the beginning of the program in the fall and at the end in the spring. Children in some Head Start Programs without a Planned Variation model (NPV) were also tested, so that a comparison could be made between model effects and those of typical non-sponsored Head Start programs. In addition, during the final year (1971-72) only,

a group of "Control" children who were not in any preschool program was tested in fall and spring, allowing the possibility of estimating the absolute effect of a Head Start program.

This report is concerned with the data from the third year of the HSPV study, the academic year 1971-72. As the result of improvements in the data collection throughout the three years of the study, these data are potentially more informative than those collected on the first two cohorts. In particular, the battery of tests is more extensive and hopefully more appropriate for program evaluation. Additionally, we have the Control children who were not enrolled in any preschool program.

We had originally hoped to study the impact of Head Start on both cognitive and socio-emotional development. For reasons to be detailed later in this chapter, the non-cognitive measures used in the study proved unsuitable for use in program evaluation. This is not a reflection on those at the Stanford Research Institute and the Huron Institute who designed the study. Good measures of affective characteristics for preschool children which can be routinely administered within the constraints of the Head Start setting are simply unavailable. According to Walker (1972) "until the major theoretical questions and issues

are answered within a comprehensive theory of socio-emotional development, socio-emotional measures for young children cannot be meaningfully developed." This report will focus, then, on the short-term cognitive effects of various types of Head Start programs.

The remainder of this chapter consists of six sections. The first outlines the major questions addressed in this report. The second contains brief descriptions of the HSPV models. The third describes the study design. The fourth discusses the data collected and explains which measures we have selected for our evaluation and why. The fifth section summarizes briefly the relevant findings of other reports in this series. Finally, we present a brief summary of this chapter and overview of the remainder of the report. We do not present an overall review of the literature on the effects of preschool programs. The interested reader is referred to Stearns (1971) and White et. al., (1972).

Major Questions

Head Start is much more than a training program to prepare disadvantaged children to perform better in school. It provides a wide range of services, including health care,

nutritional benefits, training in social skills, early detection of severe problems, and a focus for parent involvement with the community. Ideally, in evaluating the full impact of a Head Start program, we would need to measure many aspects of an individual throughout his life. Even if we could circumvent the measurement problems and practical difficulties in setting up an experiment designed to do this, it would take literally a generation to complete, by which time the results might no longer be relevant.

The questions which we can hope to answer are restriected to effects which are relatively short-term and limited to outcomes which can be measured relatively well, i.e. short-term cognitive effects. Improving the ability of disadvantaged children to acquire academic skills in school was one goal of Head Start. Although recent work by Jencks et al. (1972) suggests that such academic performance may not be so strongly related to future financial success as many thought, it is generally agreed to be a worthwhile and important goal. It can be argued that short-term preschool program effects do not ensure later school success. It seems reasonable, however, that a program which substantially raises cognitive skills for preschoolers can have a lasting influence if appropriately augmented during the early school years. The study of Follow Through currently being conducted by Abt Associates (1973) should

help us to understand to what extent this is realistic.

In this report our analyses will focus on three main questions:

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged pre-schoolers acquire cognitive skills?
2. Are the Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?
3. Are some PV models particularly effective at imparting certain skills?

These are essentially the same questions addressed by Smith (1973) in his report, on the 1970-71 cohort data although the battery of tests employed and the statistical analyses used are different. We will also be concerned with evidence of interactions between program effectiveness and specific child characteristics. Given the limitations imposed by the design, the methodological problems involved in eliciting such interactions are formidable. Any conclusions can only be in the form of suggestions rather than strong assertions.

Model Descriptions

There were 11 models for which child outcome information was collected during 1971-72. These models may be thought of as varying in terms of a number of dimensions along which preschool programs can be ranged. White et al., (1972) summarize the literature on such classification schemes.

For our analysis, one of the most important of these dimensions is the extent to which the acquisition of academic skills is stressed through formal, highly-structured activities. Traditional preschools and Head Start centers vary in their stress on such activities. Many reflect a developmental approach which tries to create a milieu in which the child is encouraged to explore and learn from his environment, rather than respond to demands leading to cognitive growth in a pre-specified way. At least three of the eleven models are consciously concerned with the development of specific academic skills useful in the early school years. These are the Oregon, Kansas, and Pittsburgh models.

While the models do vary along certain important dimensions, relative to the condition of no preschool program their similarities far outweigh their differences. According

to Smith (1973),

> 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.

We conclude this section with brief descriptions, taken from Smith (1973), which attempt to give the flavor of the various programs. For more complete descriptions, see Maccoby and Zellner (1970) and the Rainbow Series, published by the Office of Child Development (1972).

The Enabler Model
Office of Child Development

Sponsor Contact: Jenny Klein

The Enabler Model is not really a curricular model. Rather it is an approach involving the total community which is built 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 staff and parent participation in policy making, program planning and classroom operation.

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.

Design of the Study

During the 1971-72 academic year there were 29 tested Head Start locations. There were 28 locations containing one of the 11 models. Of these PV locations, 11 also contained non-sponsored (NPV) classrooms. One place (Des Moines) contained only NPV classrooms. In addition there were three places containing groups of Control children not enrolled in any preschool program throughout the year. These were Huntsville, San Jose, and Sacramento. These children were contacted by direct recruitment or from Head Start waiting lists.

The numbering system used to identify locations is somewhat complicated. Each location has a four digit code. Since each Head Start location is located in an area served by one of the Follow Through models, the code used to identify it is the same as that assigned by SRI to the Follow Through site, with the exception of the Enablers model, which is unique to HSPV. The first two digits identify the model, and the second two identify the site uniquely. Thus, 0711 refers to Follow Through site number 11 in the Oregon model (07). The three control locations were given codes of 2801 (Huntsville), 2802 (Sacramento), and 2803 (San Jose).

During 1970-71, each model was implemented in at least one location with both PV and NPV classrooms. In

1971-72 only seven of the models has such comparisons. In most of our analyses we pool all NPV classrooms and treat them as a representative sample of non-sponsored programs for purposes of comparison with the various PV models and the Control children.

Let us define a "site" as a group of children in a particular location undergoing a particular kind of pre-school experience. Thus, we have 28 PV sites, 12 NPV sites, and 3 Control sites, for a total of 43 sites. These 43 constitute a convenient set of units of analysis for some purposes. If this study did not exist in the context of other Head Start and Follow Through studies, we might number these 43 sites in some convenient and logical way. As it is, we have decided to retain the old SRI coding system. Thus, we are stuck with the awkwardness of having, for example, a site 0711 PV and an 0711 NPV. A complete description of the design is provided by Tables I-1 and I-2.

For convenience throughout this report we shall often refer formally to the NPV children and the Control children as model or program groups. They are, of course, not models or programs in the same sense as the PV models, but it is awkward in terms of reporting results to continually make this distinction. Thus, from an experimental viewpoint we can think of our 43 sites distributed across 13 programs to be compared (11 PV models, NPV, Controls).

Table I-1

PLANNED VARIATION SITES

Model	Site	Code	# Tested Classes	Year Joined
Far West	Duluth	0204	6	70
	Salt Lake	0209	6	69
	Tacoma	0211	6	70
U. of Arizona	Lafayette	0308	6	69
	Lakewood	0309	6	69
	Lincoln	0316	6	70
Bank Street	Tuskegee	0510	6	69
	Wilmingon	0511	6	69
	Elmira	0512	6	70
U. of Oregon	Tupelo	0711	4	69
	E. Las Vegas	0714	5	70
U. of Kansas	Portageville	0804	4	69
	Mounds	0808	5	70
High/Scope	Ft. Walton Bch.	0902	5	69
	Central Ozarks	0904	6	69
	Greeley	0906	4	70
U. of Florida	Jonesboro	1002	3	69
	Chattanooga	1007	6	70
	Houston	1010	5	70
EDC	Paterson	1106	7	70
	Johnston Co.	1108	6	69
U. of Pittsburgh	Lock Haven	1203	6	70
	Montevideo	1204	4	71
REC	Kansas City	2001	6	70
Enablers	Newburgh	2702	6	70
	Bellows Falls	2703	7	70
	Billings	2704	6	70
	Colorado Spring	2705	6	70

Table I-2

NON-PLANNED VARIATION AND CONTROL SITES

Site	Code	# Tested Classes	Year Joined
NPV HS			
Des Moines	0305	8	71
Tupelo	0711	4	69
W. Las Vegas	0714	4	70
Portageville	0804	11	69
Mounds	0808	2	70
Greeley	0906	4	70
Jonesboro	1002	3	69
Chattanooga	1007	4	70
Houston	1010	4	70
Paterson	1106	2	70
Johnston Co.	1108	4	69
Kansas City	2001	5	70
Control			
Huntsville	2801	---	71
Sacramento	2802	---	71
San Jose	2803	---	71

Data Collected

In this section we describe briefly all the instruments used in this study. A concise picture of the entire data collection effort is presented in Table I-3. Some information was collected on the full sample (F) of children in a tested classroom. Other information was collected on a partial sample (P). This partial sample consisted of a random sample of one third of the children in the class unless there were fewer than 18 children, in which case 6 were tested. Some tests were given in both the fall and spring, while others were given in the spring only. The test battery for Control children was somewhat different from that for Head Start children. All data collection was carried out by the Stanford Research Institute (SRI) with advice from the Huron Institute. A more complete description of data collection activities can be found in the SRI final report (1972).

In the analyses in this report, we focus exclusively on eight tests as outcome measures. These are the 32-item Preschool Inventory (PSI), the Peabody Picture

Table I-3

Data Collection Activities for the 1971-72 Year of the Head Start Planned Variation Study

F=Data for entire class.
P=Data for randomly selected 1/3 of class.

	FALL '71		SPRING '72	
Instrument	Head Start	Control	Head Start	Control
Preschool Inventory	F	F	F	F
Peabody Picture Vocabulary	F	F	F	F
Wide Range Achievement[3]	F	F	F	F
ITPA Verbal Expression	P		P	
ETS Enumeration	P		P	
8-Block Sort Task	P	F	P[1]	F
IDS Self-Concept	F	F	F[1]	F
Classroom Behavior Inventory	F		F	
Motor Inhibition	P		P	
Gumpgookies			F	
Relevant Redundant Cue			P	
Classroom Information Form	F		F	
Parent Information Form	P	F[2]	P	F[2]
Teacher Information Form	F		F	

[1]Data collected only at sites: 0204, 0316, 0512, 0711, 0804, 0902, 1002, 1010, 1106, 1203

[2]Different form for controls

[3] spelling, oral arithmetic, written arithmetic, word reading tests given in spring only.

Vocabulary Test (PPV), four subtests of the Wide Range Achievement Test (WRAT), the Verbal Expression subtest of the Illinois Test of Psycholinguistic Ability (ITPA), and the Educational Testing Service Enumeration Test (ETS).

We focus on these tests partly as a result of problems in the way the data were collected, but more importantly because we felt there were crippling limitations on the usefulness and appropriateness of the other measures as evaluative instruments. In the brief descriptions presented below, we give specific reasons for excluding each measure which is not used in the analyses. Not surprisingly, the tests which are suitable all measure skills in the cognitive domain. We would like to be able to study effects in the affective domain, but we have concluded with reluctance that the instruments used in this study need further refinement before they can be relied on. Since these instruments are experimental, we felt that it would be more valuable to look at the data from the viewpoint of what we can learn about the tests and how they might be useful in future evaluations, rather than what we can learn about program effects. These analyses are presented in detail by Walker et al. (1973). The test battery used during 1971-72 was completely different from that used in 1970-71. Thus we cannot replicate previous findings nor

make comparisons across cohorts. It is hoped, however, that the test battery represents an improvement and will be a more sensitive detector of program differences. In particular, several of the tests measure specific academic skills as opposed to general intellectual ability or achievement. Programs may differ more in their effects on such skills.

An important consideration in selecting tests was the fact that a study was planned to follow up many of the children in our study into their first school year in a Follow Through program.* Thus tests suitable for slightly older as well as pre-school children were sought, so that the developmental process over at least a two-year period could be studied.

We begin our test descriptions with the eight outcome measures selected for use. All correlations mentioned are taken from a table in Walker's report (1973) which we have for convenience reproduced here as Table I-4. These correlations are based on the fall test results for the entire Head Start sample. All other information quoted can be found in Walker's report, and we provide no further documentation.

---

*This study will be carried out by Abt Associates, beginning September 1973.

E I-4

INTERCORRELATIONS OF FALL 1971 SCORES FROM THE PPVT, WRAT SUBTESTS, 32-ITEM PSI, ITPA VERBAL EXPRESSION SUBTEST, ETS ENUMERATION SUBTESTS, BROWN, MI-TRUCK SUBTEST, AND EIGHT-BLOCK SORT SUCCESS SCORES[1]

	PPVT	WRAT-COPY MARKS	WRAT-RECOG. LETTERS	WRAT-NAME LETTERS	WRAT-READ #'s	WRAT-DOT COUNT.	PSI 32 - ITEM	ITPA-VERBAL EXPRESS.	ETS. ENUM. TOTAL	ETS. ENUM. COUNT.	ETS ENUM. TOUCH.	ETS ENUM. SAME # MATCH.	BROWN UNADJ.	BROWN ADJ.	MI - TRUCK	EIGHT-BLOCK PLACE.	EIGHT-BLOCK REASON
- MARKS	.413 (2881)																
- G. LETTERS	.537 (2881)	.575 (2995)															
- LETTERS	.546 (268[illegible]	.558 (299[illegible]	.502 (2995)														
- NUMBERS	.407 (2881)	.412 (2995)	.323 (2995)	.603 (2995)													
- COUNTING	.433 (2881)	.463 (2995)	.419 (2995)	.544 (2995)	.451 (2995)												
(32-item)	.665 (2855)	.551 (2860)	.481 (2860)	.414 (2860)	.508 (2860)	.589 (2860)											
- AL EXPRESSION	.487 (1147)	.339 (1172)	.371 (1172)	.276 (1172)	.341 (1172)	.388 (1172)	.506 (1138)										
ENUMERATION L	.475 (1075)	.508 (1097)	.427 (1097)	.307 (1097)	.446 (1097)	.542 (1097)	.584 (1073)	.459 (1115)									
ENUMERATION TING	.492 (1075)	.504 (1097)	.422 (1097)	.359 (1097)	.500 (1097)	.620 (1097)	.625 (1073)	.384 (1115)	.781 (1135)								
ENUMERATION NTING	.282 (1075)	.358 (1097)	.293 (1097)	.196 (1097)	.271 (1097)	.383 (1097)	.382 (1073)	.308 (1115)	.721 (1135)	.590 (1135)							
ENUMERATION # MATCHING	.237 (1075)	.225 (1097)	.199 (1097)	.095 (1097)	.176 (1097)	.148 (1097)	.232 (1073)	.298 (1115)	.664 (1135)	.257 (1135)	.202 (135)						
N- DJUSTED	.322 (2689)	.162 (2753)	.243 (2753)	.145 (2753)	.173 (2753)	.270 (2753)	.325 (2689)	.261 (1145)	.228 (1073)	.271 (1073)	.160 (1073)	.054 (1073)					
N- JUSTED	.239 (2689)	.127 (2753)	.166 (2753)	.100 (2753)	.124 (2753)	.194 (2753)	.259 (2689)	.215 (1145)	.159 (1073)	.172 (1073)	.134 (1075)	.034 (1073)	.637 (2879)				
RUCK	.174 (607)	.061 (625)	.048 (625)	.083 (625)	.121 (625)	.006 (625)	.164 (608)	.032 (637)	.136 (597)	.155 (597)	.047 (597)	.107 (597)	.118 (610)	.109 (610)			
T-BLOCK ENENT	.304 (1119)	.222 (1148)	.271 (1148)	.145 (1148)	.207 (1148)	.304 (1148)	.305 (1090)	.303 (1096)	.322 (1032)	.415 (1032)	.200 (1032)	.180 (1032)	.212 (1115)	.183 (1115)	.005 (575)		
T-BLOCK ON	.445 (1119)	.364 (1148)	.333 (1148)	.286 (1148)	.372 (1148)	.390 (1148)	.443 (1090)	.418 (1096)	.405 (1032)	.402 (1032)	.258 (1032)	.211 (1032)	.178 (1115)	.168 (1115)	.063 (575)	.520 (1211)	
T-BLOCK ESS TOTAL	.439 (1119)	.346 (1148)	.351 (1148)	.257 (1148)	.344 (1148)	.404 (1148)	.440 (1090)	.422 (1096)	.422 (1032)	.416 (1032)	.266 (1032)	.226 (1032)	.220 (1115)	.200 (1115)	.046 (575)	.839 (1211)	.901 (1211)

ple size for each correlation is included in parenthesis. Children in sample are those with adequate information
in Level I sites.

S ENUMERATION Scores sum of counting, touching and same number matching subtest scores.
scores are log transformations of slow times.

Caldwell Preschool Inventory (PSI)

The PSI is designed to assess general achievement in skills useful for later school success. In 1971-72 a 32-item version was used, consisting of a subset of the items in the 64-item version used in 197-71. Our best estimate for a reliability coefficient is .83. Correlations with all other cognitive tests in the battery are quite high, the highest being .665 with the PPV. For the Head Start population there appear to be no ceiling (test too easy) or floor (test too hard) effects, and the distribution of scores is quite symmetrical. With its generally excellent psychometric properties and the fact that it taps very general information processing skills and preschool achievement, the PSI is potentially the most useful test in our battery for program evaluation. Its very generality may, however, make it insensitive to program differences.

Peabody Picture Vocabulary Test (PPV)

The PPV contains a maximum of 150* test items designed to measure receptive vocabulary. For each item the stimulus word (noun or verb) is presented orally and the child is required to indicate which of 4 pictures corresponds to the

---

*Only 100 given in Fall.

word. Items increase in difficulty and the child continues until he makes 6 errors out of 8 consecutive items. His score is then the number of correctly answered items. Reliability of the PPV is in the .7 to .8 range. Since the test has effectively no upper limit for young children, ceiling effects are not a problem, nor are there floor effects. Correlations with other tests in the battery are generally high. The highest are .665 with the PSI and .537 with the WRAT Recognizing Letters. Although the PPV probably taps general intelligence and language ability, Walker recommends that the test be used only as a measure of passive vocabulary at this time.

Wide Range Achievement Test (WRAT)

Four of the WRAT subtests administered to the full sample in both the fall and spring were used in the analyses. The WRAT subtests measure specific academic skills, and it seemed reasonable to treat them as separate measures. The PSI provides a good measure of general achievement. By looking at the various WRAT subtests individually we can obtain a more detailed profile of cognitive program effects.

1. WRAT Copying Marks Subtest (WRTC)

In a one minute time interval the child copies as many of a series of 18 marks as he can. He is given credit for the number judged by the tester to be copied correctly. There are possible tester biases. Although our best estimate of internal reliability is about .8, a severe floor effect, particularly in the fall, renders this figure less impressive. Highest correlations are .551 with the PSI and .508 with the ETS. Although it is not clear exactly what useful skills are related to being able to copy abstract markings accurately and quickly, the WRTC probably measures motor coordination and a component of general school readiness.

2. WRAT Recognizing Letters (WRTR)

The child is required to recognize and match letters. The tester points to a series of letters in a row, and the child picks out the matching letter from a different series. There are 10 items. Our reliability estimate is around .8, but there are a substantial number of children scoring 0 and 10. Highest correlations are .537 with the PPV and .481 with the PSI. This test seems to measure the ability to recognize letters and also, possibly the ability to match shapes.

3. WRAT Naming Letter (WRTN)

The child is asked to name each of a series of 13 letters. Reliability is estimated at around .85. There is, however, a severe floor effect, particularly in the fall. Highest correlations are .600 with the WRAT Reading Number subtest and .414 with the PSI.

4. WRAT Reading Number (WRTD)

The child is asked to read aloud the numbers "3, 5, 6, 17, 41." Reliability is estimated at about .6, but there is a floor effect in the fall. Also there were almost no children in either fall or spring capable of identifying the number "41." Highest correlations are .600 with the WRTN and .508 with the PSI.

Since these four WRAT subtests measure fairly specific skills, have reasonable reliability, and were given in both fall and spring, we have decided to include them in the major analyses. The floor and ceiling effects will, however, raise problems in some of the analyses.

Illinois Test of Psycholinguistic Ability: Verbal Expression subtest (ITPA)

The ITPA measures a child's ability to express himself verbally. The ITPA is a diagnostic test, and its use for evaluative purposes is experimental. The child is

handed four familiar objects (ball, block, envelope, button) one at a time and asked by the tester to "tell me all about this." The score is the total number of distinct descriptors used by the child. Reliability is estimated to be between .6 and .8. Highest correlations are .506 with the PSI and .487 with the PPV. Although a child's ability to express himself would seem to be an important skill for later school success, Walker cautions that its usefulness in evaluation is questionable "because of the large variance, overall low mean response rate, and test administration problems."

Educational Testing Service Enumeration Test (ETS)

The ETS as used in this study consists of 3 subtests designed to measure components of the processes involved in learning the concept of number. The first subtest (Counting) requires the child to count dots (for one point) and say how many there are (for one point) for each of 3 items. The second subtest (Touching) has 6 items which require the child to touch each of the dots on a page one time only. The third subtest (Same Number Matching) consists of 8 items requiring the child to find the picture out of three with the same number of objects as the stimulus picture. A total score (maximum of 20) is found from the three

subtests. A fourth subtest (Same Order Matching) was originally included, but eliminated because it had low reliability and low correlations with the other subtests.

Reliability is estimated at about .75. Highest correlations are .584 with the PSI, and .508 with the WRTC. Although possibly subject to tester effects, the ETS has good psychometric properties and attempts to measure aspects of a developmental process which probably bears on future school success. It is one of the stronger tests in our battery.

As indicated in Table I-3, several other tests were also administered to all or part of the HSPV sample. We shall briefly describe these tests and give our reasons for not including them in our analyses of program effects.

The Eight-Block Sort Task (8-Block) examines maternal teaching style and mother-child interaction. The test consists of two parts, one of which has a floor effect and the other a ceiling effect. Reliability estimates are high, but correlations with other tests in the battery are low. The test was given in the spring at only 10 sites in 9 models. Since replication across sites is important in assessing model effects, we decided not to use the 8-block in our analyses.

The Brown IDS Self-Concept Referents Test (IDS) attempts to measure a child's self-concept. The distribution of scores is negatively skewed and displays ceiling effects. There is evidence that for Head Start age children the test measures cognitive (especially vocabulary) skills as well as self-concept, and that children try to select socially desirable responses rather than those applying to themselves. Walker concludes that "because of ...theoretical problems and the conflicting technical findings..., the Brown not be used in this form in future large-scale evaluations."

The Classroom Behavior Inventory (CBI) assesses child behavior in three areas: task orientation, extraversion, and hostility. For each of 15 items, a rater (usually the teacher) rates the child on a seven point scale. Test-retest reliabilities are adequate, but inter-rater reliabilities are moderate, and it is clear that different raters have different scales of reference, making cross classroom comparisons impossible. Since it appears impossible to obtain an absolute measure comparable across classrooms, there seems to be no way to use the CBI as an outcome measure.

The Motor Inhibition Test (MI) attempts to measure the ability to inhibit movement to conform to task demands. Only one of three parts (Tow Truck Task) was given in 1971-72.

Reliability does not appear adequate, and there are significant tester effects. Walker suggests that "the Truck subtest be dropped from future large-scale evaluations."

Gumpgookies (GG) is designed to measure achievement motivation. Reliability estimates range from .7 to .9. The short form of the GG used in HSPV is experimental, and, considering that it was administered only in the spring, we felt it would be misleading to use it for evaluative purposes.

The Relevant Redundant Cue Concept Acquisition Test, or "Zings and Poggles" (Z & P) tests a child's ability to master a particular abstract concept. Reliability for Head Start age children is very low, and it seems that there is much guessing. The test may be good for older children, but is just too difficult for children this young. Furthermore, it was given only in the spring.

Besides the four WRAT subtests discussed above, one other was given in both fall and spring. The WRAT Dot Counting (WRTU) subtest requires the child to count a series of 15 dots arranged in a row. The score is the highest number counted correctly. There are both floor and ceiling effects (a substantial number of children scoring 0 and 15). Moreover, the subtest consists of essentially one item.

Four other WRAT subtests were given in the spring only. These are Spelling, Oral Arithmetic, Written Arithmetic, and Word Reading. All except the Oral Arithmetic subtest were clearly too difficult for Head Start children. Since the estimated reliability of the Oral Arithmetic was only .55 and we had no fall scores, we decided not to use it.

We conclude our discussion of tests with mention of the Hertzig-Birch scoring system, which was applied to the PSI in 1971-72. This elaborate scoring system notes not only whether an item is answered correctly or incorrectly, but assesses the child's style of response to cognitive demands. The system is experimental and potentially quite informative, but more data is needed before its usefulness for evaluation can be determined.

Some of the data which we found not useful for our HSPV evaluation may prove useful in the study of Follow Through, and in particular in the study which will follow our HSPV sample into Follow Through. Accordingly, we have forwarded this data to Abt Associates. We would also encourage others to study the various experimental tests in our battery, so that more refined instruments will be available for future studies.

We conclude this section with brief descriptions of other sources of information used in our analyses. A complete listing of the specific items used can be found in Appendix A.

The Classroom Information Form (CIF) was our primary source of information on child background characteristics, such as age, sex, and ethnicity. These forms were filled out by teachers.

The Parent Information Form (PIF) was administered to parents to elicit information about home environment, parent and child attitudes, and the extent of parent involvement in Head Start and other activities. Since Control children were not in classrooms, their parents were given a modified version of the PIF which served also as the primary source of child background information.

The Teacher Information Form (TIF), filled out by teachers, requested information on teacher background, teaching experience, and attitudes towards the PV model (if any) with which they were working.

In addition, several items of information were provided by sponsors and local Head Start directors. Of these, we utilize only the ratings (on 0 to 9 scale) of classrooms in terms of the degree of implementation.

Other Reports

This report is one of a series being prepared under Grant # H 1926 from the Office of Child Development. In this section we discuss briefly relevant results from other reports in this series.

Smith (1973) analyzed the 1970-71 cohort data. His outcomes consisted of three measures of cognitive achievement, one measure of general intelligence, and one measure of motor control The achievement measures were a 64-item version of the Caldwell Preschool Inventory (PSI), and the NYU Booklets 3D and 4A. The PSI is a test of general achievement in areas deemed necessary for later school success. The NYU books tap more specific cognitive skills. The Stanford-Binet IQ was used as a measure of intelligence. The Motor Inhibition test was used to measure a child's ability to inhibit physical activity in order to perform a specified task. The interested reader is referred to Walker's report (1973) for more detail on these tests.

On the basis of his analyses using these measures, Smith concluded that:

1. The Head Start experience substantially improved performance on all 5 outcome measures.
2. There were no differences in effects between the PV programs (taken together) and the NPV programs on any of the measures.
3. No model stood out as being overall more or less effective than the others.

A more detailed breakdown of the inter-model comparison results appears in Table I-5, reproduced from Smith's report. Smith found two instances of outstanding model effectiveness. The High/Scope model was extraordinarily successful at boosting Stanford-Binet IQ scores. The Kansas model was highly effective in raising scores on the Book 4A. Only a few other effects are cited in this table, and on the whole, it appears that there is not much difference in overall effectiveness of the various models in terms of the measures used.

Featherstone (1973) has attempted to relate program effectiveness to child characteristics; that is, to detect model-by-child characteristic interactions. Using the 1969-70, and 1970-71 data and studying the PSI and Stanford-Binet only, she finds no consistent, interpretable interactions involving fixed child characteristics, such as sex, ethnicity, and socio-economic status. She suggests that characteristics such as age, prior preschool experience, and cognitive style, which describe the child at a particular point in his development, may relate to relative model effectiveness. She concludes that "the strategy which works best for a child today is not necessarily the one which will be optimum next month or next year."

Table I-5

(Reproduced from Smith, 1973)

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		+

Lukas and Wohlleb (1973) have considered the process of program implementation. They explore the basic questions:

1. How well are the models implemented?
2. What factors affect the process of implementation?

It was originally assumed that models were well-defined carefully specified educational packages which, given sufficient time, could be replicated completely at any chosen site. Lukas and Wohlleb have found that the implementation process is much more complex, involving a large number of people (sponsors, local administrators, parents, teachers, teacher-aides) with varying, and sometimes conflicting interests, goals, and philosophies. Moreover, the models themselves are not that well explicated by the sponsors. We cannot be sure exactly what treatment is being received by children in a given classroom simply by knowing the model name. Classrooms with the same model may vary considerably.

Summary and Look Ahead

In this chapter we have tried to give the reader a picture of the HSPV experiment. We began with the back-

ground of the study, showing how it evolved out of earlier Head Start programs and evaluations, and discussed its relation to the Follow Through program. We then set out the three major questions we hope to answer:

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged preschoolers acquire cognitive skills?
2. Are the Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?
3. Are some PV models particularily effective at imparting certain skills?

We then provided brief descriptions of the 11 PV models to be studied. It was noted that an important dimension in describing these models is the extent to which the acquisition of academic skills is stressed. We noted that at least three of the models are consciously concerned with the development of specific academic skills useful in the early school years. These are the Oregon, Kansas, and Pittsburgh models. We shall refer to these throughout this report as the "academic" models. As an important subquestion of the third of our major questions, we shall ask whether the academic models are overall especially effective.

Following the model descriptions, we set out the basic design for the study. We defined a site as a group of children in a particular location undergoing a particular kind of preschool experience. There are 28 PV sites, each with one of our 11 models, 12 ordinary non-sponsored (NPV) sites, and 3 Control sites, with no preschool program. We also decided for convenience in presenting results throughout this report to refer to the NPV children taken together and the Control children as program groups.

We then discussed the data collected. A brief description of each instrument was given. For each test we gave our reason for including or not including it in our analyses. We found that only eight tests were suitable for program evaluation. These are the Preschool Inventory (PSI), Peabody Picture Vocabulary Test (PPV), four subtests of the Wide Range Achievement Test (WRAT), the Illinois Test of Psycholinguistic Ability Verbal Expression Subtest (ITPA) and the Educational Testing Service Enumeration Test (ETS). Unfortunately, these tests all measure skills in the cognitive domain, frustrating the hope that information on socio-emotional development might also be obtained.

Finally, we presented brief summaries of relevant findings from other reports in this series. With all this

as background, we are now ready to look at the data.

The remainder of this report consists of 8 additional chapters. Chapter II contains a descriptive presentation of much of the data. We present background characteristics of children and teachers, and summaries of the distributions of fall and spring test scores for models and sites. Chapter III is devoted to general methodological considerations. We discuss limitations placed on the analysis by the design and the tests themselves, and our general approach to the problem of evaluating educational programs. Chapters IV through VII consist of four different analyses of the data, each with certain strengths and weaknesses. Through this variety of approaches, we hope to gain answers to the major questions mentioned above. In Chapter VIII we consider what evidence we have relating to the question of interactions between individual child characteristics and program effects. Finally, in Chapter IX we summarize the results of the various analyses and present our conclusions.

Chapter II

Descriptive Presentation of the Data

Introduction

In this chapter we present summaries of the data in order to give a general idea of the sample in terms of background characteristics and distributions of test scores. The reader is forewarned that the chapter is largely a rather dry compilation of facts, included primarily for the sake of completeness and future reference.

Note that there is no one analysis sample to which we can always refer. The samples suitable for the different analyses described in this report may differ slightly. Not all the same information has been collected on each child, and the minimal data requirements for the various analyses differ. We shall always make clear the criteria used in selecting an analysis sample. In general, the data collection was well done, and there is little missing data on key variables.

Child Background Characteristics

Tables II-1 through II-3 present some important background characteristics by model and site for a sample of 3,361 children. This represents all children whose sex is known, who are either Black, White, Mexican American or Puerto Rican, and who have a valid fall or spring test score on at least one test. A test score was considered valid if

Table II-1

BACKGROUND CHARACTERISTICS BY MODEL

Model	%F	%B	%W	%MA	%PR	%Eng Spk.	%PS Exp.	ME	Oct. 1 Age (mos.)	HH Size	Inc.	n
Far West	47.1	13.6	74.3	12.1	0	96.9	14.7	11.1	56.0	5.0	35.9	257
U. of Arizona	45.9	29.9	61.6	.7	7.8	92.5	27.5	10.4	58.5	5.4	41.2	281
Bank Street	55.4	83.3	16.7	0	0	99.7	40.9	10.5	56.0	5.4	36.4	305
U. of Oregon	45.3	43.6	10.1	46.4	0	58.1	18.6	9.9	64.7	5.5	30.4	179
U. of Kansas	46.7	54.1	45.9	0	0	100.0	7.6	10.3	54.1	5.9	43.2	135
High Scope	46.8	27.7	55.0	17.3	0	91.3	18.3	10.4	58.0	5.4	33.9	231
U. of Florida	50.9	58.1	23.9	17.9	0	91.5	16.0	9.4	60.8	5.3	30.0	234
EDC	57.9	74.2	21.6	0	4.2	95.8	45.3	9.6	60.7	5.6	38.0	190
U. of Pittsburgh	43.9	0	100.0	0	0	100.0	22.8	11.1	53.4	5.6	44.1	139
REC	54.1	26.8	20.4	51.0	0	25.7	15.5	10.5	54.2	5.7	33.9	98
Enablers	49.2	30.5	47.3	20.0	2.2	92.4	16.2	10.3	57.8	5.1	39.0	315
NPV	47.3	53.1	31.6	15.2	.1	91.8	25.6	10.0	59.1	5.7	34.4	858
Control 28	52.5	43.9	40.3	15.8	0	91.9	23.7	10.5	49.6	5.4	37.4	139
Total	49.0	44.7	40.3	13.8	1.1	91.9	23.8	10.2	57.9	5.5	36.2	3361

Table II-2

BACKGROUND CHARACTERISTICS BY SITE (PV)

Site	%F	%W	%B	%MA	%PR	%Ind	%Eng	%PS Exp.	ME	Oct. 1 Age	HH Size	Inc.	n
0204	53.7	91.5	7.3	1.2	0	0	100.0	35.9	11.1	55.7	4.7	41.5	82
0209	41.6	59.6	9.0	31.5	0	0	91.0	4.5	10.9	56.3	5.3	31.6	89
0213	46.5	73.3	24.4	2.3	0	0	100.0	5.8	11.2	56.1	4.9	35.3	86
0308	44.7	69.1	30.9	00.0	0	0	100.0	66.0	9.3	65.9	5.4	35.9	94
0309	48.3	20.7	54.0	00.0	25.3	0	77.0	10.5	10.5	55.4	5.5	57.4	87
0316	45.0	90.0	8.0	00.0	2.0	0	99.0	6.0	11.1	54.5	5.4	37.8	100
0510	57.0	11.4	88.6	00.0	0	0	100.0	57.1	10.1	66.2	5.8	30.6	114
0511	55.9	00.0	100.0	00.0	0	0	100.0	24.5	10.6	51.8	5.3	33.1	102
0512	52.8	42.7	57.3	00.0	0	0	98.9	39.3	11.0	47.7	5.0	47.7	89
0711	42.2	13.3	86.7	00.0	0	0	100.0	30.0	9.4	64.9	5.7	26.4	90
0714	48.3	6.7	00.0	93.3	0	0	15.7	6.9	10.5	64.4	5.4	34.5	89
0804	39.1	64.1	35.9	00.0	0	0	100.0	00.0	9.9	56.0	5.9	39.1	64
0808	53.5	29.6	70.4	00.0	0	0	100.0	14.9	10.6	52.4	5.9	46.9	71
0902	51.8	24.7	75.3	00.0	0	0	100.0	14.1	10.4	53.3	5.3	31.7	85
0904	45.4	100.0	00.0	00.0	0	0	100.0	12.4	10.5	63.0	5.4	34.5	97
0906	40.8	18.4	00.0	81.6	0	0	59.2	37.5	10.2	56.6	5.6	36.7	49
1002	56.4	67.3	32.7	00.0	0	0	100.0	3.8	8.6	67.5	5.4	32.1	55
1007	57.4	20.2	79.8	00.0	0	0	100.0	37.2	10.2	61.6	5.4	29.5	94
1010	40.0	00.0	50.6	49.4	0	0	76.5	00.0	9.0	55.6	5.1	29.2	85
1106	61.4	1.1	89.8	00.0	9.1	0	90.9	00.0	10.4	53.0	5.2	47.8	88
1108	54.9	39.2	60.8	00.0	0	0	100.0	84.3	8.9	67.4	6.0	33.1	102
1203	40.7	100.0	00.0	00.0	0	0	100.0	17.8	11.0	52.0	5.7	41.9	91
1204	50.0	100.0	00.0	00.0	0	0	100.0	32.6	11.6	55.9	5.3	47.9	48
2001	54.1	20.4	28.6	51.0	0	0	85.7	15.5	10.5	54.2	5.7	33.9	98
2702	51.2	8.5	80.5	2.4	8.5	0	93.9	9.9	10.8	53.4	4.9	47.4	82
2703	45.1	100.0	00.0	00.0	0	0	100.0	21.4	10.5	55.5	5.1	43.2	71
2704	51.3	77.6	2.6	19.7	0	0	100.0	32.9	10.4	67.7	4.8	29.2	76
2705	48.8	14.0	32.6	53.5	0	0	77.9	3.6	9.6	55.2	5.6	38.2	86

Table II-3

BACKGROUND CHARACTERISTICS BY SITE (NPV AND CONTROL)

	%F	%B	%W	%MA	%PR	%Ind	%Eng	%PS Exp.	ME	Age	HH Size	Inc.	n
0305	44.0	50.9	39.7	9.5	0	0	100.0	19.6	10.7	56.6	5.2	34.8	51
0711	39.7	22.2	77.8	00.0	0	0	100.0	63.5	8.9	64.6	6.1	34.6	25
0714	50.7	1.3	00.0	98.7	0	0	30.7	12.3	10.7	64.4	5.6	34.3	38
0804	49.4	43.3	56.7	00.0	0	0	100.0	8.2	9.3	54.9	5.8	31.5	89
0808	60.9	60.9	39.1	00.0	0	0	100.0	52.2	11.0	53.8	5.9	50.2	14
0906	48.9	44.4	0.00	55.6	0	0	75.6	8.9	9.5	56.4	5.9	48.5	45
1002	48.8	85.4	14.6	00.0	0	0	100.0	7.3	8.5	68.3	6.5	33.8	41
1007	36.7	5.0	95.0	00.0	0	0	100.0	29.3	10.2	66.6	5.3	29.2	60
1010	42.9	00.0	69.6	30.4	0	0	90.9	72.0	10.4	56.3	5.4	35.6	56
1106	58.8	00.0	97.1	00.0	2.9	0	97.1	00.0	10.7	52.5	4.8	37.2	20
1108	57.3	18.7	81.3	00.0	0	0	100.0	56.0	9.4	67.0	6.4	32.0	43
2001	42.2	36.7	60.0	3.3	0	0	98.9	9.0	11.0	53.5	5.6	32.4	38
2801	56.9	43.1	56.9	00.0	0	0	98.2	17.2	8.9	50.9	6.2	26.2	33
2802	47.5	39.0	42.4	18.6	0	0	91.5	25.4	11.6	47.4	4.7	49.3	28
2803	54.5	36.4	13.6	50.0	0	0	77.3	36.4	11.6	52.8	4.9	40.3	12

the tester indicated that the test was completed. Although reasons for failure to complete a test were noted by the tester, no incomplete test results were used in our analyses, regardless of the reason.

Looking first at Table II-1, we see that there is an approximately even split of boys and girls in all models. The ethnic compositions of the sites vary greatly. The Pittsburgh model, for example, contains only Whites, while the Bank Street model is 83.3% Black. The Mexican Americans are distributed throughout seven of the models, with Oregon and REC containing by far the largest number. While most children come from homes where English is spoken, Oregon has 41.9% where this is not the case. The percentage of children with some prior preschool experience varies considerably. EDC (45.3) and Bank Street (40.9) are high, and Kansas (7.6) is low. The average number of years of mother's education varies from 9.4 for Florida to 11.1 in both Far West and Pittsburgh. The mean age (on October 1, 1971) varies from 53.4 months for Pittsburgh to 64.7 for Oregon. The Control children are a bit younger than the Head Start children, averaging 49.6 months. Household size varies little, from an average of 5.0 in Far West to 5.9 in Kansas. Mean family income also shows little variation, ranging from $3000 per year in Florida to $4,410 in Pittsburgh.

Tables II-2 and II-3 give the same background information broken out by sites. For all variables, with the

possible exception of sex and household size, there is considerable site-to-site variation within models. There is only one site with a mean age between 57 and 63 months. Essentially there are two distinct types of sites, those in which children were to enter kindergarten following Head Start, and those in which they were to enter first grade directly. Smith (1973) suggests that children in "entering-first" sites may undergo systematically different experiences from those in "entering-kindergarten" sites. This is an interesting hypothesis. Unfortunately entering level is severely confounded in our design with age, region, and model, all of which would have to be controlled in order to tease out the entering-level effect. Thus we have not taken any explicit account of entering level in our analyses. We do, of course, recognize age as a potentially important influence on measured outcomes.

Teacher Background Characteristics

In Table II-4 we present background information by model for all teachers. Nearly all the teachers are women and are either Black or White. The percentage of Black teachers ranges from 0 in Pittsburgh and REC to 66.7 in EDC. The percentage of teachers living in a neighborhood similar to that of the children they teach ranges from 0 in REC to 88 in High/Scope. The percentage of teachers certified ranges from 0 in Oregon and Kansas to 50 in REC. Mean teacher age ranges from 26.8 years in Far West and

Table II-4

Teacher Background Characteristics*

Program	% Female	% Black	% Same Neighborhood	% Certified	Mean Age	Mean Yrs. Ed	n
Far West	100	5.6	16.7	33.3	26.8	16.0	18
Arizona	100	37.5	50.0	37.5	36.5	15.3	16
Bank St.	100	60.7	46.7	36.7	38.0	15.2	30
Oregon	100	22.2	66.7	0	37.4	13.6	9
Kansas	100	25.0	81.3	0	36.7	13.7	16
High/Scope	100	16.0	88.0	28.0	40.1	14.4	25
Florida	100	50.0	22.2	44.4	34.8	16.2	18
EDC	100	66.7	77.8	33.3	34.9	14.9	9
Pitts-burgh	100	0	44.4	33.3	34.9	16.0	9
REC	50	0	0	50.0	26.8	16.5	4
Enablers	85.7	19.0	12.0	28.6	41.5	14.8	21
NPV	96.2	25.0	38.5	45.3	37.7	15.2	52
Total	96.9	29.3	55.5	33.3	36.7	15.1	227

*This table is based on all teachers, including some HSPV classes in which no testing was done.

REC to 41.5 in the Enablers. Mean years of education varies from 13.6 in Oregon to 16.5 in REC.

Outcome Measures

Tables II-5 through II-28 present summary statistics for the distributions of fall and spring test scores for models and sites. For each model or site and each test, we present the mean, median, lower quartile, upper quartile, standard deviation, and sample size. The sample used for each test consists of all children with valid fall and spring scores on that test.

It is evident that there are substantial differences among models and among sites within models on background characteristics and fall test scores. Our ability to make fair comparisons among programs will depend on our ability to take account of and adjust for these pre-program differences. Since experimental equalization has apparently failed, we must rely on statistical techniques. We shall discuss how and to what extent this is feasible in the following chapter.

Many researchers feel comfortable in describing gains or effects in terms of standard deviations. As a rough rule of thumb, one-half a standard deviation is sometimes taken as a criterion for educational significance. For convenient reference, we list in Table II-29 the standard deviations based on the entire fall sample of children in Head Start programs.

Table II-5

SUMMARY STATISTICS FOR PSI DISTRIBUTION BY MODEL

Model	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
Far West	14.5	14	10	18	5.4	20.8	21	18	24	4.9	6.3	177
U. of Arizona	15.7	15	12	20	5.9	20.3	21	17	24	5.4	4.6	204
Bank Street	14.3	13	8	20	7.4	17.1	17	12	21	6.1	2.7	239
U. of Oregon	17.1	18	13	21	6.1	23.3	24	20	27	4.6	6.1	157
U. of Kansas	14.1	14	11	18	5.3	18.8	19	14	24	6.4	4.6	101
High Scope	15.8	15	10	21	6.8	19.7	20	15	24	6.1	3.9	181
U. of Florida	14.0	13	10	18	5.3	18.5	19	14	23	5.5	4.5	153
EDC	14.9	14	10	19	5.6	19.5	20	15	24	5.6	4.6	162
U. of Pittsburgh	13.6	13	9	18	6.1	19.8	20	16	23	5.5	6.1	116
REC	12.6	11	9	16	4.3	17.7	18	14	21	4.8	5.1	72
Enablers	14.9	14	10	19	6.3	19.7	20	15	25	6.1	4.8	221
NPV	14.6	14	10	19	6.2	18.9	19	15	24	5.8	4.3	669
Control 28	12.2	11	8	15	5.7	15.2	14	11	20	6.4	3.0	111

Table II-6

SUMMARY STATISTICS FOR PSI DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	15.9	15	12	19	5.1	22.6	23	19	26	4.0	6.7	53
0209	14.6	14	9	20	5.7	20.1	20	17	23	5.1	5.5	58
0213	13.4	13	10	16	5.3	20.0	21	16	23	5.0	6.6	66
0308	18.8	20	15	23	5.6	22.8	23	20	26	4.2	4.0	73
0309	12.4	12	9	15	5.1	17.2	17	14	21	5.4	4.8	58
0316	15.3	15	12	18	5.2	20.3	21	17	24	5.2	5.0	73
0510	20.9	21	16	26	5.6	21.5	21	16	26	5.0	.6	101
0511	9.5	8	6	12	3.9	13.3	12	10	16	4.3	3.6	71
0512	9.7	9	6	13	4.6	14.7	15	12	18	5.0	5.0	67
0711	17.5	18	14	22	5.8	22.8	23	20	27	5.2	5.3	80
0714	16.7	17	12	21	6.3	23.8	24	20	26	3.9	7.1	77
0804	14.3	14	10	18	5.8	21.7	23	17	26	5.6	7.4	52
0808	14.0	14	11	16	4.7	15.8	17	12	20	5.9	1.8	49
0902	11.5	10	7	15	5.3	15.0	15	11	19	4.8	3.5	62
0904	20.5	21	16	24	5.8	24.1	25	21	27	4.6	3.6	79
0906	13.1	12	10	15	4.5	18.1	18	15	21	4.3	5.0	40
1002	15.7	15	12	20	5.7	19.8	22	16	24	5.9	4.1	42
1007	15.6	15	12	19	4.7	19.0	19	15	23	5.4	3.4	61
1010	10.5	10	8	13	4.1	16.8	17	12	20	4.8	6.3	50
1106	11.4	11	9	14	4.2	15.4	15	13	18	4.2	4.0	65
1108	17.2	17	13	21	5.3	22.3	22	20	26	4.6	5.1	97
1203	11.3	10	8	15	4.8	17.8	18	15	21	5.2	6.5	75
1204	18.0	18	13	22	5.9	23.4	23	20	27	4.2	5.4	41
2001	12.6	11	9	16	4.3	17.7	18	14	21	4.9	5.1	72
2702	12.1	11	8	15	5.0	16.9	17	14	21	5.0	4.8	55
2703	12.2	12	9	15	4.9	17.7	19	13	21	5.0	5.5	35
2704	20.3	20	16	25	5.0	25.2	26	24	28	3.8	4.9	74
2705	12.4	11	9	16	5.3	16.4	16	14	20	4.9	4.0	57

Table II-7

SUMMARY STATISTICS FOR PSI DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD	Mean Gain	N
0305	15.5	16	11	19	5.3	19.2	20	15	23	5.0	3.7	75
0713	17.7	18	14	21	4.8	21.6	22	19	25	4.9	3.9	58
0714	20.0	20	19	23	5.6	22.4	24	18	26	5.2	2.4	63
0804	11.1	11	8	14	5.3	15.7	16	12	20	5.9	4.6	143
0808	14.4	14	8	18	5.9	18.9	20	13	23	6.4	4.5	17
0906	14.6	15	9	19	6.0	19.6	20	15	23	5.5	5.0	35
1002	17.1	17	12	21	5.7	21.9	23	17	26	5.7	4.8	34
1007	18.8	18	15	23	5.6	22.3	23	19	26	4.7	3.5	46
1010	12.2	10	8	18	6.2	16.4	17	13	20	5.8	4.2	41
1106	11.2	11	8	14	4.1	14.1	13	11	17	4.1	2.9	32
1108	16.6	17	12	21	5.6	21.2	21	18	24	4.8	4.6	65
2001	10.4	10	7	13	4.1	17.4	17	14	20	4.3	7.0	60
2801	11.2	11	7	13	5.1	14.0	13	9	18	6.0	2.8	37
2802	12.5	11	8	16	6.4	15.3	14	10	20	7.0	2.8	53
2803	13.4	14	9	16	5.0	17.3	19	13	22	5.3	3.9	21

Table II-8

SUMMARY STATISTICS FOR PPV DISTRIBUTION BY MODEL

Model	FALL Mean	Med.	LQ	UQ	SD	SPRING Mean	Med.	LQ	UQ	SD	Mean Gain	N
Far West	37.3	39	30	46	12.1	48.2	49	43	54	9.0	10.9	172
U. of Arizona	35.2	36	28	43	11.6	45.7	48	49	53	11.0	10.4	195
Bank Street	27.8	27	18	37	12.3	37.6	38	30	47	11.7	9.7	243
U. of Oregon	35.5	37	25	45	13.8	46.1	45	40	53	10.0	10.6	137
U. of Kansas	31.1	33	20	39	12.9	41.3	40	36	49	11.0	10.2	98
High Scope	35.2	34	24	38	14.0	44.2	46	36	53	13.1	8.9	170
U. of Florida	30.2	31	19	39	13.1	41.1	41	34	50	11.2	10.9	143
EDC	29.6	29	21	36	11.7	39.8	40	33	48	11.4	10.2	160
U. of Pittsburgh	32.1	32	21	42	13.1	45.9	47	39	53	9.6	13.8	113
REC	29.2	31	17	38	11.9	42.9	43	35	52	10.2	13.7	71
Enablers	35.7	36	24	47	14.2	44.1	46	35	54	12.5	8.4	214
NPV	29.4	29	19	40	13.0	41.1	42	34	51	11.7	11.7	629
Control 28	28.5	25	17	40	13.7	39.2	41	30	49	11.9	10.6	113

Table II-9

SUMMARY STATISTICS FOR PPV DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	39.7	41	35	46	10.2	51.5	52	48	57	6.5	11.8	50
0209	39.3	40	35	48	11.2	46.1	48	40	51	9.2	6.8	56
0213	33.9	33	24	45	13.5	47.4	49	42	54	9.9	13.5	66
0308	36.5	36	30	44	12.0	46.1	48	39	52	11.4	9.6	66
0309	31.1	32	20	38	11.6	40.9	40	34	50	11.0	9.8	56
0316	37.3	39	33	45	10.6	48.9	50	43	55	9.5	11.6	73
0510	33.8	34	24	44	12.5	42.2	42	36	51	10.7	8.4	94
0511	22.6	20	15	30	9.5	34.2	35	28	41	10.5	11.6	77
0512	25.7	24	15	34	11.6	35.0	36	27	43	12.2	9.3	72
0711	31.7	32	21	41	12.6	33.2	43	37	48	8.6	11.5	63
0714	38.7	39	29	49	14.2	48.6	51	41	54	10.5	9.9	74
0804	34.0	35	20	41	12.4	45.9	47	39	54	8.0	11.9	49
0808	28.2	32	18	38	12.9	36.8	38	31	45	11.9	8.6	49
0902	24.5	24	17	31	9.8	33.2	34	22	40	10.8	8.7	59
0904	45.8	46	38	55	10.6	53.0	53	50	57	8.7	8.2	74
0906	31.2	31	24	36	10.7	44.2	45	40	50	10.7	13.0	37
1002	41.6	39	34	51	10.3	50.6	52	46	56	7.9	9.0	39
1007	30.3	28	23	40	11.0	39.6	40	34	48	10.6	9.6	50
1010	21.9	19	13	30	10.3	35.5	36	29	42	9.6	13.6	54
1106	22.3	20	16	28	8.9	32.2	33	24	40	10.8	9.9	66
1108	34.7	33	26	42	10.8	45.1	45	39	52	8.7	10.4	94
1203	27.4	27	18	37	12.4	42.9	43	37	50	9.5	15.5	73
1204	40.7	41	33	47	9.8	51.4	52	46	55	7.1	10.7	40
2001	29.2	31	17	38	12.0	42.9	43	35	52	10.3	13.7	71
2702	29.3	27	20	40	12.2	34.3	35	26	41	11.1	5.0	58
2703	31.3	32	19	38	12.5	42.2	42	35	49	10.6	10.9	36
2704	48.0	50	40	55	9.5	54.8	55	51	59	6.7	6.8	71
2705	28.6	28	19	36	11.9	41.7	42	35	51	10.2	13.1	49

Table II-10

SUMMARY STATISTICS FOR PPV DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	32.8	33	22	40	12.4	44.8	46	39	52	10.1	12.0	71
0711	33.3	32	26	41	9.5	44.4	46	37	51	9.1	11.1	47
0714	21.6	17	13	27	13.1	44.8	49	37	53	12.0	23.2	65
0804	26.4	24	17	36	12.4	35.5	36	27	45	12.2	9.1	145
0808	31.1	28	22	40	12.9	39.6	37	27	50	13.5	8.5	16
0906	32.9	38	21	46	15.5	46.6	50	38	55	11.7	13.7	38
1002	39.6	40	33	49	11.3	47.8	51	42	54	10.4	8.2	36
1007	34.9	32	26	42	11.9	47.6	49	44	53	9.4	12.7	28
1010	25.7	26	18	32	9.3	33.8	34	28	39	8.8	8.1	39
1106	26.0	24	20	32	8.6	30.6	33	23	36	8.9	4.6	27
1108	36.0	36	26	43	11.0	44.6	44	38	52	9.1	8.6	63
2001	23.7	21	15	33	12.2	40.0	39	34	45	8.7	16.3	54
2801	22.9	22	14	26	10.9	35.4	36	28	45	10.9	12.5	41
2802	30.1	29	17	40	14.8	40.0	41	31	50	12.7	9.9	52
2803	36.4	40	30	43	11.8	45.1	47	41	51	9.6	8.7	20

Table II-11

SUMMARY STATISTICS FOR WRTC DISTRIBUTION BY MODEL

Model	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
Far West	1.7	1	0	3	2.6	5.2	5	3	7	3.6	3.5	182
U. of Arizona	1.8	1	0	3	2.2	5.3	5	3	7	3.2	3.5	216
Bank Street	2.0	0	0	3	3.2	4.3	3	0	7	4.4	2.3	259
U. of Oregon	3.9	3	1	5	3.2	8.2	8	5	11	4.2	4.4	154
U. of Kansas	1.3	0	0	1	2.5	6.6	6	4	10	4.2	5.3	106
High Scope	2.1	1	0	3	2.8	5.4	4	1	9	4.9	3.3	191
U. of Florida	2.4	2	0	4	2.6	5.0	4	2	8	3.9	2.6	114
EDC	2.4	1	0	4	3.0	5.9	6	4	8	3.2	3.5	171
U. of Pittsburgh	1.3	0	0	2	2.4	4.5	4	1	7	3.9	3.2	117
REC	0.8	0	0	1	1.6	3.5	3	0	6	3.1	2.8	82
Enablers	2.1	1	0	3	3.0	5.2	4	2	8	4.2	3.1	236
NPV	1.8	1	0	3	2.3	4.8	4	1	8	4.4	3.0	669
Control 28	1.4	0	0	2	2.2	2.7	2	0	4	2.9	1.3	88

Table II-12

SUMMARY STATISTICS FOR WRTC DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	1.8	1	0	3	2.5	6.1	6	3	8	3.8	4.3	59
0209	2.2	1	0	3	3.1	5.4	4	3	8	3.6	3.2	56
0213	1.2	0	0	2	2.0	4.2	4	2	6	3.1	3.0	67
0308	2.3	2	0	4	2.9	6.7	7	4	9	3.4	3.9	76
0309	1.2	1	0	2	1.4	5.0	5	3	7	2.9	3.8	63
0316	1.3	0	0	2	1.8	4.2	4	2	6	2.8	2.9	77
0510	4.2	3	1	6	3.9	7.7	8	4	10	4.4	3.5	104
0511	0.2	0	0	0	0.6	1.8	1	0	3	2.1	1.6	80
0512	0.7	0	0	1	1.5	2.2	1	0	3	2.9	1.5	75
0711	2.8	2	1	4	2.5	6.5	6	3	9	4.2	3.7	77
0714	5.0	4	3	6	3.5	10.0	10	7	12	3.4	5.0	77
0804	1.0	0	0	1	1.8	6.8	6	4	9	3.7	5.8	52
0808	1.6	0	0	2	3.0	6.3	6	2	10	4.6	4.7	54
0902	0.6	0	0	1	1.2	2.1	1	0	3	2.7	1.5	68
0904	3.4	3	1	4	3.3	8.9	9	5	13	4.9	5.5	81
0906	1.8	1	1	2	2.1	3.8	3	1	6	3.1	2.0	42
1002	2.6	1	0	4	3.0	5.5	5	2	8	4.2	2.9	42
1007	2.9	2	1	3	3.1	6.8	7	3	8	3.9	3.9	11
1010	2.1	2	0	3	2.2	4.3	4	1	7	3.5	2.2	61
1106	1.1	1	0	1	1.7	4.9	5	3	7	3.1	3.8	73
1108	3.4	3	1	5	3.4	6.6	6	4	9	3.1	3.2	98
1203	0.7	0	0	1	1.3	3.3	3	0	5	3.1	2.6	76
1204	2.4	1	0	4	3.3	6.5	6	3	9	4.3	4.1	41
2001	0.8	0	0	1	1.3	3.3	3	0	5	3.1	2.6	76
2702	1.2	0	0	2	1.8	4.1	4	3	6	2.6	2.9	62
2703	0.7	0	0	0	1.6	2.7	2	0	4	2.5	2.0	40
2704	4.8	4	2	6	3.5	8.8	9	5	12	3.9	4.0	75
2705	0.8	0	0	1	1.6	3.6	2	0	4	4.0	3.8	59

Table II-13

SUMMARY STATISTICS FOR WRTC DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	1.9	1	0	3	2.2	4.9	4	1	8	4.4	3.0	76
0711	2.3	2	0	4	2.1	5.8	6	3	9	3.8	3.5	57
0714	3.5	3	1	5	2.8	11.0	10	9	14	3.3	7.5	65
0804	1.1	0	0	1	1.7	2.9	2	0	4	3.4	1.8	149
0808	1.4	1	0	2	1.5	6.1	4	4	8	3.9	4.7	18
0906	1.2	0	0	2	1.7	3.6	3	1	6	3.1	2.4	39
1002	1.9	1	0	3	1.8	7.3	8	3	10	4.3	5.4	36
1007	2.4	2	0	4	2.4	6.4	5	3	9	4.5	4.0	24
1010	1.9	1	0	3	2.3	2.6	1	0	5	3.4	.7	44
1106	0.7	0	0	1	1.1	2.7	2	1	4	2.0	2.0	33
1108	3.3	3	0	4	3.5	6.3	6	3	9	3.9	3.0	64
2001	0.6	0	0	0	1.5	1.6	1	0	2	2.4	1.0	64
2801	1.5	1	0	1	2.4	2.4	1	0	4	2.6	.9	26
2802	1.1	0	0	1	2.1	2.0	1	0	3	2.8	1.0	42
2803	2.1	2	0	2	2.3	4.6	4	3	6	2.9	.2	20

Table II-14

SUMMARY STATISTICS FOR WRTR DISTRIBUTION BY MODEL

Model	FALL Mean	Med.	LQ	UQ	SD	SPRING Mean	Med.	LQ	UQ	SD	Mean Gain	N
Far West	7.1	8	5	10	2.8	8.8	10	8	10	1.9	1.7	182
U. of Arizona	6.2	7	4	9	3.4	8.9	10	9	10	2.0	2.7	216
Bank Street	6.4	7	4	9	3.4	8.1	9	7	10	2.6	1.8	259
U. of Oregon	8.1	9	7	10	2.6	9.3	10	9	10	1.5	1.2	154
U. of Kansas	6.2	7	4	9	3.0	9.2	10	9	10	1.7	3.1	106
High Scope	6.8	8	5	9	3.1	8.4	9	7	10	2.1	1.6	191
U. of Florida	6.5	7	5	9	3.2	8.6	9	8	10	2.1	2.0	114
EDC	7.2	8	6	10	2.9	9.2	10	9	10	1.3	2.1	171
U. of Pittsburgh	6.2	7	4	9	3.1	9.3	10	9	10	1.6	3.1	117
REC	6.8	7	5	9	2.9	8.9	9	8	10	1.4	2.1	82
Enablers	7.4	8	6	10	2.8	8.5	9	8	10	2.2	1.1	236
NPV	6.0	7	3	9	3.4	8.1	9	8	10	2.7	2.1	669
Control 28	5.0	5	1	8	3.5	6.9	7	5	9	2.9	1.8	88

Table II-15

SUMMARY STATISTICS FOR WRTR DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	6.8	8	5	10	3.1	9.1	10	9	10	2.2	2.3	59
0209	7.5	8	6	9	2.3	8.4	9	8	10	2.1	.9	56
0213	6.9	8	5	10	3.0	8.8	9	8	10	1.4	1.9	67
0308	7.4	8	6	10	2.5	9.6	10	10	10	1.2	2.2	76
0309	4.2	3	0	8	3.8	8.2	9	8	10	2.8	4.0	63
0316	6.7	8	4	9	3.1	8.1	9	8	10	1.7	1.4	77
0510	7.9	9	7	10	2.8	9.1	10	9	10	2.0	1.2	104
0511	5.1	6	1	8	3.5	7.6	9	6	10	2.7	2.5	80
0512	5.6	6	3	8	3.3	7.4	8	6	9	2.8	1.8	75
0711	7.1	8	5	10	3.1	8.8	10	8	10	1.9	1.7	77
0714	9.0	10	9	10	1.6	9.7	10	9	10	0.6	.7	77
0804	6.8	8	5	9	2.7	9.7	10	10	10	0.8	2.9	52
0808	5.6	6	3	8	3.2	8.8	9	9	10	2.2	3.2	54
0902	5.0	5	2	8	3.3	7.0	7	6	9	2.6	2.0	68
0904	8.3	9	7	10	2.1	9.5	10	9	10	0.9	1.2	81
0906	6.7	7	5	9	2.9	8.4	9	7	10	1.8	1.7	42
1002	7.5	8	6	9	2.5	9.4	10	9	10	1.1	1.9	42
1007	8.5	10	8	10	2.9	8.5	9	8	10	2.2	0	11
1010	5.5	6	2	8	3.3	8.0	9	7	10	2.5	2.5	61
1106	6.5	7	5	9	2.9	9.0	10	8	10	1.4	2.5	73
1108	7.7	9	7	10	2.8	9.4	10	9	10	1.2	1.7	98
1203	5.6	6	4	8	3.1	9.0	10	9	10	1.9	3.4	76
1204	7.3	8	6	9	3.0	9.9	10	10	10	0.4	2.2	41
2001	6.8	7	5	9	2.9	8.9	9	8	10	1.4	2.1	82
2702	7.2	8	5	10	3.0	8.1	9	8	10	2.8	.9	62
2703	5.9	6	4	8	2.5	7.5	8	6	10	2.7	1.6	40
2704	8.9	10	8	10	1.9	9.6	10	9	10	0.8	.7	75
2705	6.7	7	5	9	2.8	8.2	9	7	10	1.9	1.5	59

Table II-16

SUMMARY STATISTICS FOR WRTR DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	6.8	7	5	9	2.9	8.2	9	7	10	2.5	1.4	76
0711	7.4	8	5	9	2.2	8.4	9	8	10	2.3	1.0	57
0714	2.5	1	0	3	3.5	8.9	9	9	10	1.8	6.4	65
0804	5.4	6	2	8	3.3	6.9	8	5	9	3.2	1.5	149
0808	6.5	7	4	8	2.7	8.8	9	8	10	1.0	2.3	18
0906	5.9	7	4	8	2.7	8.6	9	8	10	1.7	2.7	39
1002	7.5	8	5	10	2.9	9.1	10	9	10	2.0	1.6	36
1007	5.4	5	1	10	4.2	9.3	10	9	10	1.1	3.9	24
1010	6.3	8	2	9	3.6	7.2	8	5	10	3.3	.9	44
1106	7.3	8	6	9	2.5	8.6	10	8	10	2.1	1.3	33
1108	8.1	9	7	10	2.5	9.5	10	9	10	1.2	1.4	64
2001	4.6	4	1	8	3.3	6.7	8	3	9	3.6	2.1	64
2801	3.8	3	1	7	3.4	5.9	6	4	9	3.1	2.1	26
2802	5.2	5	1	8	3.6	6.6	7	5	9	3.2	1.4	42
2803	6.4	6	4	9	3.3	8.5	9	7	10	1.4	2.1	20

Table II-17

SUMMARY STATISTICS FOR WRTN DISTRIBUTION BY MODEL

	FALL					SPRING						
Model	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD	Mean Gain	N
Far West	1.7	0	0	1	3.2	3.8	2	0	7	4.5	2.1	182
U. of Arizona	1.3	0	0	1	2.6	5.3	4	1	10	4.7	4.0	216
Bank Street	1.5	0	0	1	3.0	3.3	1	0	5	4.2	1.8	259
U. of Oregon	1.4	0	0	1	2.9	4.7	3	1	9	4.6	3.3	154
U. of Kansas	1.0	0	0	1	2.4	4.0	3	0	6	4.1	2.9	106
High Scope	1.4	0	0	1	2.9	3.6	1	0	6	4.7	2.2	191
U. of Florida	1.1	0	0	1	2.3	3.1	2	0	5	3.7	1.9	114
EDC	1.1	0	0	1	2.5	5.4	3	1	11	4.9	4.3	171
U. of Pittsburgh	1.1	0	0	1	2.8	3.4	2	0	5	4.1	2.2	117
REC	0.9	0	0	1	1.9	3.3	2	0	5	3.8	2.4	82
Enablers	1.0	0	0	1	2.3	2.9	1	0	4	4.1	1.9	236
NPV	1.0	0	0	1	2.4	2.9	1	0	4	3.5	1.9	669
Control 28	1.4	0	0	2	2.9	1.9	0	0	2	4.0	0.5	88

Table II-18

SUMMARY STATISTICS FOR WRTN DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	1.5	0	0	1	2.8	3.9	1	0	6	4.5	2.4	59
0209	2.1	0	0	2	4.1	4.0	2	0	6	4.8	1.9	56
0213	1.4	0	0	1	2.8	3.6	2	0	7	4.3	2.2	67
0308	1.4	0	0	2	2.3	7.4	8	3	11	4.3	6.0	76
0309	1.2	0	0	1	3.1	3.6	2	0	5	4.1	2.4	63
0316	1.2	0	0	1	2.5	4.5	2	0	8	4.8	3.3	77
0510	2.7	1	0	4	3.9	6.0	5	1	10	4.7	3.3	104
0511	0.5	0	0	0	1.6	1.2	0	0	1	2.4	.7	80
0512	0.9	0	0	1	2.2	1.9	1	0	2	3.1	1.0	75
0711	1.2	0	0	1	2.6	5.0	3	1	8	4.4	1.4	77
0714	1.6	0	0	1	3.2	4.4	2	0	9	4.8	1.6	77
0804	0.7	0	0	1	2.1	4.8	4	0	8	4.2	4.1	52
0808	1.3	0	0	1	2.8	3.2	2	0	4	3.9	1.9	54
0902	0.4	0	0	0	1.0	0.9	0	0	1	2.0	.5	68
0904	2.7	1	0	3	3.9	7.1	7	2	12	5.0	4.4	81
0906	0.7	0	0	1	1.4	1.6	0	0	2	3.1	.9	42
1002	1.7	1	0	2	3.0	4.3	3	0	8	4.3	2.6	42
1007	0.8	0	0	0	1.8	2.9	2	1	3	3.0	2.1	11
1010	0.8	0	0	1	1.9	2.3	1	0	3	3.1	1.5	61
1106	1.0	0	0	1	2.6	3.0	2	0	5	3.9	2.0	73
1108	1.2	0	0	1	2.4	7.2	7	3	12	4.8	6.0	98
1203	0.9	0	0	1	2.3	3.0	1	0	3	4.3	2.1	76
1204	1.6	0	0	1	3.6	4.0	4	1	5	3.8	2.4	41
2001	0.9	0	0	1	1.9	3.3	2	0	5	3.9	2.4	82
2702	0.6	0	0	1	1.5	1.9	0	0	3	3.1	2.5	62
2703	0.9	0	0	1	1.8	2.5	1	0	3	4.2	1.6	40
2704	1.9	0	0	2	3.5	5.1	3	1	10	4.9	3.2	75
2705	0.4	0	0	1	0.7	1.5	0	0	2	2.6	1.1	59

Table II-19

SUMMARY STATISTICS FOR WRTN DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	2.0	0	0	2	3.6	4.2	2	0	7	4.7	2.2	76
0711	0.7	0	0	1	1.4	2.9	1	0	4	3.7	2.2	57
0714	0.9	0	0	1	1.9	2.7	2	0	3	3.4	1.8	65
0804	0.8	0	0	0	2.1	1.5	0	0	2	2.8	.7	129
0808	0.8	0	0	2	1.1	1.5	0	0	2	2.1	.7	18
0906	1.3	0	0	0	3.3	2.6	1	0	4	3.8	1.3	39
1002	1.8	0	0	3	2.9	5.3	3	1	10	4.8	3.5	36
1007	1.2	0	0	1	2.6	4.3	3	0	8	4.4	3.1	24
1010	0.4	0	0	0	1.0	1.3	0	0	1	2.9	.9	44
1106	0.6	0	0	0	1.3	1.7	0	0	2	2.8	1.1	33
1108	1.3	0	0	1	2.8	7.0	7	2	11	4.5	5.7	64
2001	0.6	0	0	0	1.7	1.6	0	0	2	2.9	1.0	64
2801	0.5	0	0	0	1.1	1.3	0	0	1	3.0	.8	26
2802	1.8	0	0	2	3.5	2.2	0	0	2	4.0	.4	42
2803	1.9	1	0	2	3.0	2.0	1	0	2	3.3	.1	20

Table II-20

SUMMARY STATISTICS FOR WRTD DISTRIBUTION BY MODEL

Model	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
Far West	0.8	0	0	1	1.2	2.0	3	1	3	1.4	1.2	182
U. of Arizona	0.7	0	0	1	1.1	2.2	3	1	3	1.4	1.4	216
Bank Street	0.6	0	0	1	1.1	1.4	1	0	3	1.6	0.8	259
U. of Oregon	0.8	0	0	1	1.3	3.6	4	3	4	1.0	2.8	154
U. of Kansas	0.6	0	0	1	1.0	2.6	3	2	3	1.2	2.1	106
High Scope	0.6	0	0	1	1.2	1.7	2	0	3	1.6	1.0	191
U. of Florida	0.8	0	0	1	1.2	1.6	1	0	3	1.5	0.8	114
EDC	0.7	0	0	1	1.2	2.2	3	1	3	1.5	1.4	171
U. of Pittsburgh	0.5	0	0	1	0.9	2.1	3	1	3	1.4	1.7	117
REC	0.2	0	0	0	0.7	1.3	1	0	2	1.3	1.1	82
Enablers	0.7	0	0	1	1.2	1.7	2	0	3	1.5	1.0	236
NPV	0.5	0	0	0	1.0	1.6	1	0	3	1.5	1.1	669
Control 28	0.4	0	0	0	0.9	0.8	0	0	1	1.3	0.4	88

Table II-21

SUMMARY STATISTICS FOR WRTD DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	0.7	0	0	1	1.1	2.1	3	1	3	1.4	1.4	59
0209	1.0	0	0	2	1.4	2.0	2	0	3	1.5	1.0	56
0213	0.8	0	0	1	1.2	2.1	2	1	3	1.4	1.3	67
0308	0.9	0	0	1	1.1	2.4	3	1	3	1.3	1.5	76
0309	0.7	0	0	1	1.1	1.8	2	0	3	1.4	1.1	63
0316	0.6	0	0	1	1.1	2.2	3	1	3	1.5	1.4	77
0510	1.2	1	0	3	1.4	2.3	3	1	4	1.6	1.1	104
0511	0.3	0	0	0	0.8	0.8	0	0	1	1.3	.5	80
0512	0.2	0	0	0	0.5	0.8	0	0	1	1.2	.6	75
0711	0.6	0	0	0	1.2	3.5	4	3	4	1.2	2.9	77
0714	1.1	0	0	2	1.4	3.8	4	3	4	0.7	2.7	77
0804	0.6	0	0	1	1.1	2.7	3	2	3	1.2	2.1	52
0808	0.5	0	0	1	0.9	2.6	3	2	3	1.2	2.1	54
0902	0.2	0	0	0	0.5	0.7	0	0	1	1.1	.5	68
0904	1.2	0	0	3	1.5	2.7	3	2	4	1.5	1.5	81
0906	0.3	0	0	0	0.8	1.3	1	0	3	1.4	1.0	42
1002	1.1	0	0	3	1.5	2.1	3	0	3	1.6	1.0	42
1007	1.4	1	0	3	1.4	1.9	3	0	3	1.5	.5	11
1010	0.4	0	0	1	0.8	1.2	1	0	3	1.2	.8	61
1106	0.4	0	0	0	0.9	1.5	1	0	3	1.4	1.1	73
1108	1.0	0	0	2	1.3	2.7	3	2	4	1.4	1.7	98
1203	0.3	0	0	0	0.7	1.8	2	0	3	1.4	1.5	76
1204	0.7	0	0	1	1.2	0.8	3	2	4	1.3	.5	41
2001	0.2	0	0	0	0.7	1.3	1	0	2	1.3	1.1	82
2702	0.4	0	0	0	0.9	1.5	1	0	3	1.4	1.1	62
2703	0.4	0	0	0	0.9	1.0	0	0	2	1.4	.6	40
2704	1.5	1	0	3	1.5	2.8	3	3	4	1.3	1.3	75
2705	0.2	0	0	0	0.7	1.1	1	0	2	1.3	.9	59

Table II-22

SUMMARY STATISTICS FOR WRTD DISTRIBUTION BY SITE (NPV AND CONTROL)

	FALL					SPRING						
Site	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD	Mean Gain	N
0305	0.7	0	0	1	1.3	2.0	3	0	3	1.5	1.3	76
0711	0.5	0	0	1	1.0	1.7	2	0	3	1.3	1.2	57
0714	0.8	0	0	1	1.4	2.5	3	2	3	1.3	1.7	65
0804	0.3	0	0	0	0.8	0.8	0	0	1	1.2	.5	149
0808	0.3	0	0	0	0.6	1.8	1	1	3	1.3	1.5	18
0906	0.7	0	0	1	1.1	1.5	1	0	3	1.4	.8	39
1002	0.7	0	0	1	1.2	2.3	3	0	4	1.6	1.6	36
1007	0.7	0	0	1	1.0	2.4	3	2	3	1.1	1.7	24
1010	0.2	0	0	0	0.5	1.1	1	0	2	1.3	.9	44
1106	0.3	0	0	0	0.7	1.3	1	0	3	1.2	1.0	33
1108	0.6	0	0	1	1.1	2.7	3	0	4	1.5	2.1	64
2001	0.4	0	0	0	0.9	1.1	0	0	2	1.3	.7	64
2801	0.1	0	0	0	0.6	0.5	0	0	0	1.2	.4	26
2802	0.5	0	0	1	1.0	0.8	0	0	1	1.4	.3	42
2803	0.5	0	0	0	1.1	1.2	0	0	2	1.4	.7	20

Table II-23

SUMMARY STATISTICS FOR ITPA DISTRIBUTION BY MODEL

	FALL					SPRING						
Model	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD	Mean Gain	N
Far West	11.8	12	7	15	5.2	15.4	15	11	19	5.9	3.6	79
U. of Arizona	14.4	14	8	19	7.5	17.0	16	12	21	7.7	2.5	75
Bank Street	9.3	9	6	12	4.1	13.3	13	9	17	5.3	4.0	99
U. of Oregon	12.9	14	9	17	5.0	16.9	18	12	22	5.9	4.0	57
U. of Kansas	10.2	10	5	15	5.3	14.3	15	11	17	4.7	4.1	37
High Scope	11.0	10	7	14	5.8	14.1	13	11	17	5.1	3.1	75
U. of Florida	11.1	11	7	15	4.8	15.9	16	12	20	5.9	4.8	49
EDC	12.3	12	9	15	4.7	16.9	16	13	20	5.4	4.6	63
U. of Pittsburgh	8.8	9	5	11	3.8	15.6	15	11	20	6.3	6.7	51
REC	12.8	12	10	15	4.4	14.4	15	12	17	4.1	1.6	32
Enablers	12.1	11	9	15	4.6	14.2	14	10	18	5.7	2.1	117
NPV	10.8	10	7	14	4.5	15.5	15	11	19	5.9	4.7	263
Control 28												

Table II-24

SUMMARY STATISTICS FOR ITPA DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	9.3	8	6	11	4.1	14.3	14	8	17	6.8	5.0	28
0209	10.4	10	7	12	4.4	17.5	18	14	19	5.2	7.1	19
0213	14.9	14	12	18	5.1	15.1	14	11	19	5.4	..2	32
0308	22.4	22	19	25	5.3	22.7	21	19	26	6.4	.3	27
0309	10.0	9	7	14	4.6	14.8	13	10	17	7.1	4.8	25
0316	10.0	10	7	12	3.7	12.7	13	8	15	5.6	2.7	23
0510	9.1	8	6	12	4.0	16.9	17	12	20	5.4	7.8	36
0511	10.2	10	7	13	4.1	11.2	10	9	14	3.9	1.0	34
0512	8.4	8	5	11	4.3	11.3	11	8	16	4.4	2.9	29
0711	12.7	14	9	16	4.4	16.9	18	12	21	5.7	4.2	25
0714	13.1	12	9	18	5.6	17.0	17	12	22	6.3	3.9	32
0804	11.2	11	7	15	5.5	15.2	15	11	17	4.8	4.0	19
0808	9.2	8	5	13	5.2	13.3	14	11	17	4.6	4.1	18
0902	10.2	10	9	12	3.3	13.4	13	11	14	4.2	3.2	23
0904	13.8	13	9	17	6.8	15.2	15	10	20	5.7	1.4	31
0906	7.9	7	4	10	4.6	13.3	12	10	16	5.2	5.4	21
1002	14.2	16	10	18	4.3	17.5	18	12	22	6.0	3.3	15
1007	10.1	8	6	14	4.3	15.9	16	10	17	4.0	5.8	29
1010	9.7	10	5	12	4.6	15.0	14	9	20	6.5	5.3	25
1106	11.0	11	8	14	3.9	16.2	16	13	18	3.9	5.2	25
1108	13.2	12	10	16	5.0	17.4	17	13	22	6.3	4.2	38
1203	8.2	7	5	9	3.5	15.8	15	9	20	7.2	7.6	32
1204	9.9	10	5	12	4.3	15.3	15	12	16	4.7	5.4	19
2001	12.8	12	10	15	4.4	14.4	15	12	17	4.2	1.6	32
2702	11.6	12	10	13	3.3	11.9	11	9	15	5.8	.3	35
2703	12.7	10	8	16	6.6	12.9	11	10	17	5.2	.2	18
2704	13.3	12	10	17	4.0	8.2	18	15	21	3.9	4.9	37
2705	10.9	10	8	13	5.1	2.8	12	9	15	5.3	1.9	27

Table II-25

SUMMARY STATISTICS FOR ITPA DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	8.4	8	7	10	3.0	15.0	14	12	17	4.4	6.6	27
0711	13.9	14	10	16	4.5	20.0	19	16	24	6.4	6.1	24
0714	13.4	13	10	16	5.1	13.0	13	10	15	3.7	-.4	22
0804	9.3	8	6	12	4.1	12.2	11	8	15	5.1	2.9	54
0808	11.1	11	6	14	3.9	14.4	14	7	19	5.8	3.3	9
0906	7.6	7	6	9	3.1	16.7	15	13	19	6.0	9.1	19
1002	13.6	15	8	16	5.2	19.1	18	12	25	7.9	5.5	17
1007	9.5	9	8	12	3.9	12.9	13	11	15	3.7	3.4	11
1010	10.2	9	7	12	4.6	15.9	16	10	21	6.2	5.7	19
1106	9.2	10	8	10	1.9	15.8	15	10	19	5.6	6.6	9
1108	13.4	13	11	15	3.6	17.9	18	15	20	4.1	4.5	28
2001	10.6	10	7	14	4.4	16.2	15	12	19	5.5	5.6	24
2801												
2802												
2803												

Table II-26

SUMMARY STATISTICS FOR ETS DISTRIBUTION BY MODEL

Model	FALL Mean	Med.	LQ	UQ	SD	SPRING Mean	Med.	LQ	UQ	SD	Mean Gain	N
Far West	9.0	9	7	12	3.5	13.2	14	11	16	3.8	4.2	75
U. of Arizona	8.1	7	5	11	4.1	13.6	14	11	16	3.5	5.5	80
Bank Street	7.8	8	6	11	3.3	11.7	11	88	16	4.6	3.9	90
U. of Oregon	11.9	12	9	14	3.8	16.3	16	15	18	2.3	4.4	55
U. of Kansas	7.0	7	5	99	3.7	13.2	14	10	16	4.3	6.2	36
High Scope	9.7	9	6	13	4.5	12.3	13	9	15	4.4	2.6	69
U. of Florida	8.7	9	4	12	4.4	12.5	13	10	15	3.7	3.8	48
EDC	11.4	11	8	15	3.8	13.9	15	12	16	3.8	2.4	65
U. of Pittsburgh	8.4	7	5	12	4.5	13.4	14	11	16	3.9	5.0	47
REC	10.5	10	8	12	2.6	11.4	11	10	14	3.6	0.9	30
Enablers	11.4	11	8	15	4.7	12.9	14	10	16	4.5	1.5	78
NPV	9.2	9	6	12	3.9	12.3	13	9	15	4.0	3.1	255
Control 28												0

Table II-27

SUMMARY STATISTICS FOR ETS DISTRIBUTION BY SITE (PV)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0204	9.7	10	9	12	3.6	14.1	15	11	17	4.1	4.4	23
0209	9.2	9	6	13	4.0	11.3	12	7	14	3.9	2.1	21
0213	8.4	8	7	10	3.0	13.8	14	12	16	3.1	5.4	31
0308	11.7	12	10	14	3.1	15.2	16	14	17	3.1	3.5	27
0309	6.1	5	4	8	2.9	13.0	14	12	15	3.4	6.9	27
0316	6.4	66	4	7	3.6	12.5	12	10	16	3.6	6.1	26
0510	8.4	99	6	11	3.2	15.2	16	13	17	3.2	6.8	37
0511	7.9	7	6	10	3.1	9.0	8	7	11	3.6	1.1	30
0512	6.5	7	3	8	3.5	9.6	9	7	11	4.2	3.1	23
0711	10.5	11	8	12	2.8	15.4	15	13	16	2.3	4.9	22
0714	12.8	13	10	15	4.2	16.9	17	15	19	2.2	4.1	33
0804	9.4	8	7	12	2.9	15.2	15	13	18	3.5	5.8	18
0808	4.7	5	2	7	3.0	11.3	11	9	14	4.3	6.6	18
0902	7.6	9	5	9	2.9	10.0	10	6	13	4.6	2.4	21
0904	12.7	14	8	16	4.5	14.9	15	13	18	3.5	2.2	29
0906	7.7	7	4	10	3.6	10.9	11	8	14	3.6	3.2	19
1002	12.3	13	9	15	3.1	14.9	15	14	18	3.4	2.6	15
1007	6.9	5	3	10	4.5	12.9	13	11	14	1.9	6.0	9
1010	7.2	7	4	11	3.8	10.9	11	9	14	3.7	3.7	24
1106	9.6	10	7	11	3.7	11.2	12	7	14	3.8	1.6	27
1108	12.7	13	10	16	3.4	15.3	16	15	17	2.5	2.6	38
1203	7.4	7	5	11	3.7	12.0	13	8	15	3.8	4.6	29
1204	9.9	8	6	13	5.4	15.6	16	12	18	3.1	5.7	18
2001	10.5	10	8	12	2.7	11.4	11	10	14	3.7	.9	30
2702	0											
2703	8.9	9	6	11	3.8	9.9	11	5	13	3.9	1.0	17
2704	15.2	16	14	17	3.2	16.1	16	15	18	2.9	.9	35
2705	7.9	8	6	10	2.9	10.7	11	8	14	4.0	2.8	26

Table II-28

SUMMARY STATISTICS FOR ETS DISTRIBUTION BY SITE (NPV AND CONTROL)

Site	FALL					SPRING					Mean Gain	N
	Mean	Med.	LQ	UQ	SD	Mean	Med.	LQ	UQ	SD		
0305	8.9	10	5	11	4.0	12.6	12	10	16	4.0	3.7	29
0711	10.7	11	8	13	3.5	12.4	14	11	14	3.3	1.7	22
0714	10.4	10	9	11	2.3	13.7	14	12	15	3.3	3.3	20
0804	8.3	9	5	11	3.6	10.6	11	8	14	4.4	2.3	55
0808	6.1	5	4	6	2.9	12.1	12	10	13	2.6	6.0	8
0906	8.5	8	6	11	3.3	12.2	14	9	14	3.9	3.7	19
1002	13.0	13	9	16	3.3	14.8	15	13	16	2.1	1.8	17
1007	9.4	8	6	11	4.1	15.9	16	15	16	1.7	6.5	8
1010	7.2	8	3	10	4.9	10.4	9	6	14	5.4	3.2	17
1106	10.2	9	8	12	3.5	12.8	12	11	15	2.7	2.6	10
1108	11.1	11	8	13	4.3	14.2	15	13	16	2.6	3.1	27
2001	7.0	7	5	9	3.1	11.0	11	8	14	4.3	4.0	23
2801												
2802												
2803												

Table II-29

Means and Standard Deviations for Entire Head Start Sample in Fall 1971*

Test	Mean	Standard Deviation	n
PSI	14.59	6.16	2972
PPV	31.53	13.26	2996
WRTC	1.92	2.67	2980
WRTR	6.55	3.21	2980
WRTN	1.20	2.63	2980
WRTD	.61	1.10	2980
ITPA	11.28	5.16	1204
ETS	11.65	4.84	1129

*This sample includes all children on whom age information was available.

Chapter III

General Methodological Issues

Introduction

In Chapter I we gave an overall picture of the HSPV study. In Chapter II we took a first look at the data collected. Before going on to the analyses carried out, we thought the perspective provided by a general discussion of methodological issues would be valuable. We begin with a discussion of the major difficulties resulting from the study design. We then discuss the bag of statistical tricks usually used in attempts to overcome such difficulties. We at first ignore the thorny problem of measurement error, and later discuss its effects on the various statistical techniques. Finally, we discuss the general research strategy we have adopted.

Design Problems

Undoubtedly the most serious design problem in this study is the lack of randomization. If a group of experimental units is divided randomly into two or more groups, then providing the groups are sufficiently large, there is only a small probability that they differ significantly on any given variable, measured or unmeasured. Of course,

we can never be sure that the groups are equivalent with respect to all variables, but randomization is our best protection that there are no relevant group differences. If allocation to treatment groups is random we can be fairly confident that comparisons among group outcomes are unbiased even if no explicit account of pre-treatment variables is taken. We may still wish to use pre-treatment information to increase the precision of our comparisons, but with random allocation this information is more a luxury than a necessity.

In the HSPV study we would ideally have liked the group of children assigned to each model to be a representative sample of potential Head Start children. Since we cannot transfer children around the country at will, the smallest unit in which children can be assigned to models is the site. Thus, for purposes of model comparisons, randomization would have to be employed in the assignment of sites to models. Since there may well be systematic differences among the pools of children at different sites, it would be necessary to have several sites assigned to each model. With only 2 or 3 sites per model, substantial differences in the children assigned to various models would be likely even if randomization were employed.

Since we have so few sites and assignment was not

random, we cannot assume that the children assigned to various treatments are sufficiently alike to allow direct comparisons. A quick glance through the tables in the previous chapter reveals that, in terms of at least some variables undoubtedly associated with academic performance, there are some obvious and pronounced differences. There is clearly variation among models and among sites within models in ethnicity, age, mother's education, prior pre-school experience, and, most important, fall test scores. It will be necessary to in some fashion take account of these differences in our analyses. We will never know for certain, of course, whether our adjustments suffice to provide fair comparisons of program effects, but we hope to make a convincing case for their adequacy.

The unbalanced nature of the design in terms of background characteristics causes particular problems for the measurement of interaction effects. If we wish to relate program effectiveness to various background variables, we would like to have the distribution of these variables similar in the various programs, and representative of the full range of variation in the Head Start population. As an extreme example, suppose we are trying to relate model effects to ethnicity. Since the Pittsburgh model has only White children assigned to it, we have no data to address

the question of how its effect varies for different ethnic groups.

Of particular concern in connection with the lack of random allocation is whether the Control children differ in any systematic way from the Head Start children. We wish to use the Control results to estimate what would happen to a potential Head Start child if not enrolled in a pre-school program. Since the selection mechanism is of necessity different from that for Head Start, there may be important differences. For example, we note with some concern the fact that the Control children tend to be younger than the Head Start children. In fact, there are a substantial number of very young (less than 4 years old) children in the Control sample. These may well be waiting-list children deemed not yet old enough for Head Start. The fact that their mothers are applying so early may indicate that there is something special about such children. We really don't know, but we cannot be sure that Control children are sufficiently similar to Head Start children to be used to measure absolute effects of Head Start. Our attitude in general will be to compare Controls with Head Start children, but to be circumspect in interpreting the results.

approach is impossible to justify unless we have strong evidence that the groups are a priori equivalent.

The simplest approach which takes some account of pre-treatment differences is to compare average gain scores. A gain score is simply the difference between post-test and pre-test scores. By using gain scores we implicitly assume a mathematical model which states that on the average, if treatment effects were equal, the post-score would equal the pre-score plus some constant. This is a very restrictive model. It says, for example, that given the same program, if children with a fall score of 10 on the PSI obtain 17 on the average in the spring, then children with a score of 20 will on the average obtain 27. It is rare that a test is calibrated so as to make such an assumption reasonable.

Gain scores also, of course, take no account of background variables other than the pre-test. Analyses using gain scores as outcomes and adjusting for other variables are possible. It seems, however, more natural to use approaches which are more flexible in the way pre-tests and other variables are used together to adjust post-test scores. Three such approaches are how considered. In describing them we sl .ll refer to all variables used for adjustment of the post-test scores as "covariates."

A simple way of comparing programs is to cross-classify subjects on the basis of the covariates and directly compare the average scores for subjects in the same class in the two treatment groups. Suppose, for example, that we stratify children by ethnicity, age, and pre-test score. Then Whites between 50 and 55 months of age with fall scores between 10 and 15 in the two programs could be directly compared. Such comparisons will be unbiased with respect to the covariates used in the cross-classification. The approach is simple and the results easily understandable, but it generates a mass of information which may be difficult to use.

Suppose we can meaningfully specify a reference population (in terms of covariate distribution) which is of interest (often the entire sample is used for this purpose). Then by appropriately weighting the subgroup means for the two treatments, we can estimate the average outcome score which would result from applying each treatment to the reference population. This technique is known as <u>direct standardization</u>. For example, suppose we sub-divide according to sex and pre-score, and obtain the hypothetical results illustrated in Figure III-1. Note that the overall post-score mean for Model A is 19.3 and for B is 18.2, even though the mean for each sub-class is at least as high

for B as for A. Now suppose we apply the observed sub-group means to a standardized population with 25% in each of the four sub-classes. The standardized mean for Model A is now 19.0 and for B 19.5. These numbers present a fairer comparison, in that effects resulting from imbalances in sex and prior pre-scores have been removed.

The major difficulty with sub-classification approaches, including direct standardization, is that in order to exercise greater control over biases, we must sub-divide the sample more finely. This leads to fewer observations per sub-group and less precise estimation.

Possibly the most popular approach at the present time is the analysis of covariance (ANCOVA). It is based on the assumption that the expected value of an individual's post-test score is a linear function of a set of measurable variables. These may be continuous variables, dummy variables* representing membership in various classificatory groupings, or variables representing interactions among directly measured variables or transformations of them. Thus the expected outcome can in principle be expressed as a function of dummy variables corresponding to the programs we wish to

---

*A dummy variable is one which assumes a conventional value (usually 1) for all individuals with some specified property and another value (usually 0) for those without the property.

Figure III-1

Illustration of Direct Standardization*

re-score	Model A Male	Model A Female	Model B Male	Model B Female
≤ 12	15.0 (10)	18.0 (15)	15.0 (20)	19.0 (10)
> 12	23.0 (15)	20.0 (10)	24.0 (5)	20.0 (15)
ean	19.3	$(=\frac{10}{50} \times 15 + \frac{15}{50} \times 18 + \frac{15}{50} \times 23 + \frac{10}{50} \times 20)$	18.2	$(=\frac{20}{50} \times 15 + \frac{10}{50} \times 19 + \frac{5}{50} \times 24 + \frac{15}{20} \times 20)$
tandardized Mean	19.0	(= 25 X 15 + .25 X 18 + 25 X 23 + .25 X 20)	19.5	(= .25 X 15 + .25 X 19 + .25 X 24 + .25 X 20)

Reference Population

re-score	Male	Female
≤ 12	.25	.25
> 12	.25	.25

For each sub-class, the top number represents the corresponding mean and the number in parentheses the sample size.

compare as well as a variety of covariates. We can then estimate the relative effects of the treatments after "adjustment" for covariate differences, estimate the proportion of total post-test variance explained by program differences over and above that explained by the covariates, and test the significance of adjusted program differences. If the ANCOVA model is approximately correct, it is a powerful and flexible instrument for group comparisons. We shall discuss the theory underlying ANCOVA in more detail in Chapter VI.

A common approach which avoids the necessity to specify a particular mathematical form for the relationship between outcomes and covariates is matching. In its simplest form matching involves finding pairs of subjects in different treatment groups with effectively identical covariate values. Any difference between post-test scores of the members of such pairs cannot be attributed to differences on the covariates. Each pair provides an unbiased comparison between two treatments and by averaging we obtain an estimate of program difference. Since in practice we can almost never find exact matches, the efficiency of the matching procedure depends on our ability to find "good" matches. This can be a serious problem. The most up-to-date theory on this subject is by Rubin (1973).

If the assumptions of ANCOVA are approximately correct, it uses the data much more efficiently than matching. Matching, on the other hand, has the advantage of robustness. That is, it requires almost no assumptions on the form of the relationship between covariates and post-test scores to be valid. Combinations of matching and ANCOVA techniques are also possible. The interested reader is referred to Rubin (1973) and Smith (1973).

Effects of Measurement Error on Standard Analyses

Up to this point in discussing standard approaches to the problem of accounting for initial differences between treatment groups, we have not considered the fact that what we measure may be only an approximation to a true variable of interest. Classical measurement theory (see Lord and Novick, 1968) defines the reliability of a variable as the percentage of its variance (over some specified population) attributable to variation in the true score. This notion is meaningful if we assume that the observed score is the sum of true and error components, where the error has mean 0 and is uncorrelated with the true score.

In general, we are much more concerned about the

reliability of covariates than of the outcome measure. Under the classical measurement model, at least, the random noise introduced by errors in the post-test score tends to make our inferences less precise, but does not introduce systematic biases. Error in the covariates, on the other hand, causes serious problems. In the standardization approach, for example, we try to create relatively homogenous subclasses. If our classification is on the basis of variables measured with error, the subclasses may be less homogeneous than we believe. Substantial misclassification may result in serious biases.

Effects of measurement errors on ANCOVA can be equally devastating. Suppose that an ANCOVA model using the (unavailable) true covariate scores would accurately describe the situation. In the econometric literature, the equations relating expected outcomes to true covariates are known as structural equations. If we use our observable variables to fit a linear model, the resulting parameter estimates turn out to be biased estimates of the structural parameters. A biased treatment comparison will result, with the nature of the bias depending upon the nature of the measurement error.

In the one covariate situation, Lord (1960) and Porter (1971) assume the classical measurement model and suggest

techniques for obtaining fair comparisons. Other recent work (De Gracie and Fuller, 1972; Stroud, 1972) has addressed the question of "correcting" linear models under various assumptions about the errors. All these approaches are mathematically complex, and it is not clear at this point which, if any, are really suitable for educational quasi-experiments.

Instead of assuming the existence of a true model involving structural equations, we can decide to deal only with observables, and to build the best model we can. Under this approach the only way to insure against possible biases is to find covariates with high reliability, as well as a strong relation to outcomes. This approach has the advantage of simplicity. A sophisticated statistical correction which can be implemented only crudely may well be more misleading than no correction at all.

In matching, if there are errors in the covariates, we will be matching on the basis of possibly incorrect values. A true match would occur if the members of a matched pair had identical true scores. Under the classical measurement model, an individual's true score on a variable is on the average somewhere between his observed score and the mean for the population from which he is selected. Thus, two individuals may have the same observed

score on a variable but true scores which differ. If this variable has an effect on outcome scores, this (unobserved) difference may affect the observed post-test difference.

General Analysis Approach

In this section we discuss the general principles guiding our analysis plan. If we were dealing with a carefully designed randomized experiment, the analysis strategy would derive naturally from the design. Unfortunately, as explained above, we do not have such a situation. Campbell and Erlebacher (1970) have argued that the problems caused by lack of randomization combined with imperfect covariate reliability are virtually insurmountable. They seem to agree with Lord (1967) that "no logical or statistical procedure can be counted on to make proper allowances for uncontrolled pre-existing differences between groups." Certainly there is no substitute for a randomized experiment, but we feel that the "randomization or bust" reaction is overstated. By applying several alternative analysis strategies to our data, we feel it will be possible to obtain a fair assessment of the relative impacts of various preschool experiences. Each of our analyses will have its own strengths and weaknesses in terms

of the ability to detect real "effects." Each depends for its validity on a set of assumptions. These assumptions correspond to certain mathematical models which describe aspects of children's learning processes. We shall try to make the assumptions and corresponding models as explicit as possible, so that the reader can judge for himself the validity of the various analyses. At the very least, it should be possible to make conditional inferences of the form "if assumption A is true, hypothesis B is supported by the data." Moreover, the pattern of results from the whole set of analyses will hopefully give us more insight than would be possible with a single analysis strategy. In particular, for any one analysis, it is quite possible that a mathematical artifact will pass for a real effect. It is far less likely that an effect which shows up in several analyses based on different mathematical models is an artifact.

Use of multiple analyses, then is one of the principles guiding our analysis plan. Another principle is conservatism. Like Smith (1973), we intend to be cautious and conservative in declaring differences among models. We would rather risk missing a marginally significant difference than declare a difference significant when it really is not. There are two main reasons for this policy. First, on the basis of

both our own intuition and Smith's results, our expectation is that relative to the no-preschool condition, Head Start programs are quite homogeneous. Second, because of the implementation problems alluded to in Chapter I, we can never be sure that apparent effects are really the result of programs. In comparing models, we assume a strong common component to the experiences of children in a given model. In reality, a model is a complex combination of the sponsor's original conception and many factors which affect its implementation in a particular site, or class. In many ways it is preferable to consider the model-site combination as the treatment. Thus we shall look for effects which are consistent not only across analyses, but also across sites within models. Unfortunately, one or our models (REC) was implemented in 1971-72 at only one site (Kansas City). Although the data from this site will be analyzed and results presented, we shall not draw any conclusions about the REC model's relative effectiveness.

A third principle underlying our analyses is an emphasis on <u>estimation</u> of effects rather than formal testing of hypotheses. We feel that the resulting information is more useful. Given the large sample sizes with which we are dealing, statistical significance may be achieved by an effect which is educationally insignificant.

It is of course important to know whether an apparent effect could be the result of chance variation. In the setting of a complex quasi-experiment with multiple analyses, however, this is no easy matter. We can, in effect, use our data to test many hypotheses, using tests that are often not independent. Thus, even if the mathematical model underlying a testing procedure is correct, the formal significance level is illusory. There really is no true significance level. Significance levels should therefore be used as suggestive indicators rather than formal certifications of real effects.

In the next four chapters we attempt to implement the analysis plan discussed in this section. In Chapter IV we describe a "ranking analysis" which is intuitive and valid under minimal assumptions, though possibly conservative in detecting effects. In Chapter V we present a "residual analysis" which attempts to partition the observed gains for different models into a part attributable to natural maturation and a residual attributable to program effects. Chapter VI describes a conventional analysis of covariance. Chapter VII discusses a "resistant analysis" less sensitive to certain departures from the assumptions on which ANCOVA is based. The analyses described in Chapters IV, V and VII have not to our knowledge been used before in educational evaluation. We hope others may find these approaches useful additions to the standard "bag of tricks" described above.

Chapter IV

RANKING ANALYSIS

Theory of Ranking Analysis

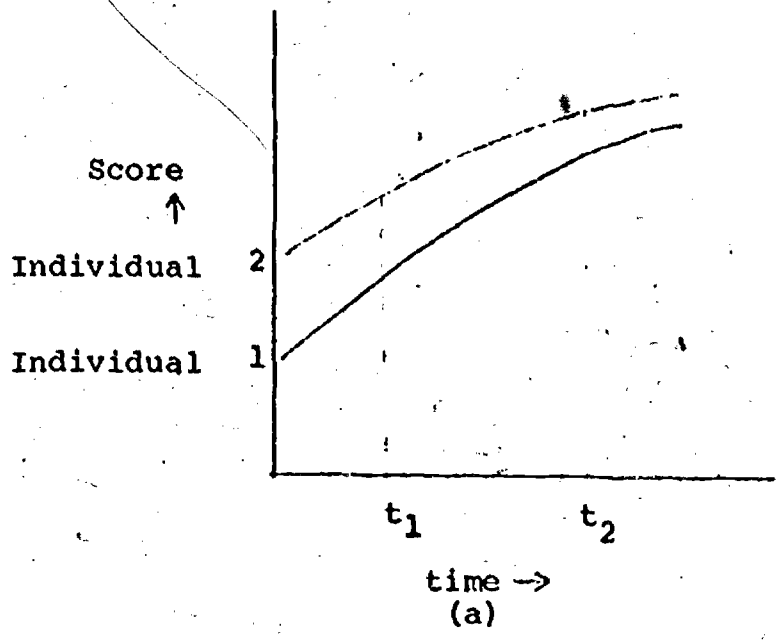
The ranking analysis is simple, intuitively appealing, and valid under minimal assumptions. It does not, however, provide a precise numerical measure of program effectiveness and may be conservative in detecting model differences. The other techniques described in this report provide more precise measurements at the cost of more stringent assumptions.

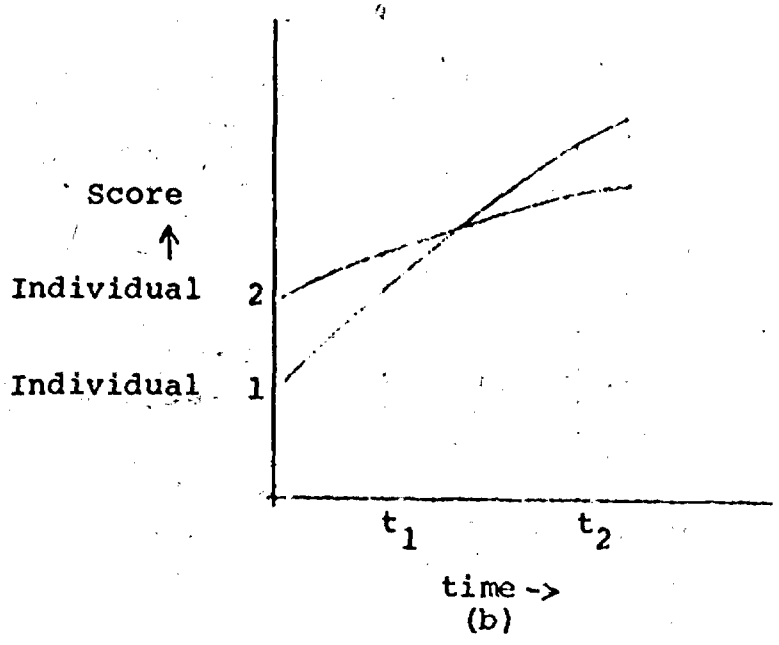
The ranking analysis is based on the idea that if two individuals are exposed to equally effective programs, the relative order of the individuals in terms of their true scores on a measure should remain the same from pre- to post-test. Suppose that each individual has a true score T which is an increasing function of time. Suppose further that we can measure the true scores of two individuals (say individual 1 and individual 2) in different programs at two points in time (t_1 and t_2). The relative order of these two individuals may remain the same (see Figure IV-1a) or be reversed (see Figure IV-1b).

In order to judge whether individual 1 has gained relative to 2 in situation (a), we would need some notion of

FIGURE IV-1

Two Possibilities for the Relative Growth of Two Individuals





the kind of growth patterns we would expect under the "null hypothesis" of no program differences. Without this knowledge we can make no inferences from (a).

In situation (b), on the other hand, it is clear that 1 has gained relative to 2, since he started below but ended above. Note that this argument is plausible regardless of any differences in the individuals' background characteristics. Individual 1 may be below 2 at time t_1 for any number of reasons (e.g. younger, mother's education less, less prior preschool experience). It seems reasonable that whatever factors cause him to be behind at t_1 will continue to operate so as to keep him behind at t_2, unless he is exposed to a more effective program. If programs are equally effective, the developmental process may be such as to change the difference between scores over time, but it seems unlikely that the growth curves will cross.

Basically, then, we assume that situations like (b) are evidence for differential program effects. Of course, there may be meaningful effects with (a), but we must rely on other analyses to detect them.

Let us now extend the above argument to the case of several models* and several individuals per model. Suppose we rank

---

*Recall that we consider the NPV sites pooled together as a model.

order the entire sample for both the pre-test and post-test. Let the individual with the highest score have rank 1, the second highest rank 2, and so on. If a particular model is effective (ineffective), then the ranks of the individuals in that model should tend to decrease (increase) from pre-test to post-test. Looking at the average rank on pre- and post-tests would be a simple way of assessing program effectiveness.

There is only one possible flaw in this argument. Suppose there are strong interactions between the relative effectiveness of the programs and some characteristic of individuals. Then the individual differences may be confounded with program effectiveness. For example, suppose program A is highly effective for boys but not for girls, and B is highly effective for girls but not boys. Then the observed relative effectiveness of the two programs will depend on the proportions of boys and girls in the two programs.

Even assuming there are no interactions between program effects and child characteristics, there is still another serious problem in implementing the above approach. We know that the reliabilities of our measures range from about .6 to .85. Thus, the observed rankings of individuals would be determined in part by random error variation.

To explore mathematical models to describe this situation would be a useful exercise, and might lead to statistical procedures which could be used to test the significance of observed rank changes. At present, however, no such procedure is available, and we have therefore taken a somewhat different tack in dealing with the problem of less than perfect reliability.

Often the reliability of certain group means is high, even when the reliability for individuals is quite low. By using the classroom, or site, as our unit of analysis instead of the individual, we may obtain higher reliability at the cost of fewer degrees of freedom. Since there is no sampling theory to provide significance tests, having a large number of degrees of freedom is not especially useful. We felt that our primary goal must be to achieve virtually perfect reliability. Site means met this requirement reasonably well.

The reliability of site means depends on the reliability of the test for individuals, the number of individuals per site, and the percentage of total individual variance which lies between sites. Thus measures will vary in terms of the reliability of site means, and the closer the reliability is to 1.0, the more confidence we will have in the analysis. For more detail on the way site mean reliabilities

were estimated, the reader is referred to Appendix B.

Using site means still leaves us with the problem that if model effectiveness is related to child characteristics, site variation in terms of such characteristics may come to be confounded with program effects. To get a partial handle on this problem, we performed the ranking analysis not only for the whole sample, but also for various sub-samples. We performed separate parallel analyses for Blacks only, Whites only, Mexican-Americans only, males, females, children with prior preschool experience and children with no prior preschool experience. If model effectiveness were strongly related to background characteristics, we would expect substantial differences in the results of these various analyses. By and large, the results were quite consistent, increasing our confidence in the validity of the ranking procedure applied to the whole sample.

In declaring a model particularly effective (or ineffective) on the basis of this analysis, we will take into account both the amount of the improvement in site ranks from fall to spring and the consistency across sites within a model. Note that since we are banking on virtually perfect site-mean reliability, we do not, in theory, expect any "random" component to the changes in rank. However,

since reliabilities are not actually perfect and there may be some interaction effects, we must expect a little variation not attributable to programs. Moreover, sites starting out low have more opportunity to improve their position "by chance," and sites starting out high have more opportunity to lose ground. Also, since the distribution of site mean scores is probably most concentrated near the center of the distribution, we would expect more change for sites nearer the center. It is difficult to weigh these factors. Our judgments are subjective, and the reader is encouraged to draw his own conclusions from Tables IV-1 through IV-9.

We conclude this section with a succinct reiteration of the two assumptions on which the validity of the ranking analysis depends.

Assumption 1: Developmental growth curves (in terms of true scores) for two site means will not cross during the period of program exposure unless the programs at the two sites differ in effectiveness.

Assumption 2: Site means have high enough reliability that the ranking of observed means is virtually identical to that of true score means.

TABLE IV-1

RESULTS OF PSI RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .90$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	14	7		
2.09	20	17		
2.13	24	18		
3.05			17	21
3.08	6	5		
3.09	27	29		
3.16	18	16		
5.10	1	14		
5.11	40	40		
5.12	39	38		
7.11	9	6	8	13
7.14	12	3	4	8
8.04	22	12	36	35
8.08	23	34	21	23
9.02	32	37		
9.04	2	2		
9.06	25	24	19	20
10.02	15	19	11	11
10.07	16	22	5	9
10.10	37	31	30	32
11.06	33	36	35	39
11.08	10	10	13	15
12.03	34	25		
12.04	7	4		
20.01	26	27	38	28
27.02	31	30		
27.03	29	26		
27.04	3	1		
27.05	28	33		
			19.8	20.5

TABLE IV-2

RESULTS OF PPV RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .88$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	5	3		
2.09	7	12		
2.13	15	10		
3.05			19	16
3.08	10	13		
3.09	23.5	27		
3.16	9	6		
5.10	16	24		
5.11	37	36		
5.12	34	34		
7.11	20	21	17	19
7.14	8	7	40	17
8.04	14	14	31	33
8.08	29	31	23.5	30
9.02	35	38		
9.04	2	2		
9.06	22	20	18	11
10.02	3	5	6	8
10.07	25	29	12	9
10.10	39	32	33	37
11.06	38	39	32	40
11.08	13	15	11	18
12.03	30	22		
12.04	4	4		
20.01	27	23	36	28
27.02	26	35		
27.03	21	25		
27.04	1	1		
27.05	28	26		
			23.3	22.2

TABLE IV-3

RESULTS OF WRTC RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .85$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	20	16		
2.09	15	20		
2.13	25	26		
3.05				
3.08	9	9	18	23
3.09	27	21		
3.16	24	25		
5.10	3	5		
5.11	40	39		
5.12	36	37		
7.11	10	12	14	18
7.14	1	2	4	1
8.04	31	7	29	33
8.08	22	15	23	17
9.02	39	38		
9.04	5	3		
9.06	21	28	26	29
10.02	11	19	17	6
10.07	8	8	13	13
10.10	16	24	19	36
11.06	30	22	34	34
11.08	6	10	7	14
12.03	37	32		
12.04	12	11		
20.01	33	31	38	40
27.02	28	27		
27.03	35	35		
27.04	2	4		
27.05	32	30		
			20.1	22.0

TABLE IV-4

RESULTS OF WRTR RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .60$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	18	13		
2.09	10	26		
2.13	17	18		
3.05			21	31
3.08	12	5		
3.09	39	29		
3.16	23	22		
5.10	6	11		
5.11	36	34		
5.12	32	36		
7.11	16	21	11	28
7.14	1	2	40	16
8.04	19	3	34	39
8.08	31	19.5	25	19.5
9.02	37	38		
9.04	4	7		
9.06	24	27	28	24
10.02	8	9	9	12
10.07	3	25	35	10
10.10	33	33	27	37
11.06	26	14	13	23
11.08	7	8	5	6
12.03	30	15		
12.04	14	1		
20.01	20	17	38	40
27.02	15	32		
27.03	29	35		
27.04	2	4		
27.05	22	30		
			23.8	23.8

TABLE IV-5

RESULTS OF WRTN RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .79$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	10	17		
2.09	3	16		
2.13	11	18		
3.05			4	14
3.08	12	1		
3.09	16	19		
3.16	17	10		
5.10	1	5		
5.11	37	39		
5.12	24	31		
7.11	20	8	33	24
7.14	8	11	26	26
8.04	31	9	29	37
8.08	14	21	30	35
9.02	40	40		
9.04	3	3		
9.06	32	33	15	27
10.02	7	12	6	6
10.07	27	25	18	13
10.10	28	29	38	38
11.06	21	22	36	32
11.08	19	2	13	4
12.03	23	23		
12.04	9	15		
20.01	25	20	35	34
27.02	34	30		
27.03	22	28		
27.04	5	7		
27.05	39	36		
			23.6	24.2

TABLE IV-6

RESULTS OF WRTD RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .75$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	13	17		
2.09	7	19.5		
2.13	11	18		
3.05			14	19.5
3.08	9	11		
3.09	17	22		
3.16	19	15		
5.10	3	13		
5.11	30	38		
5.12	40	37		
7.11	22	2	23	25
7.14	6	1	10	10
8.04	21	7	35	39
8.08	24	9	34	23
9.02	38	40		
9.04	4	6		
9.06	32	30	18	26
10.02	5	16	12	14
10.07	2	21	15	12
10.10	28	32	39	34
11.06	27	27	33	31
11.08	8	8	20	5
12.03	31	24		
12.04	16	4		
20.01	36	29	25	33
27.02	26	28		
27.03	29	36		
27.04	1	3		
27.05	37	35		
			23.2	22.7

TABLE IV-7

RESULTS OF ITPA RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .43$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	32	28		
2.09	21	6		
2.13	2	22		
3.05			36	24
3.08	1	1		
3.09	26	25		
3.16	27	36		
5.10	35	10		
5.11	24	40		
5.12	37	39		
7.11	13	11	4	2
7.14	11	9	8	32
8.04	16	20	31	37
8.08	34	30.5	17	26
9.02	22	29		
9.04	5	21		
9.06	39	30.5	40	12
10.02	3	7	6	3
10.07	25	16	30	33
10.10	29	23	23	15
11.06	18	14	33	17
11.08	10	8	7	5
12.03	38	18		
12.04	28	19		
20.01	12	27	20	13
27.02	15	38		
27.03	14	34		
27.04	9	4		
[illegible]	19	35		

TABLE IV-8

RESULTS OF ETS RANKING ANALYSIS

1 = Highest Site Mean
40 = Lowest Site Mean

$r_s = .66$

	PV		NPV	
SITE	FALL	SPRING	FALL	SPRING
2.04	15	14		
2.09	19	27		
2.13	24	15		
3.05			20	20
3.08	7	7		
3.09	38	17		
3.16	36	21		
5.10	23	9		
5.11	27	39		
5.12	35	38		
7.11	10	6	9	22
7.14	3	1	12	16
8.04	17	8	25	34
8.08	39	28	37	24
9.02	29	36		
9.04	5	11		
9.06	28	31	22	23
10.02	6	10	2	12
10.07	34	18	18	3
10.10	31	32	32	35
11.06	16	29	13	19
11.08	4	4	8	13
12.03	30	25		
12.04	14	5		
20.01	11	26	33	30
27.02				
27.03	21	37		
27.04	1	2		
	26	33		

Table IV-9

Results of Ranking Analysis for Control Children

1 = Highest Site Mean

43 = Lowest Site Mean

F = Fall

S = Spring

Site	PSI		PPV		WRTC		WRTR		WRTN		WRTD	
	F	S	F	S	F	S	F	S	F	S	F	S
2801	37	42	39	36	24	37	42	43	40	40	43	43
2802	28	38	27	30	31	41	37	42	7	30	25.5	38
2803	24	29	11	15	17	24	27	25	5	31	25.5	33

Results of Ranking Analysis by Test

In this section we present a brief summary of the results of the ranking analysis for each test. An overall summary of the ranking analysis with more interpretation of the results will be provided in the following section.

In order to estimate the relative effects of various models on a given test, we calculated the fall and spring ranks for all Head Start sites. These are displayed in Tables IV-1 through IV-8. Recall that the site with the highest mean is given rank 1 and the lowest rank 40. Thus a decrease in rank is evidence that a site has improved its position relative to the other Head Start sites.

As a rough measure of how much change has occurred overall, we have computed for each test the Spearman rank correlation (r_s) between the fall and spring rankings. The rank correlation measures the degree of similarity between two rankings of the same set of objects. Thus a value of r_s near 1 would indicate that the relative position of the sites has changed little from fall to spring, implying that model effects are quite homogeneous.

While we do not wish the Control results to influence our inter-model comparisons, we are interested in how the Control children perform relative to the Head Start children. We therefore calculated the ranks out of 43 total sites that

would have been occupied by the 3 Control sites had they been included with the 40 Head Start sites in the analysis. These results appear in Table IV-9.

Since there are 40 Head Start sites, the expected rank for a randomly selected site would be 20.5. If the NPV sites as a group do not differ from the PV sites, we would expect their average rank to be around 20.5 in both fall and spring. By looking at the average ranks for the NPV sites, we can get an idea whether they differ from the PV sites in initial level or effectiveness.

Preschool Inventory. We estimate the reliability of site means for the PSI to be between .98 and .99. The fall-spring rank correlation is .90. The mean rank for the NPV sites is 19.8 in the fall and 20.5 in the spring.

Looking now at the individual models, we find that Oregon and Pittsburgh show rank decreases (i.e., improvement) in both of their sites, and Far West in all 3 of its sites. Thus there is evidence that these 3 models are particularly effective in improving PSI scores. No model seems consistently ineffective, though Bank Street is something of a puzzle. In the fall, one Bank Street site (Tuskegee) has the highest mean of all 40 sites, while the other sites (Wilmington and Elmira) have the lowest means. The rank for Tuskegee slips from 1 to 14 in the spring, while

Wilmington and Elmira remain near the bottom. Although the Bank Street performance is rather poor, the fact that two sites start out so low leads us to suspect that some peculiarities of these sites may be more responsible than the model.

All 3 Control sites experience substantial increases in rank from fall to spring. This implies that Head Start programs were generally more effective than the Control "program" in raising PSI scores.

Peabody Picture Vocabulary Test. We estimate the site mean reliability of the PPV to be between .97 and .99. The rank correlation between fall and spring is .88. Mean ranks for the NPV sites are 23.3 in the fall and 22.2 in the spring. No model stands out as particularly effective or ineffective in raising PPV scores. Of the Control sites, one decreases in rank slightly, and the other two increase slightly. On the whole, the Control sites appear no less effective than the Head Start sites.

WRAT Copying Marks. As explained previously, the WRTC suffers from floor effects, so that the classical measurement model underlying our reliability estimates is probably inappropriate. This comment applies to the other WRAT subtests as well. Thus, although we calculate the site

mean reliability as .98, we are not sure that this figure is really meaningful. The fall-spring rank correlation is .85. The mean ranks for the NPV sites are 20.1 in the fall and 22.0 in the spring. The Kansas model stands out strongly, as both sites improve their positions dramatically. There is also a suggestion that the Pittsburgh model may be particularly effective, and the Florida model ineffective. All 3 control sites have much worse positions in the spring than in the fall.

WRAT Recognizing Letters. Our best estimate of site mean reliability is around .96. The rank correlation between fall and spring is only.60. The mean rank for NPV sites is 23.8 in both fall and spring. The Arizona, Kansas, and Pittsburgh models seem particularly effective, and the Enabler model particularly ineffective. There is a suggestion that High/Scope and Oregon are ineffective. Of the 3 Control sites, one decreases slightly and two increase slightly in rank from fall to spring.

WRAT Naming Letters. Site mean reliability is estimated at .96. The fall-spring rank correlation is .79. The mean ranks for NPV sites are 23.6 in the fall and 24.2 in the spring. Far West and Bank Street seem relatively ineffective. The Control sites do very poorly.

WRAT Reading Numbers. The estimated site mean reliability is only.92. The fall-spring rank correlation is .75. Mean ranks for the NPV sites are 23.2 in the fall and 22.7 in the spring. The Oregon, Kansas, and Pittsburgh models perform very well. Far West and Florida do poorly. The Control sites also perform poorly.

ITPA Verbal Expression. We estimate the site mean reliability to be between .95 and .98. The fall-spring rank correlation is only .43, indicating either that our reliability estimate is inflated, or that there is considerable variation in program effectiveness among sites (and possibly models). Mean ranks for the NPV sites are 21.3 in the fall and 18.3 in the spring. The Pittsburgh model appears most effective. There is a suggestion that Oregon and EDC are also effective, and that the Enabler model is ineffective. The ITPA was not administered to the Control children.

ETS. The estimated site mean reliability is .98. The fall-spring rank correlation is .66. Mean ranks for the NPV sites are 19.2 in the fall and 10.9 in the spring. Oregon, Kansas, and Pittsburgh seem particularly effective, High/Scope seems ineffective. The ETS was not administered to the Control children.

Summary of Ranking Analysis Results

In Chapter I we stated the three major questions on which our analyses will focus. In this section we present what evidence the ranking analysis provides bearing on these questions.

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged pre-schoolers acquire cognitive skills?

Our evidence here comes from the performance of the Control sites relative to the Head Start sites as a whole (PV and NPV). Of the six tests for which we have data on both groups, the Control children clearly lose ground relative to the Head Start children on four (PSI, WRTC, WRTN, WRTD). On the PPV, Head Start and Control children perform comparably. The PPV measures very general skills which are perhaps not easily taught in a pre-school program. On the WRTR, two Control sites (Huntsville and Sacramento) drop from near the bottom to the bottom two rungs. The third site (San Jose) has a rank of 27 in the fall and 25 in the spring. Tables II-15 and II-16 indicate, however, that because of the ceiling effect, there is little variability in spring site means. Many children are at the maximum score of 10. Thus the exact ranks have little meaning, since a difference of a few tenths of a point may

correspond to a large difference in rank. This also probably explains why the rank correlation for the WRTR is only .60. The upshot of all this is that we do not have very good evidence on whether Head Start is effective in teaching the ability to recognize letters.

2. Are the Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?

The evidence here is that overall PV and NPV sites are very comparable. The NPV sites as a group perform just about the way we might expect a randomly selected subset of 12 out of the 40 Head Start sites to do. In fact, according to the theory behind the Wilcoxon test (see e.g., Snedecor and Cochran, 1972), under this null hypothesis the mean NPV rank for a given test would have an approximately normal distribution with mean 20.5 and standard deviation 2.8. Since fall and spring rankings are not independent we cannot formally test the significance of mean rank changes. It is worth noting, however, that the NPV means for all tests, both fall and spring, lie comfortably within 2 standard deviations of 20.5.

3. Are some PV models particularly effective at imparting certain skills?

Table IV-10 presents a summary based on the information in Tables IV-1 through IV-9. There are a fair number of apparently strong positive or negative model "effects." In terms of tests, most of these effects (18 out of 22) occur in 4 of the 8 tests (PSI, WRTR, WRTD, ETS). These tests may be more sensitive to program differences.

In terms of models, it is interesting that of the 15 positive effects, 12 are for the "academic" models (Oregon, Kansas, Pittsburgh), which also show no negative effects. Moreover, all 8 ++'s are for these models. Thus, we have at least tentative evidence suggesting that the academic models may be generally more effective in transmitting academic skills. This may be the result of a test battery more sensitive to model differences, but we withhold final judgment until we have the results of our other analyses.

Besides the three academic models, the overall performance of the Arizona model is also somewhat encouraging. Arizona does not do really poorly on any test, and does well on the WRTR and ETS. No model does consistently poorly across the board.

Table IV-10

SUMMARY OF RELATIVE MODEL EFFECTIVENESS
BASED ON RANKING ANALYSIS*

++ Indicates model appears to be highly effective.
+ Indicates evidence for above average effectiveness.
- Indicates evidence for below average effectiveness.
-- Indicates model appears to be highly ineffective.

Model	PSI	PPV	WRTC	WRTR	WRTN	WRTD	ITPA	ETS
Far West	+				--	--		
Arizona				+				+
Bank Street					-			
Oregon	+					++		+
Kansas			++	++		++		+
High/Scope								-
Florida						---		
EDC								
Pittsburgh	+			++		++	++	+
Enablers				-				-

* REC not included because with only one site we felt it unfair to draw any conclusions.

Chapter V

RESIDUAL ANALYSIS

Introduction

The ranking analysis makes rather minimal assumptions on the nature of children's growth processes and how they are affected by the HSPV programs. It results in useful inferences about relative model effectiveness, but does not provide a precise numerical measure of program effectiveness. The "residual analysis" described in this chapter makes more stringent assumptions on the nature of growth processes and tries to provide a precise measure of the absolute effect of a Head Start program.

The basic idea is to estimate for each child the spring test score he would have obtained had he not been in a preschool program. Comparing this projected spring score with his actual fall and spring scores, we can estimate how much of his growth is the result of "natural" maturation and how much is a residual effect attributable to the program in which the child was enrolled.

Smith (1973) used this approach to estimate overall Head Start effects. We have refined his method and provided more theoretical underpinning. The basic theory of the residual analysis is presented heuristically in the follow-

ing section. The statistically sophisticated reader is referred to the more mathematical discussion in Appendix D.

Theory of Residual Analysis

The major assumption underlying the residual analysis is that variation displayed by fall test scores in our sample reflects developmental trends over time. More specifically, suppose we could look at a sub-sample with identical values of all background characteristics except age. Suppose we observe the mean score for such individuals in our sample as a function of age. Then we are assuming that the resulting curve is very similar to the natural developmental growth curve for such children as they grow older.

In general this will be true unless our sample has selection biases which imply a relationship between age and ability. Suppose, for example, that the younger children in our sample tend to be particularly clever as a result of the way the sample was selected. Then children who are, say, 52 months old at fall testing will on the average do better at spring testing (say 6 months later) than those children in our sample who are 58 months old in the fall do in the fall.

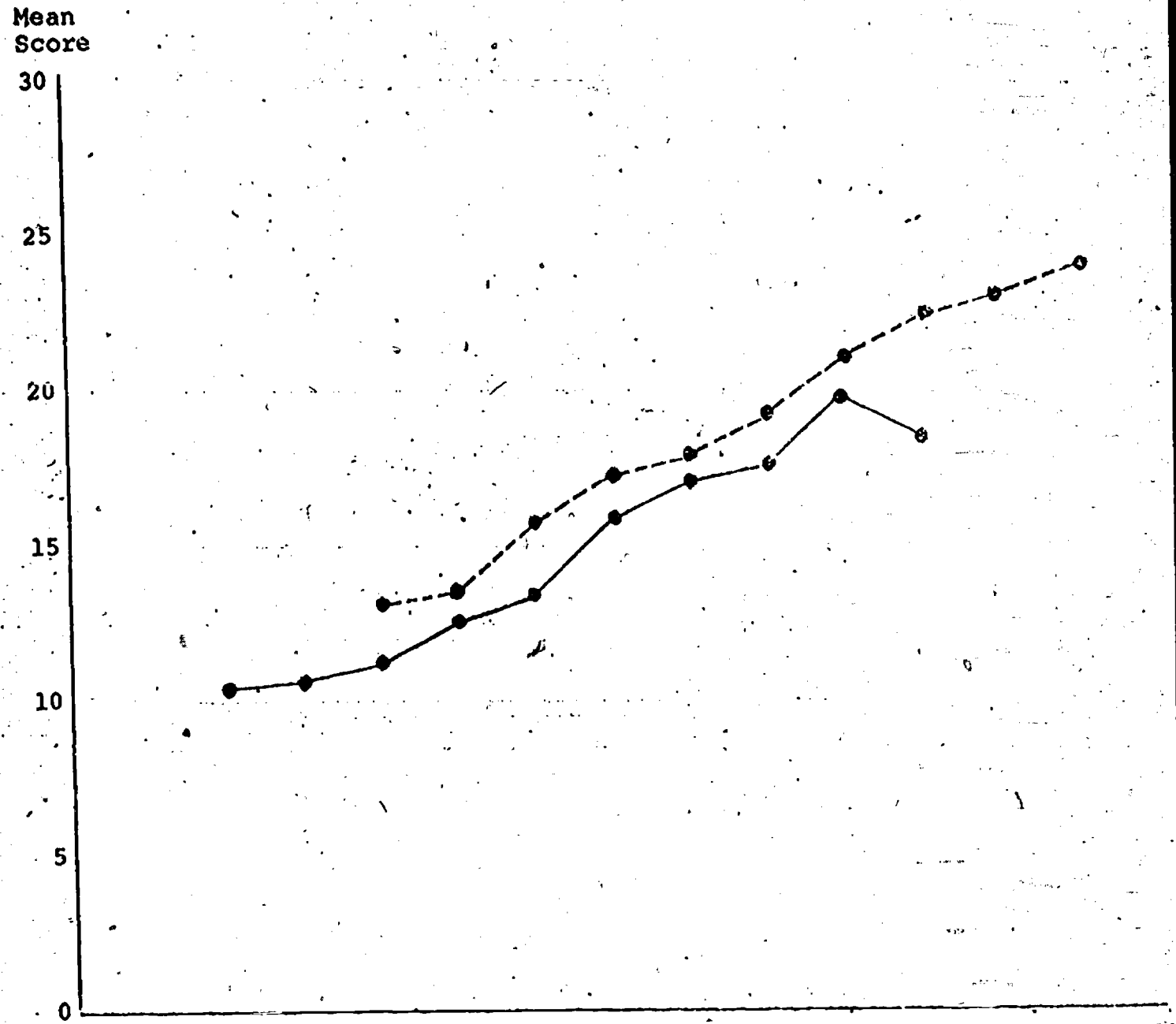
As an introduction to the full-blown residual analysis, we first present a more intuitive graphical analysis which will provide a rough idea of the overall effect of Head Start. Suppose we graph the mean outcome score for a particular group of Head Start children as a function of age for both fall and spring. Figure V-1, for example, presents the results on the PSI for all children with no prior pre-school experience. We have divided the sample into 3-month age groupings and plotted the mean for each group.

If Head Start had no effect for these children, we would expect the fall and spring curves to be similar. Of course, some age groups may turn out to be a bit cleverer than others and there will be sampling fluctuations so that the curves will not be identical, even if Head Start has no effect. Suppose, however, we find that the spring curve is consistently above the fall curve. Then unless there is a selection effect in the sample implying a consistent negative relationship between age and ability, we have evidence that Head Start has raised the level of the growth curve. The difference between the curves provides a rough estimate of the value added by Head Start over and above that expected on the basis of natural maturation.

Figure V-1

PSI - FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE

Fall ———
Spring ------



ge	40		46		52		58		64		70		76		8
pre	4	16	63	207	374	397	368	257	162	165	119	52	2	1	
post		3	5	6	19	94	284	311	313	300	196	146	124	98	1

For example, according to Figure V-1, children with no prior preschool between 51 and 53 months of age in the fall averaged approximately 11.3 on the PSI. Suppose for simplicity that there were 7 months between fall and spring testing for all children.* Then we would expect these children to obtain an average score in the spring of about 14.5 without Head Start. Their average spring score was in fact about 16.5. The difference of 2 points represents a residual effect, possibly attributable to Head Start, over and above the expected natural growth.

The results of the various graphical analyses we carried out are difficult to summarize verbally. The interested reader is referred to Appendix C, where the resulting graphs are presented.

If in fact certain age groups are cleverer than others, such differences in ability may be at least partially associated with various background characteristics. Thus, selection biases can be reduced by carrying out the graphical analysis on various sub-classes of the total sample. For example, one sub-class might be Black males with prior pre-school experience. The difficulty, of course, is that the more refined we make our subclassifications the smaller the sample sizes become and the more complex

---

*The actual time varied from 6 to 9 months.

the interpretation. Thus, the growth curves would be estimated very imprecisely. One way to avoid this dilemma is to define a mathematical model to describe the developmental process which, if correct, makes more efficient use of the data. This brings us to the full residual analysis.

Our first task is to build a mathematical model to enable us to predict the expected value of the test score of a child not in any pre-school program on the basis of his age and other background characteristics. We do this by the technique of regression analysis. The details are described in the next section.

Suppose, now, that we have such a model Then for any child in our sample we can compute Δ , the increase in score he would be expected to achieve between fall and spring testing on the basis of natural maturation. This is done by up-dating his age the appropriate number of months, leaving other background variables unchanged, and calculating the effect this would have according to our model.

For example, let Y and Y' represent fall and spring test scores respectively. Let AGE1 be the age at fall testing, and MOMED be the number of years of mother's education. Suppose our regression model is given by

$$Y' = 4 + .2 \times AGE + .1 \times MOMED + e \qquad (5.1)$$

where e represents random error. In this case Δ is simply .2 times the number of months between fall and spring testing.

Having calculated Δ, we can calculate an estimate of the child's expected spring score in two reasonable ways. One is to simply add Δ to the observed fall score (Method 1). For example, suppose we have the model described by equation (5.1), that there are 7 months between tests, and that for a given child AGE1 is 50, MOMED is 10, and his fall score is 17. Then:

Method 1 Expected Spring Score = $Y + \Delta = 17 + .2 \times 7 = 18.4$

The second way (Method 2) to estimate the expected score uses the regression model directly. We simply substitute the appropriate values of background variables (including the age at spring testing) in our regression equation. We obtain an estimate of the expected score which we shall call $\hat{Y}'$. Note that if we do the same thing using the age at fall testing, we obtain a predicted fall score $\hat{Y}$, which is what we would expect the individual to have achieved in the fall on the basis of his background characteristics. Because the regression model is linear, the Method 2 expected spring score is mathematically equivalent to adding Δ

to the predicted fall score rather than the actual fall score.

In our example

$$\hat{Y} = 4 + .2 \times 50 + .1 \times 10 = 15$$

$$\hat{Y}' = 4 + .2 \times 57 + .1 \times 10 = 16.4$$

Thus

$$\text{Method 2 Expected Spring Score} = \hat{Y}'$$

$$= 4 + .2 \times 57 + .1 \times 10$$

$$= (4 + .2 \times 50 + .1 + 10) + .2 \times 7$$

$$= \hat{Y} + \Delta .$$

Finally, the residual attributable to the child's program for each method is the difference between the observed and expected spring scores. Thus, for Method 1 we have

$$r_1 = Y' - (Y + \Delta).$$

and for Method 2

$$r_2 = Y' - \hat{Y}' = Y' - (\hat{Y} + \Delta)$$

Thus in our example, suppose the spring score is 19. Then we have

$$r_1 = 19 - (17 + 1.4) + .6$$

$$r_2 = 19 - 16.4 = 2.6$$

Note that if the regression model were exact (perfect prediction, no error variance) and the test perfectly reliable, then the two methods would be equivalent and would perfectly estimate the "true" effect. In a real situation the tradeoff between the two methods hinges on whether the observed or predicted fall test is a better estimate of the true fall test score. Roughly speaking, higher test reliability favors Method 1 (using the observed fall score), while more accurate regression equations favor Method 2 (using the predicted fall score). The existence of this tradeoff suggests that a weighted combination of the estimates provided by the two methods may be optimal. We shall have more to say about the appropriate weighting in the next section. A more theoretical discussion of the issue is presented in Appendix D.

We can compare various Head Start programs by estimating the mean residuals for their program groups and comparing them. Since the residuals reflect the increase in score beyond that expected on the basis of natural maturation, they provide an absolute measure of program effect. Thus the "effect" for the Control "program" can also be calculated

and used as a check on our methods. If the Control children really are similar to Head Start children in the fall, we would expect their residuals to average close to zero. That is, we would expect their spring scores to reflect only maturation. If the Control residuals are substantial, there are at least three possible explanations. There may be some sort of test sensitization or practice effect; some of the Control children might actually be involved in a preschool program; or there may be a selection bias in the fall sample causing us to underestimate the slope of the growth curve. If the slope is underestimated, we underestimate the projected spring scores and overestimate the residuals. The upshot of this discussion is that small residuals for the Controls is evidence that our technique is working as it should. Large residuals are troubling, since they may mean either that the analysis is in some way incorrect or that our Control children fail in some way to be legitimate controls.

In concluding this section, let us summarize the two major assumptions on which the residual analysis is based.

Assumption 1: The relationship between fall test score and age in our sample accurately reflects the developmental process occurring over time.

Assumption 2: A linear regression model adequately represents the relationship between fall test score and

background characteristics (including age).

In the following section we describe the way in which the regression models used in the residual analysis were developed.

Regression Models

In this section we describe the derivation of the regression equations necessary to implement the analysis approach described in the previous section. These equations are also of some interest in their own right as descriptive statements about the developmental process for young children.

We attempted to build the best possible model to describe the relationship between expected outcome scores and our measured background characteristics in the absence of program effects. To do this, we tried to explain as much variance as possible in the pre-test scores with the measured background characteristics. By performing regression analyses using fall scores for the entire Head Start sample as the dependent variable, we eliminated program effects. Moreover, the sample size was large enough to ensure accurate estimation.

We began with some exploratory regression analysis involving as independent variables a wide variety of back-

ground variables, including age (AGE 1), sex (SEX), mother's education (MOMED), ethnicity (ETHBL, ETHWHITE), family income (FAMINC), first language (FLANG), prior preschool experience (PS, PSMNTHS), household size (HHSIZE), and sex of the head of household (SEXHH). From here on we will often for convenience, refer to the abbreviations for variables defined in Appendix A, where the exact coding for all variables can also be found.

For all tests it was found that restricting attention to AGE1, SEX, MOMED, FLANG, and the ethnicity and prior preschool variables lost very little in terms of R^2, the proportion of fall score variance which could be explained. We were concerned that the effects of first language and prior preschool might be particularly complex. We therefore decided to divide the total sample into the following 4 exclusive groups, and to fit a separate regression model for each group:

Group 1: FLANG : First language not English
Group 2: PSNHS : Non-Head Start prior preschool experience
Group 3: NOPS : No prior preschool experience
Group 4: PSHS : Prior Head Start experience

There are a few children with prior preschool experience whose first language was not English, but since

the sample was rather small to begin with, we did not separate them out. A dummy variable indicating previous preschool experience was, however, included in the regression for Group 1.

In all 4 groups we eliminated from the analysis the very few children who were not Black, White, or Spanish American (Mexican American or Puerto Rican). Since Group 1 was comprised almost entirely of Spanish Americans, we felt the results would be more meaningful if the others were eliminated. The analyses for Groups 2 and 3 contain dummy variables for both Black (ETHBL) and White (ETHWHITE). Group 4 had nearly all Blacks and Whites, and the analysis included a dummy variable for Black only.

Note that for children with some prior preschool experience, their "natural" developmental process may have been altered in a variety of ways. Thus, although our approach is probably most suitable for Group 3 (which incidentally comprises about 2/3 of the sample), we have decided to carry through the analysis for the other groups as well, but to be somewhat careful in interpreting the results.

In addition to the basic variables themselves, all 2-way interactions among them were also considered. Several analyses were run for each of the 4 groups in an attempt to

obtain a large R^2 with as few variables as possible. The final set of equations selected are presented in Tables V-1 through V-8. Each row of these tables represents one equation. Variables are specified by the column headings. The entries in any column are the regression coefficients associated with the specified variable. For the PSI for children with no prior pre-school experience, for example, the regression equation reads:

Expected PSI score = 6.518 + .0544 X AGE + .8883 X SEX + 1.549 X MOMED + .0341 X AGE1 X MOMED - 1.977 X ETHWHITE - .1931 X ETHBL X MOMED + .2596 X ETHWHITE X MOMED.

Before discussing these equations in more detail, we present two digressions which may aid the reader in interpreting them. First we consider the interpretation of the coefficient of an interaction variable, and then the significance of R^2, the proportion of variance explained by our independent variables.

Interpretation of Interaction Coefficients

Consider any two independent variables, such as AGE1 and SEX. The interaction variable is simply the product of the two, e.g. AGE1 X SEX. The coefficient for such a

RESRESSION EQUATIONS USED TO COMPUTE PSI RESIDUALS

TABLE V-1

Group	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
First Lang. Not English	5.286		-13.79	-1.376	.3764	.0363	-.8963										204	.324
Non HS Prior PS Exp	54.81	-.8108		-5.662		.1148	.1514	6.013					-3.483	-1.019			172	.416
No Prior PS Exp	6.518	.0544	.8883	-1.549		.0341			-1.977					-.1931	.2596		2087	.328
Prior HS Exp	-.49	.1426		-1.032		.0295	.1242	4.902						-.7094		.1616	529	.317

REGRESSION EQUATIONS USED TO COMPUTE PPV RESIDUALS

TABLE V-2

oup	C.	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
rst ng. t glish	100.7	-1.767		-9.274	.5625	.2122	-3.952									.8612	227	.261
n HS ior s xp	214.0	-3.18		-21.56		.4033		35.41	-8.809				-4.208				173	.429
rior s xp	-13.31	.6331		.7754					-26.40	-.0866	.3290		-1.777		1.211		2111	.340
rior s xp	-18.32	.7345		1.13	-.1873		1.136	11.28		-.1277				-1.066		.061	517	.258

REGRESSION EQUATIONS USED TO COMPUTE WRTC RESIDUALS

TABLE V-3

p	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
t ish	-6.99	.1186	-7.043		.1744	.0037	-.2773										222	.316
HS r	66.27	-1.076		-7.16		.1261		11.95	-3.145	-.2744							171	.404
r	1.763			-1.184	.0102	.0209		5.878		-.1178	-.0526				.2450		2099	.265
r	8.273	-.1404		-1.62		.0310	.1299	2.925						-.3713			517	.248

REGRESSION EQUATIONS USED TO COMPUTE WRTR RESIDUALS

TABLE V-4

p	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
st e- lish	3.481				.0420	.0048	-.3222										222	.038
HS or	-1.903	.088			.0093	.0063		6.213		-.0994		.8784		-.088	-.0475		171	.134
or	-4.488	.1624		.1562	.0065			1.341	-2.638	-.0344					.2338		2099	.138
or	1.887	.0462				.0038	.0826			.0299				-.2129			517	.106

REGRESSION EQUATIONS USED TO COMPUTE WRIN RESIDUALS

TABLE V-5

Group	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
rst g. t glish	.804		-2.236		.0307											-.384	222	.017
n HS ior p	-2.199		-7.85		.0787	.0062	.3175				.0171						171	.145
ior p	-2.031		.1531	-.1541			.0079	5.109		-.0482			.3521	-.256			2099	.083
ior p	-3.154			-.2809		.0199	.0382	6.105		-.0524				-.3354		.0277	517	.120

REGRESSION EQUATIONS USED TO COMPUTE WRTD RESIDUALS

TABLE V-6

oup	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
rst ng. t glish	-.742	.0115	-2.639	.052	.0548		-.0749										222	.076
n HS ior p	-1.307			-.2162		.0078		5.944		-.0796				-.166			171	.198
ior p	-.688			-.1708	.0027	.0051		2.289		-.028		-.0764		-.0824			2099	.135
ior p	4.8[illegible]	-.0787		-.7172		.013								-.0259		.0109	517	.161

REGRESSION EQUATIONS USED TO COMPUTE ITPA RESIDUALS

TABLE V-7

Group	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
rst g. t glish	-42.48	.7520	16.90	3.911		-.0496	-1.672										81	.205
n HS ior p	-9.902			-1.43	.098	.0541		56:20		-.6066		-5.708		-1.726	.2684		63	.431
ior p	.671	.1602	-.6656				.127		-12.92		.1802				.3577		822	.187
ior p	-20.63	.441	-1.549	.7227				22.84	-.287			1.997		-.7749		.0874	209	.207

REGRESSION EQUATIONS USED TO COMPUTE ETS RESIDUALS

TABLE V-8

...up	C	Age	Sex	Mom's Educ.	AxS	AxME	SxME	Black	White	AxB	AxW	BxS	WxS	BxME	WxME	Months HS	n	R^2
...st ...g. ...lish	-6.450	.2069		.2912	.1325		-.6182										80	.300
...HS ...or ...p	-40.04	1.099	-28.77	2.995		-.0712	2.634				-.034	-.2118					58	.398
...ior ...p	4.639			-1.38	.0103	.0311		11.79		-.1277				-.5484			757	.297
...ior ...p	-9.678	.2739				.0056		7.292		-.0866				-.2976			205	.296

variable may be interpreted as the effect that a unit increase in either of the variables has on the effect of the other variable. For example, the coefficient of AGE1 X SEX* in the PSI Group 1 equation is .3764. Thus we can say that, all else being equal, the rate of increase of PSI with age is .3764 points per month higher for girls than for boys. We can equivalently say that the advantage of being a girl rather than a boy increases by .3764 points per additional month of age (or more accurately here the disadvantage descreases).

If an interaction coefficient is statistically significant, it means that the combination of the two variables has an effect over and above what can be adequately described by simple additive effects. In our example, the advantage which accrues to an older girl is greater than the sum of the effects of being a girl and being older.

Significance of Explained Variance (R^2)

For the residual analysis to be valid, it is not necessary that R^2 be very large. Roughly speaking, the larger R^2 the smaller will be the variance of the resulting

---

*SEX is coded 1 for girls and 0 for boys.

residuals, however, and the more precise the analysis will be. In many contexts in which regression analysis is used, R^2 serves as a measure of the strength of the independent variables as predictors of the dependent variable. In the present context, the value of R^2 is determined by three factors, in addition, of course, to random fluctuations.

First there is the importance, or strength, of effects on the dependent variable attributable to the independent variables, as measured by the regression coefficients. Second, there is the variability in our sample. To take an extreme example, if there were no Blacks in our sample, the ETHBL dummy variable could explain no outcome variance regardless of the true effect of being Black.

If the distribution of the independent variables is similar to that in the population to which we wish to generalize our regression results, the R^2 for that population will be similar. If the distribution is different, R^2 may be quite different, even if the effects are the same. The values of the regression coefficients do not depend on the distribution of the independent variables in our sample, although our ability to estimate these coefficients accurately might.

The third factor on which R^2 depends is the reliability of

the outcome measure*. A test score Y can be thought of as containing a true component T corresponding to a stable characteristic of interest, and a random error component e. Part of the variance of the true part T can be related to measured independent variables via regression analysis. The higher the reliability of Y, the larger the proportion of its variance attributable to T; hence the more potentially explainable variance. Thus, the reliability sets an upper bound on the proportion of variance explainable.

The upshot of this discussion is that although maximizing R^2 is desirable for maximum precision, the value of R^2 is determined by a complex interaction of factors, making its interpretation difficult. The interpretation of the coefficients determining our regression model, on the other hand, is straightforward, and for large sample sizes the estimation of these coefficients should be accurate.

Let us look now at the regression models. It is difficult to summarize all the implications of these equations in any simple way. To measure the net effect of a variable in a particular equation, it is necessary to consider also all interaction variables with non-zero coefficients which involve this variable. Suppose, for example, we are

---

*We assume here that the independent variables have perfect reliability. Unreliability in the independent variables introduces further complications. See Chapt. VI.

interested in the effect of age on PSI score for Group 1 children. Then a one month increase in age corresponds to an average PSI increase of

$$\Delta = .3764 \times SEX + .0363 \times MOMED$$

Thus the effect of age in this case depends on the values of SEX and MOMED. For a boy with MOMED = 12,

$$\Delta = .3764 \times 0 + .0363 \times 12 = .436.$$

For a girl with MOMED = 10

$$\Delta = .3764 \times 1 + .0363 \times 10 = .739.$$

At first sight it may appear that the equations for the four groups on the same test differ wildly. Closer inspection of the net effects of each of the variables reveals that over the range of values found in our sample the equations are quite similar.

With 8 tests and 4 groups, we might not expect to find consistent patterns in the effect of a given variable across the various equations. Surprisingly, certain patterns do emerge. These are summarized in Table V-9.

It seems that SEX and interactions involving it tend not to show consistent patterns, while the effects of age, ethnicity, and mother's education are quite consistent.

In concluding this section, let us state that we believe

Table V-9

Effects of Background Variables on Fall Scores

Variable	Net Effect
AGE 1	Nearly always +
SEX	No consistent pattern except in grp. 1 where -
MOMED	Generally + except sometimes for young black children and girls in grp. 1
AGE1 x SEX	Nearly always +
AGE1 x MOMED	Nearly always +
SEX x MOMED	Generally + except for Group 1
AGE1 x ETHBL	Generally -
ETHBL x MOMED	Generally -
ETHWHITE x MOMED	Generally +
ETHBL x SEX	No consistent pattern
ETHWHITE x SEX	No consistent pattern

the regression models described in Tables V-1 though V-8 to be reasonably accurate, concise mathematical descriptions of the relationship between test scores and various background characteristics in the absence of pre-school intervention for Head Start-age children. In carrying out the residual analyses in the following section, we shall use these equations in this way. We recognize that to be formally correct, we should in some way take into account sampling errors in the estimation of the regression coefficients. This would, however, make the analysis almost impossibly complex and add little to our confidence in the results.

Implementation of Residual Analysis

In this section we present brief descriptions of the various ways in which the residual analysis was implemented and the basic tables of results. At the outset, let us remark that the residual analysis was carried out for only 6 of the 8 tests. Since the WRAT Recognizing Letters and Naming Letters subtests generally had such low values of R^2, we felt the regression models were of questionable validity, particularly in light of the floor and ceiling effects. The WRAT Reading Numbers seemed of borderline acceptability, but we decided to include it.

For the six tests, we calculated for each child in the sample with the necessary data a Method 1 and Method 2 residual. As explained above, to compute the Method 1 residual it was first necessary to calculate Δ, the expected increment. This was found by incrementing the child's age by the actual number of months between his fall and spring tests and calculating the effect this would have according to the appropriate equation. Adding Δ to the fall test score, we obtained the predicted spring score in the absence of a pre-school program. Finally, the residual was found by subtracting the predicted score from the actual spring score. To find Method 2 residuals we obtained the predicted spring score by simply substituting the child's age at spring test time in the appropriate equation.

In Tables V-10 through V-15, we present the results of the Method 1 analysis by model. For each model, we present the mean fall score $\bar{Y}$, the mean spring score $\bar{Y}'$, the mean expected increment $\bar{\Delta}$, the mean residual $\bar{r}_1$, and the sample size. Thus, we partition the mean gain $\bar{Y}' - \bar{Y}$ into two parts: the increase we would expect on the basis of natural maturation, and the residual increase over and above this which <u>may</u> be attributable to the effect of the model.

Method 2 results are presented in Tables V-16 through V-21. Here we present the mean predicted fall score $\bar{Y}$ based on the regression equations, the mean spring scores, the mean expected increment, mean residual $\bar{r}_2$ and sample size. Spring score and expected increment means would be identical for the two methods except for the fact that the analyses are based on slightly different samples. Also, because of the nature of least-squares regression analysis, the observed and predicted fall score means for the entire sample would be identical. For particular programs, however, they might differ, since a program group may on the average do better or worse in the fall than we would expect on the basis of background characteristics.

We will find an apparent model effect more credible if it is consistent across all the sites in the model. To check this, we have computed the mean residual by both methods for each site. These results appear in Tables V-22 through V-27.

Having performed these analyses it occurred to us that there might be some optimal way of combining the two methods to obtain a better estimate of the mean residual for each model. It seemed logical to consider weighted averages of the Method 1 and Method 2 residuals. If r_1 represents the residual from Method 1 and r_2 from Method

2, we can consider all combinations r of the form

$$r = w\, r_1 + (1-w)\, r_2 \quad , \quad 0 \leq w \leq 1.$$

For any value of w we can compute the combined residuals and use these as outcome measures in a one-way analysis of variance with programs as factors. It seems reasonable to try different values of w and select that value w^* which minimizes the within-group mean square. A more detailed rationale is presented in Appendix D. The resulting ANOVA also provides a measure of the statistical significance of model differences. The results of the "Combined" residual analysis are presented in Table V-28. We were particularly interested in comparisons among the means for the PV children as a whole, the NPV children, and the Control children. The results of various t-tests based on the ANOVA are presented in Table V-29.

We were somewhat concerned by the fact that the theory underlying the residual analysis might be less appropriate for children with some prior preschool experience. Since the models vary somewhat in their proportions of children in the four groups for which separate regressions were run, biased comparisons may result. As a check, we computed model means and one-way ANOVA's for the four groups separately. These results are presented in Tables V-30 through V-35.

TABLE V-10

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
PSI

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	14.69	20.79	3.22	2.87	126
Arizona	15.95	20.56	3.24	1.37	184
Bank Street	14.11	16.91	3.08	-.28	225
Oregon	17.06	23.26	3.36	2.84	141
Kansas	14.10	18.86	2.77	1.99	97
High/Scope	16.05	19.81	3.13	.63	169
Florida	13.96	18.52	2.82	1.74	138
EDC	14.88	19.55	2.68	1.99	161
Pittsburgh	13.37	19.55	3.40	2.77	99
REC	12.45	17.76	2.78	2.53	67
Enablers	14.59	19.62	3.09	1.94	181
Control	12.22	15.12	2.51	.39	105
NPV	14.92	19.23	3.01	1.31	609
TOTAL PV	14.83	19.54	3.07	1.66	1588
TOTAL	14.74	19.26	3.03	1.50	2302

TABLE V-11

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
PPV

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	37.76	48.40	6.18	4.46	156
Arizona	35.14	45.53	5.87	4.52	190
Bank Street	27.73	37.46	3.94	5.80	241
Oregon	35.47	46.06	4.69	5.91	131
Kansas	31.17	41.69	4.78	5.74	94
High/Scope	35.55	44.14	5.63	2.96	161
Florida	30.28	41.15	4.63	6.24	141
EDC	29.57	39.75	3.60	6.58	160
Pittsburgh	32.13	45.92	7.40	6.39	109
REC	29.28	43.07	4.08	9.72	69
Enablers	34.80	43.29	5.18	3.32	195
Control	28.50	39.26	4.70	6.07	106
NPV	29.64	41.50	4.76	7.10	592
TOTAL PV	32.70	43.02	5.04	5.27	1647
TOTAL	31.74	42.47	4.96	5.77	2345

TABLE V-12

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
WRTC

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	1.93	4.85	1.39	1.53	123
Arizona	1.84	5.33	1.28	2.21	194
Bank Street	1.90	4.28	1.04	1.34	243
Oregon	3.66	8.04	1.24	3.14	138
Kansas	1.23	6.47	.88	4.36	101
High/Scope	2.08	5.58	1.19	2.31	178
Florida	2.43	5.23	1.01	1.80	103
EDC	2.43	5.91	.78	2.69	169
Pittsburgh	1.18	4.17	1.99	1.42	101
REC	.81	3.48	1.19	1.49	77
Enablers	2.02	5.21	1.21	1.98	193
Control	1.39	2.74	1.09	.27	85
NPV	1.85	4.95	1.11	1.99	606
TOTAL PV	2.03	5.33	1.15	2.16	1620
TOTAL	1.96	5.14	1.14	2.05	2311

TABLE V-13

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
WRTD

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	.87	2.09	.40	.82	126
Arizona	.76	2.24	.35	1.12	195
Bank Street	.65	1.41	.30	.47	243
Oregon	.78	3.69	.26	2.65	143
Kansas	.55	2.64	.25	1.84	101
High/Scope	.64	1.68	.33	.71	183
Florida	.84	1.67	.26	.57	103
EDC	.73	2.18	.22	1.24	169
Pittsburgh	.46	2.07	.48	1.14	101
REC	.23	1.31	.28	.80	77
Enablers	.67	1.74	.34	.74	193
Control	.40	.85	.25	.19	85
NPV	.53	1.72	.29	.90	607
TOTAL PV	.67	2.03	.32	1.05	1634
TOTAL	.63	1.91	.31	.98	2326

TABLE V-14

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
ITPA

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	11.83	15.49	2.40	1.26	72
Arizona	14.65	17.31	2.16	.49	72
Bank Street	9.28	13.31	1.38	2.66	99
Oregon	12.89	16.96	1.43	2.64	54
Kansas	10.36	14.19	1.63	2.20	36
High/Scope	10.97	14.21	2.19	1.04	73
Florida	11.27	16.04	1.74	3.03	48
EDC	12.35	16.92	1.52	3.06	63
Pittsburgh	8.82	15.57	2.93	3.82	51
REC	12.81	14.44	1.58	.046	32
Enablers	12.20	14.13	1.91	.031	112
Control	-----	-----	----	-----	---
NPV	10.93	15.70	1.65	3.12	248
TOTAL PV	11.57	15.18	1.89	1.71	712
TOTAL	11.41	15.32	1.83	2.08	960

TABLE V-15

RESULTS OF RESIDUAL ANALYSIS

METHOD 1
ETS

Program	Fall $\bar{Y}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_1$	Sample Size n
Far West	8.83	13.26	2.34	2.09	58
Arizona	8.15	13.90	2.13	3.63	73
Bank Street	7.67	11.67	1.73	2.27	87
Oregon	11.72	16.32	1.75	2.85	50
Kansas	7.06	13.29	1.84	4.39	35
High/Scope	9.88	12.23	2.28	.07	66
Florida	9.16	12.61	1.85	1.59	43
EDC	11.47	13.88	1.55	.85	84
Pittsburgh	8.28	13.18	2.58	2.32	40
REC	10.47	11.40	2.03	-1.10	30
Enablers	11.47	13.22	2.25	-.50	68
Control	-----	-----	----	-----	---
NPV	9.42	12.61	1.82	1.37	253
TOTAL PV	9.46	13.17	2.02	1.68	594
TOTAL	9.45	13.02	1.97	1.60	847

TABLE V-16

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
PSI

Program	Predicted Fall $\hat{\bar{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n
Far West	14.85	20.59	3.22	2.53	135
Arizona	15.14	20.08	3.24	1.69	197
Bank Street	13.76	16.55	3.08	-.29	242
Oregon	16.24	23.13	3.30	3.58	148
Kansas	12.53	18.72	2.76	3.43	99
High/Scope	14.76	19.46	3.14	1.56	184
Florida	14.07	17.80	2.83	.90	154
EDC	14.96	19.28	2.71	1.62	170
Pittsburgh	13.83	19.34	3.41	2.11	102
REC	12.59	16.99	2.83	1.57	78
Enablers	14.64	19.33	3.10	1.59	191
Control	12.07	13.94	2.51	-.65	124
NPV	14.52	19.08	3.02	1.45	636
TOTAL PV	14.45	19.17	3.07	1.66	1700
TOTAL	14.38	18.89	3.03	1.49	2460

TABLE V-17

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
PPV

Program	Predicted Fall $\bar{\hat{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n
Far West	34.58	48.33	6.32	7.42	131
Arizona	33.48	45.05	5.83	5.74	189
Bank Street	28.77	37.32	4.05	4.50	248
Oregon	31.37	45.98	4.46	10.15	133
Kansas	28.90	41.45	4.78	7.77	98
High/Scope	33.67	43.60	5.58	4.35	178
Florida	30.63	39.58	4.45	4.51	146
EDC	30.85	39.45	3.61	4.98	166
Pittsburgh	32.71	45.26	7.55	5.01	99
REC	28.67	42.38	4.06	9.65	74
Enablers	32.75	43.49	5.02	5.72	184
Control	27.99	37.97	4.75	5.23	117
NPV	31.30	41.81	4.76	5.75	604
TOTAL PV	31.57	42.51	4.99	5.95	1646
TOTAL	31.33	42.11	4.92	5.87	2367

TABLE V-18

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
WRTC

Program	Predicted Fall $\hat{\bar{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n
Far West	1.92	4.75	1.39	1.44	130
Arizona	1.91	5.22	1.30	2.01	205
Bank Street	1.69	4.15	1.03	1.43	254
Oregon	2.94	8.02	1.21	3.87	145
Kansas	1.27	6.34	.87	4.20	103
High/Scope	2.02	5.39	1.18	2.19	190
Florida	1.91	4.99	1.00	2.08	147
EDC	2.26	5.86	.79	2.81	174
Pittsburgh	1.03	4.13	1.57	1.52	102
REC	1.35	3.42	1.21	.86	81
Enablers	2.06	5.10	1.21	1.84	202
Control	.88	2.47	1.04	.56	97
NPV	2.02	4.90	1.10	1.79	660
TOTAL PV	1.91	5.22	1.14	2.17	1733
TOTAL	1.90	5.03	1.13	2.01	2490

TABLE V-19

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
WRTD

Program	Predicted Fall $\hat{\bar{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n'
Far West	.70	2.07	.40	.97	129
Arizona	.68	2.18	.36	1.14	204
Bank Street	.54	1.39	.30	.56	254
Oregon	.69	3.65	.26	2.70	145
Kansas	.44	2.62	.26	1.93	103
High/Scope	.71	1.68	.33	.64	190
Florida	.57	1.61	.27	.77	147
EDC	.63	2.16	.22	1.31	174
Pittsburgh	.52	2.05	.48	1.06	102
REC	.48	132	.29	.55	81
Enablers	.70	1.68	.34	.64	202
Control	.42	.74	.26	.07	97
NPV	.62	1.75	.29	.84	660
TOTAL PV	0.61	1.99	0.32	1.05	1731
TOTAL	.61	1.88	.31	.96	2488

TABLE V-20

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
ITPA

Program	Predicted Fall $\hat{\bar{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n
Far West	11.57	16.37	2.39	2.42	59
Arizona	11.86	16.48	2.22	2.41	73
Bank Street	10.76	13.08	1.40	.93	98
Oregon	12.58	17.04	1.50	2.96	55
Kansas	10.55	14.19	1.63	2.02	36
High/Scope	11.56	13.96	2.17	.23	76
Florida	11.40	14.72	1.78	1.54	57
EDC	12.65	16.68	1.51	3.12	65
Pittsburgh	10.23	15.55	3.01	2.31	44
REC	10.14	13.47	1.64	1.69	36
Enablers	11.68	14.03	1.89	.46	104
Control	-----	-----	----	----	---
NPV	11.48	15.44	1.65	2.32	273
TOTAL PV	11.39	14.95	1.90	1.65	703
TOTAL	11.42	15.09	1.83	1.84	976

TABLE V-21

RESULTS OF RESIDUAL ANALYSIS

METHOD 2
ETS

Program	Predicted Fall $\hat{\bar{Y}}$	Spring $\bar{Y}'$	Expected Increment $\bar{\Delta}$	Residual $\bar{r}_2$	Sample Size n
Far West	8.97	12.88	2.46	1.46	134
Arizona	9.39	13.23	2.28	1.56	202
Bank Street	8.51	11.53	1.81	1.21	234
Oregon	10.37	16.26	1.77	4.11	148
Kansas	7.92	12.98	1.73	3.33	102
High/Scope	9.39	12.12	2.20	.53	184
Florida	8.92	12.14	1.77	1.44	152
EDC	9.69	13.57	1.76	2.12	169
Pittsburgh	7.86	11.86	2.59	1.41	102
REC	8.06	10.94	2.01	.86	77
Enablers	9.63	12.61	2.21	.78	186
Control	----	-----	----	----	---
NPV	9.21	12.31	1.89	1.21	638
TOTAL PV	9.08	12.77	2.05	1.63	1690
TOTAL	9.12	12.65	2.00	1.52	2328

TABLE V-22

RESULTS OF RESIDUAL ANALYSIS

BY SITE

PSI

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	2.66	4.83		
02.09	2.91	2.00		
02.13	2.93	1.91		
03.05			.46	1.69
03.08	1.09	2.20		
03.09	1.59	.17		
03.16	1.49	2.44		
05.01				
05.10	-2.32	.11		
05.11	.43	-.96		
05.12	1.77	-.07		
07.11	2.29	3.71	.74	1.80
07.14	3.43	3.45	-2.02	1.12
07.19				
08.02				
08.04	5.02	5.48	2.75	.97
08.08	-1.51	1.06	1.18	1.33
09.02	.12	.02		
09.04	.75	2.95		
09.06	1.31	1.17	1.81	2.10
09.10				
10.01				
10.02	1.15	.15	2.35	2.95
10.07	.70	1.31	.14	2.61
10.10	3.42	1.02	.76	.19
11.05				
11.06	1.15	1.35	-.04	.38
11.08	2.54	1.81	1.82	.95
12.03	3.13	1.10		
12.04	1.85	4.67		
20.01	2.53	1.57	4.18	2.08
27.01				
27.02	1.66	1.57		
27.03	3.89	.89		
27.04	2.95	3.09		
27.05	.47	.46		
28.01	.56	-2.06		
28.03	.97	.16		
28.02	.04	.20		

RESULTS OF RESIDUAL ANALYSIS

BY SITE

PPV

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	5.86	10.36		
02.09	1.28	6.00		
02.13	6.19	7.16		
03.05			5.83	6.21
03.08	4.74	4.13		
03.09	4.59	4.71		
03.16	4.26	7.89		
05.01				
05.10	4.76	3.69		
05.11	7.87	6.27		
05.12	4.90	3.72		
07.11	6.89	7.66	6.46	7.49
07.14	5.00	12.26	17.28	7.35
07.19				
08.02				
08.04	6.97	9.18	5.39	3.84
08.08	4.40	6.30	3.57	1.78
09.02	3.70	1.95		
09.04	.73	4.44		
09.06	7.18	9.06	8.26	9.59
09.10				
10.01				
10.02	2.46	6.38	1.58	3.46
10.07	5.19	2.65	8.31	9.93
10.10	9.99	5.07	4.72	1.74
11.05				
11.06	6.03	3.54	.96	2.97
11.08	6.97	10.47	4.77	8.81
12.03	7.92	3.58		
12.04	3.42	8.61		
20.01	9.72	9.65	11.49	5.75
27.01				
27.02	2.01	3.89		
27.03	4.92	4.85		
27.04	1.42	4.80		
27.05	7.60	9.08		
28.01	9.74	2.31		
28.03	2.52	6.29		
28.02	4.84	7.44		

TABLE V-24

RESULTS OF RESIDUAL ANALYSIS

BY SITE

WRTC

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	2.00	1.76		
02.09	1.58	2.16		
02.13	1.31	.86		
03.05			1.81	2.03
03.08	2.83	2.73		
03.09	2.57	2.03		
03.16	1.34	1.29		
05.01				
05.10	2.58	3.29		
05.11	.76	.12		
05.12	.35	.41		
07.11	2.76	3.25	2.64	2.39
07.14	3.53	4.46	5.48	5.29
07.19				
08.02				
08.04	5.04	4.25	1.02	.68
08.08	3.64	4.15	3.25	2.93
09.02	.64	.15		
09.04	4.31	4.52		
09.06	.64	.54	.71	.03
09.10				
10.01				
10.02	1.97	2.11	4.67	4.26
10.07	2.68	1.85	3.36	2.38
10.10	1.49	2.30	-.73	-.35
11.05				
11.06	3.08	2.99	1.17	.94
11.08	2.41	2.66	2.14	2.18
12.03	1.04	.72		
12.04	2.37	3.54		
20.01	1.49	.86	.08	-.43
27.01				
27.02	2.01	1.99		
27.03	.65	.17		
27.04	3.22	3.81		
27.05	1.37	.69		
28.01	.55	.35		
28.03	.86	1.25		
28.02	-.20	.38		

RESULTS OF RESIDUAL ANALYSIS

BY SITE

WRTD

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	1.12	1.16		
02.09	.65	.99		
02.13	.82	.88		
03.05			.97	1.08
03.08	1.31	1.25		
03.09	.86	.97		
03.16	1.14	1.20		
05.01				
05.10	.77	1.04		
05.11	.25	.24		
05.12	.30	.26		
07.11	2.84	2.69	.90	.69
07.14	2.47	2.71	1.32	1.20
07.19				
08.02				
08.04	1.86	1.84	.33	.17
08.08	1.81	2.02	1.09	.83
09.02	.27	.03		
09.04	1.14	1.21		
09.06	.57	.45	.46	.46
09.10				
10.01				
10.02	.74	1.09	1.25	1.27
10.07	.16	.61	1.39	1.81
10.10	.53	.68	.70	.44
11.05				
11.06	.97	.97	.78	.73
11.08	1.43	1.57	1.57	1.69
12.03	.96	.82		
12.04	1.58	1.66		
20.01	.80	.55	.50	.36
27.01				
27.02	.92	.73		
27.03	.26	.13		
27.04	1.02	1.16		
27.05	.51	.32		
28.01	.24	-.06		
28.03	-.32	.02		
28.02	.10	.21		

TABLE V-26

RESULTS OF RESIDUAL ANALYSIS

BY SITE

ITPA

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	2.83	3.77		
02.09	5.06	3.85		
02.13	-2.11	1.16		
03.05				
03.08	-2.07	6.42	4.49	1.68
03.09	3.22	1.38		
03.16	.67	-.24		
05.01				
05.10	6.57	2.55		
05.11	-.11	.55		
05.12	1.04	-.44		
07.11	2.78	4.20	4.49	5.91
07.14	2.52	2.00	-2.19	-2.17
07.19				
08.02				
08.04	2.22	2.16	1.91	.57
08.08	2.19	1.86	1.18	1.25
09.02	1.83	1.84		
09.04	-1.05	-.95		
09.06	3.52	-.06		
09.10			7.02	3.48
10.01				
10.02	.93	1.45	3.10	2.89
10.07	4.86	1.10	2.08	1.36
10.10	3.72	1.97	4.18	2.69
11.05				
11.06	4.06	4.91	5.40	4.54
11.08	2.40	1.93	3.15	4.00
12.03	4.56	3.12		
12.04	2.57	.20		
20.01	.05	1.69	3.94	4.08
27.01				
27.02	-.87	.45		
27.03	-2.31	1.08		
27.04	2.06	-.05		
27.05	.19	.76		
28.01				
28.03				
28.02				

RESULTS OF RESIDUAL ANALYSIS

BY SITE

ETS

	PV		NPV	
SITE	METHOD 1	METHOD 2	METHOD 1	METHOD 2
02.02				
02.04	2.40	2.99		
02.09	.34	.54		
02.13	3.06	1.42		
03.05			1.39	.79
03.08	1.67	1.20		
03.09	5.37	1.84		
03.16	3.47	1.66		
05.01				
05.10	5.28	2.33		
05.11	-.35	.12		
05.12	.99	.89		
07.11	3.33	4.04	.00	-.30
07.14	2.56	4.19	.91	.73
07.19				
08.02				
08.04	3.88	4.33	1.58	.78
08.08	4.93	2.20	3.71	.56
09.02	.27	-.04		
09.04	-.19	1.02		
09.06	.28	.43	1.06	.59
09.10				
10.01				
10.02	.43	1.30	-.05	2.08
10.07	4.56	.95	5.31	3.19
10.10	1.39	1.98	1.14	.60
11.05				
11.06	.12	2.39	1.15	2.77
11.08	1.35	1.93	1.14	2.30
12.03	1.91	.58		
12.04	3.27	3.50		
20.01	-1.10	.86	2.36	1.54
27.01				
27.02		2.41		
27.03	-1.40	.06		
27.04	-1.25	.55		
27.05	.78	-.10		
28.01				
28.03				
28.02				

TABLE V-28

RESULTS OF COMBINED RESIDUAL ANALYSIS

Program	PSI	PPV	WRTC	WRTD	ITPA	ETS
Far West	2.72	5.79	1.52	.90	1.83	1.91
Arizona	1.61	5.23	2.16	1.16	1.81	2.92
Bank Street	-.23	5.34	1.39	.50	1.80	1.82
Oregon	3.08	7.52	3.35	2.67	3.10	3.21
Kansas	2.60	6.55	4.35	1.88	2.11	3.79
High Scope	1.11	3.57	2.33	.71	.68	.22
Florida	1.54	5.77	1.93	.67	2.52	1.78
EDC	1.89	6.18	2.73	1.27	3.11	1.65
Pittsburgh	2.57	5.80	1.46	1.11	3.10	2.52
REC	2.48	9.99	1.33	.70	1.37	.01
Enablers	1.86	4.27	1.94	.71	.42	-.19
Control	.23	6.01	.43	.17	....	....
NPV	1.40	6.39	1.93	.86	2.79	1.37
Total PV	1.72	5.69	2.19	1.06	1.82	1.75
Total	1.57	5.89	2.06	.98	2.08	1.63
Within Mean-square	14.68	64.36	10.04	1.47	25.64	10.18
F	9.91	4.01	10.91	37.57	2.42	7.32
Significance Level	<.001	<.001	<.001	<.001	.005	<.001
w*	.6	.6	.7	.6	.5	.5
n	2301	2200	2311	2309	901	844

TABLE V-29

TESTS FOR SIGNIFICANT DIFFERENCES AMONG PV, NPV, AND CONTROL RESIDUAL MEANS*

	PV-NPV		PV-Control		NPV-control	
	t	Significance	t	Significance	t	Significance
PSI	1.75	Above .05	5.66	<.001	2.79	<.01
PPV	-1.77	Above .05	-.40	Above .05	.43	Above .05
WRTC	1.72	Above .05	4.26	<.001	3.63	<.001
WRTD	3.45	<.001	6.13	<.001	4.60	<.001
ITPA	-3.34	<.001	.....	.....	.....	.....
ETS	2.11	<.005	.....	.....	.....	.....

* The number of degrees of freedom for each test is so large that we refer to normal distribution tables for the significance level.

TABLE V-30

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

PSI

PROGRAM	FIRST LANGUAGE NOT ENGLISH		NON HS PRIOR PS		NO PRIOR PS		PRIOR HS	
Far West	2.11		.93		2.83	*		
Arizona	.73		-.81		2.22		.65	
Bank Street	----		-3.54	*	1.36	***	-.74	**
Oregon	3.49	**	-.55		3.66	***	.86	
Kansas	----		----		3.02	**	-2.77	*
High/Scope	1.57		-1.66		1.27		1.34	
Florida	2.85		.02		1.60		.20	
EDC	----		2.91		1.55		2.18	**
Pittsburgh	----		----		3.05	*	1.49	
REC	4.81	**	----		2.17		----	
Enablers	1.44		2.85		1.79		.57	
Control	.91		-.58		.50		----	
NPV	-.14		.33		1.87		.68	
TOTAL	1.92		.17		1.90		.64	
N	161		133		1577		430	

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance is for contrast with NPV

TABLE V- 31

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

PPV

PROGRAM	FIRST LANGUAGE NOT ENGLISH	NON HS PRIOR PS	NO PRIOR PS	PRIOR HS
Far West	8.17	-2.21	6.03	3.88 *
Arizona	4.65	1.07	5.64	5.05
Bank Street	----	1.38	6.32	4.88
Oregon	8.71 *	-13.09	7.45	5.05
Kansas	----	----	6.82	3.73
High/Scope	14.50	-.68	3.14 ***	4.01
Florida	16.16	1.46	5.47	2.06 *
EDC	-3.30 *	15.78 **	5.40	5.08
Pittsburgh	----	2.97	6.39 *	4.84
REC	17.10	5.56	8.90 *	----
Enablers	6.30	1.11	4.81	-.16
Control	3.45 *	3.40	7.40	----
NPV	13.30	.68	6.13	5.88
TOTAL	10.77	1.91	5.89	5.36
N	150	133	1518	399

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance is for contrast with NPV

TABLE V-32

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

WRTC

PROGRAM	FIRST LANGUAGE NOT ENGLISH	NON HS PRIOR PS	NO PRIOR PS	PRIOR HS
Far West	2.60	.19	1.52	2.04
Arizona	2.26	1.19	2.02	2.69
Bank Street	----	.48	1.25	1.66
Oregon	3.77	-1.38	4.53 ***	2.85
Kansas	----	----	4.53 ***	2.29
High/Scope	.93	-.50	2.54 **	3.90 *
Florida	2.29 *	2.21	1.82	2.43
EDC	-2.51 *	6.01 *	3.21 ***	2.18
Pittsburgh	----	1.03	1.64	.29
REC	2.49 *	.97	1.14	----
Enablers	1.88 *	3.05 *	1.78	1.79
Control	3.88	-.56	.35 **	----
NPV	4.82	.53	1.68	1.71
TOTAL	3.53	.99	1.99	2.02
N	173	123	1576	439

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance as for contrast with NPV

TABLE V-33

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

WRTD

PROGRAM	FIRST LANGUAGE NOT ENGLISH		NON HS PRIOR PS		NO PRIOR PS		PRIOR HS	
Far West	2.33		.35		.90		1.01	
Arizona	.57		.74		1.17	**	1.25	
Bank Street	----		.51		.41	**	.63	*
Oregon	2.78	***	.42		2.77	***	2.20	***
Kansas	----		----		1.93	***	1.32	
High/Scope	.41	*	.35		.70		1.54	
Florida	.02	**	.08		.83		.05	*
EDC	.54		1.71		1.02		1.52	***
Pittsburgh	----		.44		1.21	**	1.08	
REC	1.76		.36		.52		----	
Enablers	.56	*	.78		.74		-.08	*
Control	.85		.11		.12	***	----	
NPV	1.29		.84		.77		.94	
TOTAL	1.60		.58		.92		1.06	
N	173		123		1576		437	

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance is for contrast with NPV

TABLE V-34

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

ITPA

PROGRAM	FIRST LANGUAGE NOT ENGLISH	NON HS PRIOR PS	NO PRIOR PS	PRIOR HS
Far West	----	-2.28	1.96	3.58
Arizona	-2.52	3.10	1.64	2.25
Bank Street	----	3.02	1.49 *	2.13
Oregon	2.43 *	----	4.21	2.76
Kansas	----	----	2.10	2.17
High/Scope	2.48	.98	.73 **	-3.24 **
Florida	3.94	----	.73 **	-3.24 **
EDC	----	3.32	3.57	2.63
Pittsburgh	----	-2.56	3.87	5.72
REC	.39	1.17	1.80	----
Enablers	-4.21	.72	.69 ***	-.72
Control	----	----	----	----
NPV	-1.41	4.65	3.05	3.50
TOTAL	.57	.84	2.17	2.70
N	68	47	611	175

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance is for contrast with NPV

TABLE V-3 5

COMBINED RESIDUAL RESULTS FOR FOUR GROUPS ON WHICH REGRESSIONS ARE BASED

ETS

PROGRAM	FIRST LANGUAGE NOT ENGLISH	NON HS PRIOR PS	NO PRIOR PS	PRIOR HS
Far West	----	1.76	1.79	3.15
Arizona	-2.03	2.43	3.53 ***	1.73
Bank Street	----	-2.67	1.94	2.03
Oregon	3.11 **	----	3.05	3.89 *
Kansas	----	----	3.85 ***	3.40
High/Scope	-1.58	.48	.30 *	1.48
Florida	3.53	----	1.44	3.09
EDC	-3.35	.68	1.77	1.74
Pittsburgh	----	3.64	2.42	.75
REC	1.13	.00	-.29 *	----
Enablers	-.80	-.95	.03 **	-1.53
Control	----	----	----	----
NPV	.49	1.00	1.55	1.13
TOTAL	137	.76	1.67	1.79
N	67	43	565	169

* Indicates significance p .05
**Indicates significance p .01
***Indicates significance p .001

Significance is for contrast with NPV

Table V-36

Total Sample Size for Residual Analysis

Test	Valid Fall	Valid Spring	Method 1	Method 2	Combined
PSI	3175	2753	2302	2460	2301
PPV	3217	2660	2343	2367	2200
WRTC	3204	2792	2311	2490	2311
WCID	3204	2792	2326	2488	2309
ITPA	1210	1077	960	976	901
ETS	1135	2606	847	2328	844

One final remark concerns the samples on which the various analyses were based. The data requirements for Methods 1 and 2 differ slightly* and those for the Combined analysis are the most stringent. We would be somewhat concerned if the data collection did not allow computation of residuals for very many children. The sample sizes summarized in Table V-36 would seem to reveal no cause for concern.

Results by Test

In this section we present a summary of the results of the residual analysis for each test. These summaries are based primarily on the combined residuals.

Preschool Inventory

The average expected increment for all PV children was 3.07, for NPV children 3.01, and for Control children 2.51. The average residuals were 1.72 for PV, 1.40 for NPV and .23 for Control. Thus the growth rate for Head Start (PV and NPV) children over the period between tests increased by roughly 50%, while the rate for Control children increased negligibly. Putting it another

---

*The main difference is that Method 1 requires fall scores, while Method 2 does not.

way, we would have expected Head Start children to gain about .5 of a standard deviation (see Table II-25) without any preschool; with Head Start they gained about .75 of a standard deviation. The difference between PV and NPV was not significant at the .05 level. The overall F-test for program differences was highly significant. The mean residuals for Oregon (3.08), Far West (2.72), Kansas (2.60), and Pittsburgh (2.57), and REC (2.48) were high. Bank Street (-.23) and the Controls (.23) were low. Most of these effects seem fairly consistent across sites, but Kansas is rather puzzling. The Portageville site showed the highest mean residuals for the two methods (5.02, 5.48), while Mounds did very poorly (-1.51, 1.06). On the whole, children without prior pre-school experience had larger residuals than those with prior pre-school.

Peabody Picture Vocabulary Test

The average expected increment for all PV children was 5.04, for NPV children 4.76, and for Controls 4.70. The average residuals were 5.69 for PV, 6.39 for NPV, and 6.01 for Control Thus the growth rate for all three groups more than doubled. In terms of standard deviations, the expected growth was about .35 and the actual growth about .8 for all three groups. The differences among the three groups were not significant at the .05 level.

The overall F-test for program differences was significant ($p < .001$). The mean residuals for REC (9.99) and Oregon (7.52) were high, and those of High/Scope (3.57) and the Enablers (4.27) low. There appear, however, to be large variations among sites within models. The 150 children whose first language was not English and who were in a Head Start program had an average residual of 11.58, while the 7 in the Control group averaged only 3.45.

WRAT Copying Marks.

The average expected increment for all PV children was 1.15, for NPV children 1.11, and for Control children 1.09. The average residuals were 2.19, 1.93, and .43 respectively. Thus, while the growth rates for PV and NPV children nearly tripled, we must remember that since the mean fall score was only 2.03 on a test with a maximum of 18, the spring mean of 5.33 was still rather low. In terms of standard deviations, the expected gain was about .4 and the actual gain about .75. The PV and NPV means did not differ significantly, but both were significantly ($p < .001$) above the Control mean. The overall F-test for program difference was highly significant. The mean residual for Kansas (4.35) and Oregon (3.35) stood out on the high side, while the Controls (.43) were by far the lowest. These results were consistent across all sites within these models, although other models (most notably Bank Street

and High/Scope) showed large site to site variations.

WRAT Reading Numbers

The average expected increment for all PV children was .32, for NPV children .29 and for Control children .25. The average residuals were 1.06, .86, and .17 respectively. Thus the growth rates for PV and NPV quadrupled. Of course the projected growth rate was rather small. In terms of standard deviations, the expected gain was about .3 and the actual gain over 1.0. The PV mean was significantly ($p < .001$) higher than the NPV mean, but this was probably attributable to two outstanding PV models (Oregon and Kansas). Both PV and NPV were significantly ($p < .001$) above the Controls. The overall F-test for program differences was highly significant. Oregon (2.67) and Kansas (1.88) clearly stood out on the high side. The Controls (.17) and Bank Street (.50) were low. Results seem quite consistent across sites.

ITPA Verbal Expression.

The average expected increment for all PV children was 1.89 and for NPV children 1.65. Controls were not given the ITPA. The average residuals were 1.82 for PV and 2.79 for NPV. Thus the growth rate for Head Start children more than doubled. In terms of standard deviations, the expected

gain was about .35 and the actual gain about .7 for PV and .85 for NPV. The difference between PV and NPV was significant ($p < .001$) and somewhat perplexing. The overall F-test for program differences was significant ($p=.005$). EDC (3.11), Oregon (3.10), and Pittsburgh (3.10) had the highest mean residuals, while Enablers (.42) and High/Scope (.68) were lowest. The results seem fairly consistent across sites.

ETS Enumeration.

The average expected increment for all PV children was 2.02 and for NPV children 1.82. Controls were not given the ETS. The average residuals were 1.75 for PV and 1.37 for NPV. Thus the growth rate increased by about 75%. In terms of standard deviations, the expected gain was about .4 and the actual gain about .7. The difference between PV and NPV was barely significant at the .05 level. The overall F-test for model differences was significant ($p < .001$). Kansas (3.79), Oregon (3.21), Arizona (2.92), and Pittsburgh (2.52) were high. Enablers (-.19), REC (.01), and High/Scope (.22) were low. Effects seem fairly consistent across sites.

Summary of Residual Analysis Results

As we did for the ranking analysis, we shall summarize in this section the evidence provided by the residual analysis bearing on our three major questions:

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged pre-schoolers acquire cognitive skills?

Our evidence here is clear and direct. Children in Head Start programs apparently gained substantially more on each test than they would have without the programs. For all tests except the PPV, the Control children showed small average residual gains. Since there is bound to be some test sensitization or slight imperfection in our regression models, these results are quite consistent with what we might expect, and further evidence that the increase in growth rates for Head Start children are genuine program effects and not mathematical artifacts. We do not, however, understand why the Control children on the PPV showed an increase comparable to that of the Head Start children. From this and the ranking analysis it seems clear that the Controls performed about as well as the Head Start children on the PPV. The question is whether Head Start programs really have no effect, so that the residuals are some

kind of artifact, or whether for some reason both Head Start and Control children do better than we would expect on the basis of natural maturation. We shall explore this perplexing issue somewhat further in Appendix F. A particularly interesting finding about the PPV was the tremendous increase in scores for children from Spanish speaking families. Head Start may be functioning for these children as an effective early exposure to the English language. This effect seems to hold only for receptive and not active vocabulary, as the residuals of Spanish speaking children on the ITPA were rather low.

2. Are the Planned Variation models simply by virtue of sponsorship, more effective than ordinary, non-sponsored Head Start programs?

On three of the six tests (PSI, PPV, WRTC) the difference between PV and NPV mean residuals fails to reach significance at the .05 level. The difference for the ETS is barely significant at the .05 level. For the WRTD the PV mean is significantly ($p < .001$) higher, and for the ITPA the NPV mean is significantly ($p < .001$) higher. The difference for the WRTD can be primarily attributed to the stand-out performance of two models (Oregon and Kansas). The ITPA difference seems attributable primarily

to two models (High/Scope and Enablers) which stood out negatively. Our impression is that, on the whole, the performance of PV and NPV programs is quite comparable.

3. Are some PV models particularly effective at imparting certain skills?

Table V-37 presents a summary based on the discussion in the previous section. Of the 22 "effects" noted, 16 occur in 3 of the 6 tests (PSI, ITPA, ETS). The only test with fewer than 3 effects is the PPV, which has none. As in the ranking analysis, it appears that the PPV is not particularly sensitive to program differences. In terms of models, it is interesting to note that of the 15 positive effects, 11 are for the "academic" models (Oregon, Kansas, and Pittsburgh). Moreover, all 4 ++'s are for these models. Thus, as in the ranking analysis, the evidence suggests that the academic models may be generally more effective in transmitting academic skills.

Table V-37

Summary of Relative Model Effectiveness

Based on Residual Analysis*

++ Indicates model appears to be highly effective.
+ Indicates evidence for above average effectiveness.
- Indicates evidence for below average effectiveness.
-- Indicates model appears to be highly ineffective.

Model	PSI	PPV	WRTC	WRTD	ITPA	ETS
Far West	+					
Arizona						+
Bank Street	-			-		
Oregon	+		+	++	+	++
Kansas			++	+		++
High/Scope	-				-	-
Florida						
EDC			+		+	
Pittsburgh	+				+	+
Enablers					-	-

*REC not included because with only one site we felt it unfair to draw any conclusions.

Chapter VI

ANALYSIS OF COVARIANCE

Theory of the Analysis of Covariance

In this section we discuss the theory underlying what is currently perhaps the most popular technique for comparing the effects of educational programs in quasi-experimental situations, the analysis of covariance (ANCOVA). We begin with the more general problem of constructing linear models to describe the relationship between post-test scores and variables which can be measured prior to program exposure, including the pre-test score. Let us, for convenience, refer to all such preprogram variables as covariates. Suppose for each program we could fit a regression model which would allow perfect prediction of a child's post-test score on the basis of the available covariates. Then, in theory at least, we could compare the effects of different programs on children with any specified set of background characteristics. In practice, we can predict with only limited accuracy. Moreover, there would be a virtually infinite number of possible comparisons, one for each possible combination of child background characteristics. To summarize all this information in a meaningful way would be quite difficult.

Suppose, however, it turns out that a simpler mathematical model is adequate. Suppose that the post-test score Y' for any child can be predicted by some function (say F) of his covariate values (say V) plus an additional effect attributable to the particular program experienced. Thus, for individual i in program j we would have

$$Y_{ij}' = \alpha_j + F(V_{ij}) + e_{ij} \qquad (6.1)$$

where α_j represents a program effect and e_{ij} random error uncorrelated with the covariates. If F is a linear function of the covariates, it can be separated into a part involving the pre-test Y and a remainder, say M, involving the other covariates. Thus we have

$$Y_{ij}' = \alpha_j + \beta Y_{ij} + M_{ij} + e_{ij} \qquad (6.2)$$

If this model is appropriate, it provides straightforward treatment comparisons. We simply fit the model and compare the values of the effects α_j estimated for the various programs. Each α_j may be considered as the expected value for individuals in program j after "adjustment" for the pre-test and other covariates.

Note that the assumption that the function F (i.e. the set of regression coefficients for the covariates) is the same for all program groups is absolutely essential in

allowing straightforward program comparisons. To clarify this point, suppose for the moment we have only one covariate, the pre-test, and are comparing two programs. Then over the range of possible pre-test scores, there are essentially three possibilities, as illustrated in Figure VI-1. In situation (a) we cannot say which program is better. For children with low pre-tests program 2 is better. For those with high pre-tests 1 is better. In (b) we can say that 2 is generally better, but we have no simple measure of its superiority, since the difference between the program effects varies with pre-test score. Only in situation (c) can we say simply that program 2 is on the average $\alpha_2 - \alpha_1$ points better.

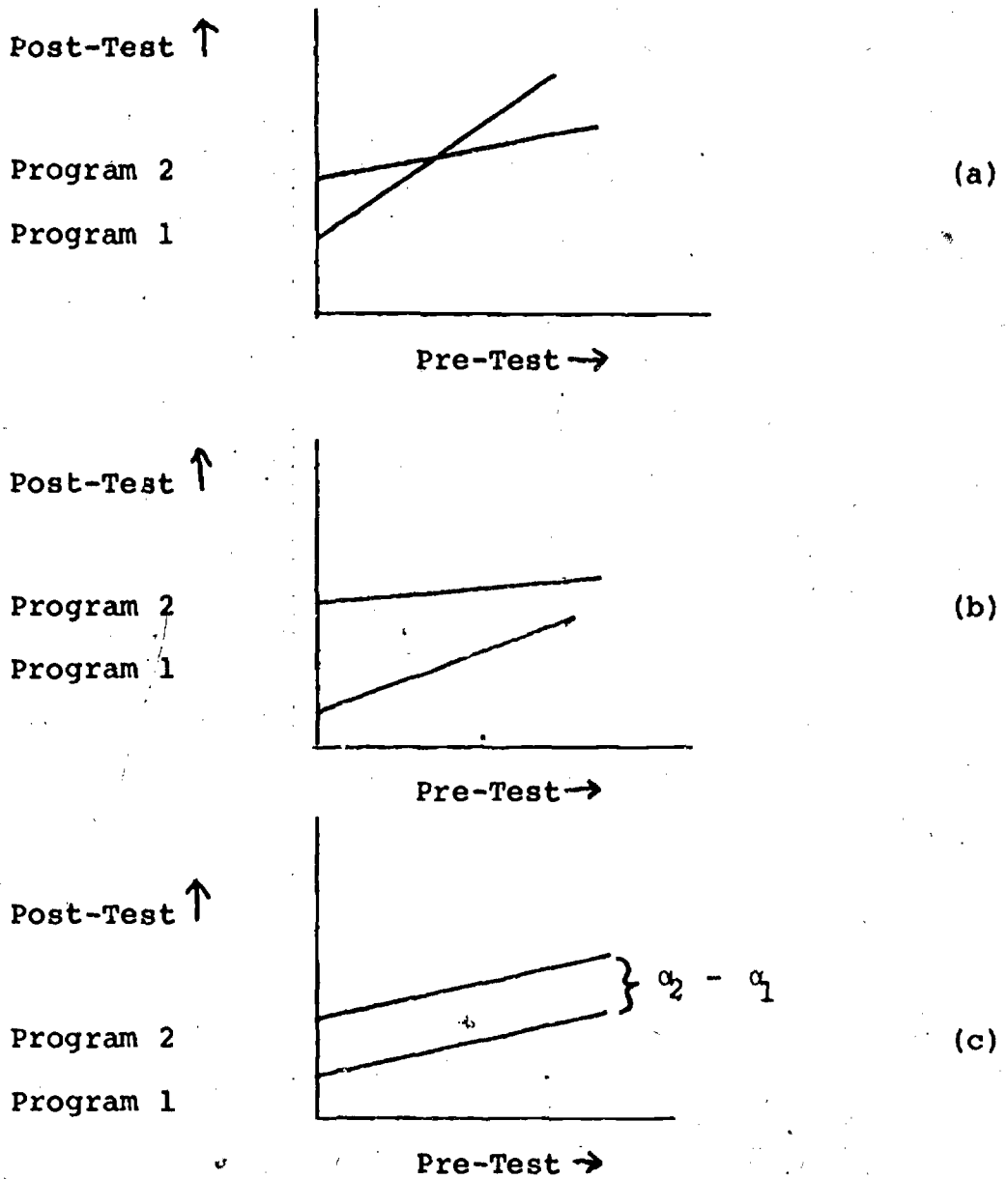
With more than one covariate the situation becomes more complex. The assumption that F is the same for all program groups becomes more difficult to check.

If the ANCOVA model is basically correct, the precision of group comparisons based on it depends on how much of the <u>within-group</u> variance can be explained by the covariates. As explained in Chapter V, the reliability ρ of the post-test is an upper bound on the proportion of variance explainable. Thus, our goal is to build models with R^2 as close to ρ as possible.

A rather thorny issue which is the focus of much current concern involves effects on the ANCOVA of unrelia-

Figure VI-1

Possible Relationships between Pre- and Post-Tests for Two Programs



bility in the covariates. If the model described by equation (6.2) is correct, there is no theoretical problem. Some researchers feel, however, that a linear model stated in terms of "true scores" rather than observed scores is more appropriate. Suppose that

$$Y_{ij} = T_{ij} + u_{ij}$$
$$Y_{ij}' = T_{ij}' + u_{ij}'$$

where T_{ij} and T_{ij}' are the true scores corresponding to Y_{ij} and Y_{ij}' respectively, and u_{ij} and u_{ij}' are random errors of measurement with mean 0 and uncorrelated with the true scores. For simplicity, suppose the pre-test is the only covariate. Then we can consider a mathematical model of the form

$$T_{ij}' = a_j + bT_{ij} \qquad (6.3)$$

which implies

$$Y_{ij}' = a_j + bY_{ij} + (u_{ij}' - bu_{ij}). \qquad (6.4)$$

In this model, the error term is correlated with the pre-test, a violation of the usual assumptions on the basis of which linear models are fit. If we try to estimate the a_j's and b using the usual least-squares procedure, we obtain biased estimates.

As we mentioned in Chapter III, several suggestions for "correcting" the ANCOVA for unreliability of the covariates have recently appeared (e.g. Lord, 1960; Porter, 1971). These corrections seem to us rather shaky for use at the present time in educational evaluations. For one thing, they depend heavily on the rather stringent classical assumptions about errors of measurement. Second, they make the somewhat arbitrary assumption that a linear model holds in terms of true scores but not observed scores. Third, they require a fairly precise knowledge of the covariate reliabilities, and finally, from a practical standpoint they are difficult to implement, particularly in the multiple covariate situation. It seems to us more fruitful to try to explain as much variance as possible using a limited number of reasonably reliable covariates. In concluding this section, let us summarize the main assumptions on which the use of ANCOVA is based.

Assumption 1: A linear model adequately represents the relationship between post-test and covariates for each program group.

Assumption 2: The coefficients of the covariates in the different program groups are approximately equal.

Assumption 3: The covariates have high enough reliability to avoid seriously biased results.

Implementation of the Analysis of Covariance

Several exploratory regression analyses were carried out with the spring test score (post-test) as the dependent variable and a variety of covariates, including the pre-test score and fall scores for other tests. For practical reasons, we limited these preliminary investigations to three tests (PSI, PPV, WRTC).

Although we felt that interpretation would be easier if we could avoid interaction variables, it became clear that interactions involving child ethnicity could not be ignored. Since we wished to avoid the introduction of many two-way and even higher order interactions, we felt it would be simpler to divide the sample into Blacks, Whites, and Spanish Americans (Mexican Americans and Puerto Ricans) and to build separate regression models. Within these ethnic groups, separate regression models were fitted for each of the thirteen program groups, (11 PV models, NPV, Controls) with a sufficient number of children*. One of the most promising sets of models is displayed as an example in Tables VI-1 through VI-9.

---

*One model (Pittsburgh) contained no Blacks, and 5 (Far West, Arizona, Kansa, EDC, Pittsburgh) had not enough Spanish Americans to carry out the analysis.

Table VI-1

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Whites

ODEL	C	PSI FALL	FPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	N	R^2
West	13.6	.545	.064	.057	.166	-.001	-.201	-.537	.103	.579	137	.572
zona	5.20	.425	.117	.091	.135	-.009	.343	-.145	-.648	1.00	125	.557
k Street	14.1	.648	.108	-.035	-.301	-.059	-.689	-.142	.233	.215	41	.802
gon												
sas	-9.18	.503	.046	.670	-.056	-.007	-4.34	-1.09	6.56	-1.12	41	.713
h/Scope	5.29	.552	.058	.095	-.170	-.026	-1.50	.058	-.468	-.895	107	.661
rida	-40.56	.726	-.213	.484	.701	-.034	-.094	1.23	.180	.929	39	.662
	46.17	.779	.157	-.271	.129	-.007	-3.49	-1.63	.764	-.272	39	.744
tsburgh	6.85	.493	.152	.011	-.100	-.045	0.0	.144	-.522	.067	119	.667
blers	8.67	.586	.140	.033	-.259	.014	.898	-.107	1.83	-.276	103	.796
trol	-6.53	.684	.086	.130	.234	0.0	0.0	0.0	2.19	-.975	54	.774
	-1.74	.481	.121	.101	.182	-.030	.976	.172	.371	-.128	205	.648

Table VI-2

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Blacks

ODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	N	R^2
West												
zona	4.63	.746	.011	.120	.003	-.025	-.211	-.243	-.453	-.071	74	.681
k Street	1.29	.442	.116	.094	.211	-.018	-.182	-.110	.114	.072	216	.732
gon	27.50	.442	.145	-.087	.070	.002	2.75	-.937	-1.23	.604	74	.585
sas	-18.01	.857	.112	.449	.575	-.100	-1.85	-.457	-8.00	.120	62	.695
h/Scope	-3.85	.087	.274	.167	.163	-.035	-.775	.097	.149	.674	54	.433
rida	-28.70	.605	.154	-.033	.198	-.060	.709	2.16	-1.81	.084	103	.575
	10.72	.446	.066	.211	.126	-.016	-2.50	-.825	-.806	.209	129	.595
tsburgh												
blers	5.15	.342	.104	.112	.131	-.031	-2.18	-.230	1.85	.860	74	.491
trol	-5.60	.650	.192	.090	.183	0.0	0.0	0.0	.180	-.187	51	.740
	-1.93	.514	.093	.099	.180	-.031	-.378	.186	.151	.138	383	.615

Table VI-3

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Spanish-American

ODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	N	R^2
West												
ona												
Street												
on	-.578	.413	.042	.222	.050	.045	0.0	-.008	.315	.299	76	.684
as												
/Scope	17.28	.489	.135	-.182	.125	.039	0.0	-.080	-.370	-.113	33	.342
ida	37.72	.339	.229	.049	.297	-.046	0.0	-2.29	0.0	3.40	30	.496
sburgh												
	53.89	.494	.154	-.204	-.142	-.020	0.0	-2.09	-.536	.564	44	.567
lers	-8.53	.436	.094	.122	-.057	-.020	.811	.757	.240	1.45	47	.681
rol	-7.90	.641	.116	.146	.392	0.0	0.0	0.0	1.53	-2.18	21	.864
	8.18	.486	.098	.144	.128	-.063	-.455	-.463	1.062	-.512	111	.652

Table VI-4

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Whites

ODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	N	R^2
West	14.6	.266	.443	-.031	.157	.032	-1.00	.768	1.12	1.37	132	.574
ona	-24.4	.493	.430	.140	.292	.003	-1.08	2.48	-1.72	-3.07	121	.540
Street	30.7	.304	.526	.018	.009	-.092	-2.20	-.453	-2.60	-.042	41	.610
on												
as	15.4	.212	.479	.542	-.312	.084	-5.42	-1.21	7.87	-1.88	42	.650
/Scope	24.5	.523	.496	.072	-.755	.024	-3.18	-.046	-3.76	-3.02	102	.720
ida	79.74	-.035	.799	.112	-.652	-.127	-.547	-3.77	-4.03	-3.47	39	.811
	76.4	.580	.628	-.253	.447	-.024	-9.77	-3.01	1.02	1.97	39	.746
sburgh	33.7	.207	.496	.037	-.093	-.002	0.0	-.439	-1.07	-.193	105	.618
lers	36.8	.580	.437	-.041	-.154	.017	.552	-.708	-2.69	-2.22	100	.693
rol	7.35	.984	.382	.279	-.310	0.0	0.0	0.0	-1.61	-5.25	48	.754
	1.69	.295	.443	.201	.657	-.050	-1.43	.470	-.778	-.694	204	.599

Table VI-5

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Blacks

-DEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	N	R^2
Wast												
=ona	47.4	.833	.599	.426	.883	-.153	.297	1.83	-5.29	.683	74	.647
= Street	6.28	.270	.585	.070	.491	-.084	1.72	.160	-.126	-.878	222	.660
=on	25.6	.702	.297	-.234	.355	-.057	-2.54	.860	-.102	-2.72	61	.714
=as	-1.68	.591	.335	.324	.711	-.388	-2.64	.374	.426	-1.19	61	.697
■/Scope	38.1	-.327	.952	.316	.440	.046	-6.28	-3.22	1.94	2.21	53	.751
=ida	-.399	.761	.425	.176	.241	-.107	2.95	.452	-6.47	-4.02	93	.558
	28.1	.334	.620	.233	-.018	-.025	-5.51	-1.37	-.862	-2.07	126	.664
=sburgh												
=lers	74.7	-.010	.779	-.151	.337	-.090	2.82	-3.53	-2.75	1.15	73	.649
=rol	3.05	.473	.571	.148	.376	0.0	0.0	0.0	3.25	-1.80	52	.603
	5.96	.362	.469	.071	.374	-.120	-3.06	.636	1.06	-1.75	365	.620

Table VI-6

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST AND OTHER BACKGROUND VARIABLES

Spanish Americans

MODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX		N	R^2
Far West													
Arizona													
Bank Street													
Oregon	-11.2	.051	.530	.554	.469	.091	0.0	-.329	2.42	.274		74	.632
Kansas													
High/Scope	42.8	.406	.721	-.411	-.42	.101	0.0	.205	-2.26	-2.11		32	.417
Florida	2.20	.737	.901	.063	.110	.256	0.0	14.03	0.0	-5.89		29	.737
EDC	52.2	.139	.697	-.498	.226	-.053	0.0	-2.52	.840	-3.23		41	.624
Pittsburgh													
REC													
Enablers	2.34	.525	.524	.159	1.71	-.040	-3.35	-.541	-1.89	4.22		41	.892
Control	44.9	-.404	.968	-.348	-.924	0.0	0.0	0.0	13.5	-12.9		19	.914
NPV	27.5	.521	.362	-.079	1.01	-.177	-.410	-.150	-6.40	-2.16		111	.481

Table VI-7

REGRESSION MODELS RELATING WRTC POST-TEST TO

PRE-TEST AND OTHER BACKGROUND VARIABLES

Whites

MODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	WRTC FALL	N	R^2
Far West	-14.5	.115	.019	.069	.296	-.012	1.40	.526	.029	1.20	.541	134	.324
Arizona	12.8	.171	.021	.018	.020	.004	-.505	-.856	.090	.622	.495	129	.520
Bank Street	-6.80	.137	.077	.132	.853	-.016	-3.05	-.648	.073	.953	-.002	43	.655
Oregon	-1.89	-.029	.079	.193	-.231	.199	.479	-.517	-9.55	3.14	.710	42	.560
Kansas													
High/Scope	-16.9	.144	.011	.287	-.021	-.002	-.944	.196	-1.03	-.889	.451	110	.660
Florida	-58.3	.241	-.162	.431	-.037	-.078	1.66	2.31	-1.21	-1.22	.986	40	.747
EDC	13.1	.569	.011	.005	.285	.023	-3.82	-1.19	-1.29	2.03	-.062	38	.753
Pittsburgh	-12.03	.089	.058	.104	.028	-.018	0.0	.421	-.534	.513	.597	118	.479
REC													
Enablers	-3.30	.169	.023	.043	-.076	.029	.872	.059	-.94	.319	.817	106	.721
Control	-3.85	.115	.068	.028	.047	0.0	0.0	0.0	.129	.851	.208	40	.701
NPV	-15.6	.193	.048	.140	.149	.032	.028	.279	.310	1.01	.573	213	.533

Table VI-8

REGRESSION MODELS RELATING WRTC POST-TEST TO

PRE-TEST AND OTHER BACKGROUND VARIABLES

Blacks

MODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX		N	R^2
Far West													
Arizona	2.17	.046	.035	.085	-.084	-.003	-.530	-.266	.025	.008	.616	77	.342
Bank Street	14.1	.169	-.006	.120	.141	.003	.565	.372	-.868	.601	.605	228	.735
Oregon	7.94	-.013	.120	-.200	-.057	-.045	-1.34	.564	.856	.454	.537	71	.367
Kansas	-20.1	-.018	.118	.195	.155	-.040	1.47	.822	-1.75	-.381	.658	66	.519
High/Scope	-5.9	.207	.035	-.070	.009	.010	-.470	.581	-.780	.137	1.09	57	.623
Florida	-11.6	.258	.016	.086	.210	-.016	-.138	.197	-1.06	.848	.560	99	.592
EDC	3.65	.061	.031	.090	.027	.003	.788	-.406	-1.97	.768	.392	134	.302
Pittsburgh REC													
Enablers	-6.35	.056	.067	-.015	.154	-.005	-.510	.356	-1.01	.412	.842	78	.566
Control	-1.31	.060	-.023	.002	.174	0.0	0.0	0.0	.156	-.076	.689	41	.639
NPV	-3.97	.050	.048	.076	-.033	-.032	.003	.010	.091	.735	.847	396	.549

Table VI-9

REGRESSION MODELS RELATING WRTC POST-TEST TO

PRE-TEST AND OTHER BACKGROUND VARIABLES

Spanish Americans

MODEL	C	PSI FALL	PPV FALL	AGE	MOM'S EDUC.	DAYS ABSENT	TEACHER BLACK	TEACHER EDUC.	PS	SEX	WRTC FALL	N	R^2
Far West													
Arizona													
Bank Street													
Oregon	18.4	.133	-.085	-.101	.235	-.089	0.0	-.390	-2.51	-.947	.647	77	.457
Kansas													
High/Scope	13.4	.217	-.117	-.062	-.038	.088	0.0	-.450	-1.30	-.006	.908	34	.650
Florida	19.5	-.068	.197	-.036	-.134	.032	0.0	-.965	0.0	.970	.255	27	.517
EDC	-12.6	.297	.045	-.146	-.146	.034	0.0	1.16	.631	-.047	.925	45	.348
Pittsburgh													
REC													
Enablers	-5.85	.079	.052	-.033	-.065	-.007	.736	.653	.298	1.01	.698	53	.420
Control	-10.0	-.227	.093	.207	.311	0.0	0.0	0.0	.980	-3.67	.720	18	.747
NPV	-17.4	.249	-.026	.438	-.174	-.036	.778	-.240	-.292	-.086	.360	117	.651

Looking over the various equations, we noticed that a few variables tended to predominate in importance. These were the pre-test score, fall scores for some other tests, and age. Using only these variables, we fit the equations displayed in Table VI-10 through VI-18. Note that generally there is only a small loss in R^2 compared with the more complex models described above. Moreover, except possibly for age, the coefficients seemed fairly constant across the different models within any ethnic group. We decided to perform ANCOVA's for each of the 8 tests for Blacks and Whites. We felt there were not enough models with a substantial number of Spanish Americans to justify running the ANCOVA for them. We also decided to eliminate age as a covariate. We ended up using as covariates the pre-test score and fall scores for the PSI and PPV. We considered the possibility of carrying out a formal statistical test of the assumption that the regression coefficients for different groups were the same. An attempt to do this would, however, have involved us in computational problems beyond the capabilities of the computer programs available to us.

The ANCOVA was carried out using a standard multiple regression program.. For Whites there were 13 treatment groups (11 models, NPV, Controls) and for Blacks, 12

Table VI-10

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Whites

Model	c	PSI Fall	PPV Fall	Age	n	R^2
Far West	6.47	.541	.074	.061	137	.554
Arizona	6.71	.463	.114	.046	125	.543
Bank Street	3.97	.574	.123	.034	41	.787
Oregon						
Kansas	-15.3	.359	.145	.463	41	.574
High/Scope	2.20	.516	.077	.114	107	.643
Florida	-15.2	.716	-.146	.448	39	.552
EDC	19.2	.638	.080	-.159	39	.628
Pittsburgh	7.66	.513	.146	.002	119	.656
REC						
Enablers	3.56	.499	.144	.062	103	.768
Control	-3.77	.644	.145	.099	54	.747
NPV	2.68	.494	.128	.090	205	.623

Table VI-11

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Blacks

Model	c	PSI Fall	PPV Fall	Age	n	R^2
Far West						
Arizona	1.86	.746	.008	.096	74	.673
Bank Street	.963	.463	.125	.095	215	.724
Oregon	13.2	.423	.145	-.038	74	.551
Kansas	-9.21	.547	.119	-.287	61	.416
High/Scope	-.815	.071	.294	.146	54	.411
Florida	5.64	.518	.141	.016	103	.508
EDC	-3.16	.417	.064	.226	129	.553
Pittsburgh						
REC						
Enablers	3.86	.414	.139	.048	74	.411
Control	-3.62	.629	.205	.088	51	7.37
NPV	1.55	.543	.104	.097	383	.599

Table VI-12

REGRESSION MODELS RELATING PSI POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Spanish Americans

Model	c	PSI Fall	PPV Fall	Age	n	R^2
Far West						
Arizona						
Bank Street						
Oregon	-1.32	.417	.047	.246	76	.675
Kansas						
High/Scope	17.97	.482	.139	-.196	33	.335
Florida	-.324	.386	.203	.130	30	.354
EDC						
Pittsburgh						
REC	15.6	.561	.130	-.169	44	.507
Enablers	-4.07	.423	.087	.239	47	.640
Control	--2.60	.670	.117	.108	21	.829
NPV	-2.65	.518	.101	.189	111	.623

Table VI-13

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Whites

Model	a	PSI Fall	PPV Fall	Age	n	R^2
Far West	29.9	.309	.419	-.025	131	.550
Arizona	28.8	.322	.480	-.072	121	.484
Bank Street	19.9	.335	.476	.043	41	.588
Oregon						
Kansas	13.4	.086	.564	.202	42	.574
High/Scope	13.1	.310	.582	.093	103	.677
Florida	-.845	-.191	.788	.304	39	.703
EDC	32.4	.272	.428	-.082	39	.537
Pittsburgh	26.4	.214	.490	.016	115	.611
REC						
Enablers	20.9	.406	.496	.009	100	.665
Control	1.97	.550	.503	.299	48	.718
NPV	18.1	.320	.476	.104	204	.561

Table VI-14

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Blacks

Model	c	PSI Fall	PPV Fall	Age	n	R^2
Far West						
Arizona	5.84	.824	.627	.066	74	.582
Bank Street	12.1	.307	.594	.067	221	.631
Oregon	37.2	.694	.346	-.237	61	.675
Kansas	15.3	.249	.478	.111	61	.438
High/Scope	-1.22	-.231	.962	.250	53	.691
Florida	10.2	.292	.452	.191	93	.439
EDC	.139	.216	.646	.272	126	.616
Pittsburgh						
REC						
Enablers	.738	.327	.576	.267	73	.481
Control	5.05	.290	.600	.214	52	.574
NPV	10.4	.397	.509	.133	365	.546

Table VI-15

REGRESSION MODELS RELATING PPV POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Spanish Americans

Model	c	PSI Fall	PPV Fall	Age	n	R^2
Far West						
Arizona						
Bank Street						
Oregon	-12.3	.097	.561	.570	74	.616
Kansas						
High/Scope	40.6	.462	.627	-.368	32	.369
Florida	-32.7	1.28	.411	.756	29	.610
EDC						
Pittsburgh						
REC	50.0	.146	.696	-.531	41	.599
Enablers	5.13	.320	.533	.272	41	.697
Control	22.5	-.210	.814	-.137	19	.694
NPV	7.82	.533	.421	.243	111	.380

Table VI-16

REGRESSION MODELS RELATING WRTC POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Whites

del	c	PSI Fall	PPV Fall	Age	WRTC Fall	n	R^2
r West	-.92	.147	.015	.040	.544	134	.262
izona	-2.63	.203	.011	.059	.462	129	.479
nk Street	-5.31	.316	.026	.081	.123	43	.545
egon							
nsas	-9.64	.060	.073	.220	.293	42	.295
gh/Scope	-17.7	.099	.036	.334	.418	110	.641
lorida	-19.8	.159	-.174	.418	.920	40	.605
C	.499	.431	-.106	.028	.112	38	.500
ttsburgh	-4.22	.094	.060	.086	.603	118	.479
C							
nablers	-2.27	.168	.010	.043	.917	106	.072
ntrol	-2.93	.147	.061	.024	.205	40	.668
V	-9.34	.223	.038	.147	.580	213	.505

Table VI-17

REGRESSION MODELS RELATING WRTC POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Blacks

del	c	PSI Fall	PPV Fall	Age	WRTC Fall	n	R^2
r West izona	-3.21	.042	.038	.088	.623	77	.316
nk Street	-5.06	.134	.010	.100	.633	228	.692
egon	16.3	.011	.116	-.239	.599	71	.326
nsas	-3.54	-.048	.124	.117	.644	66	.357
gh/Scope	4.41	.183	.031	-.115	1.13	52	.546
orida	-6.66	.238	.016	.094	.570	99	.519
C ttsburgh C	.922	.027	.026	.048	.427	134	.216
ablers	2.75	.010	.079	-.049	.937	78	.519
ntrol	.324	.031	-.005	.004	.710	41	.618
V	-4.86	.036	.053	.084	.887	396	.530

Table VI-18

REGRESSION MODELS RELATING WRTC POST-TEST TO PRE-TEST, OTHER FALL TEST SCORES, AND AGE

Spanish Americans

del	c	PSI Fall	PPV Fall	Age	WRTC Fall	n	R^2
r West							
izona							
nk Street							
egon	19.2	.126	-.076	-.182	.629	77	.359
nsas							
gh/Scope	9.30	.218	-.148	-.091	.865	34	.463
orida	.926	-.035	.197	.014	.264	27	.417
C							
ttsburgh							
C	2.30	.214	.057	-.063	.773	45	.326
ablers	.626	.031	.053	.023	.814	53	.383
ntrol	-11.9	-.251	.081	.285	.816	18	.570
V	-24.2	.205	-.022	.460	.345	117	.636

treatment groups (10 models, NPV, Controls). In each case, we could include as independent variables along with the covariates, dummy variables for all but one of the treatment groups. The coefficient of a dummy variable estimates the difference in program effects (α_j's) between the corresponding treatment group and the "base" group for which no dummy was included. In our first runs, we used the Control children as the base group for our comparisons* The results of these analyses appear in Tables VI-19 through VI-24. Since we were also interested in the significance of comparisons between PV and NPV, we ran another ANCOVA with the Controls deleted and the NPV children as our base. The results appear in Tables VI-25 though VI-32.

Results of the Analysis of Covariance by Test

In this section we present brief summaries of the ANCOVA results for each of the 8 tests in our battery. In this section, when we refer to an "effect" of a program, we mean the estimated difference between its effect and that of the Controls. When we say simply that an effect is significant, we mean at least at the .05 level.

*This could not be done for the ITPA and ETS since the Controls were not given these tests.

Table VI-19

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

PSI

Program	White		Black	
	Effect	t	Effect	t
Far West	4.63	8.38	4.24	4.60
Arizona	4.29	7.67	3.41	5.03
Bank Street	2.90	4.11	1.80	3.09
Oregon	5.94	5.77	5.86	8.52
Kansas	5.10	7.21	2.78	3.95
High/Scope	3.35	5.80	2.25	2.09
Florida	2.94	4.07	3.29	5.15
EDC	5.28	7.33	3.84	6.22
Pittsburgh	4.79	8.58		
REC	2.27	2.29	2.97	3.16
Enablers	4.14	7.15	1.97	2.91
NPV	4.09	7.84	3.37	6.04
Control				
Constant	3.29		3.23	
Fall PSI Coeff.	.543		.550	
Fall PPV Coeff.	.125		.127	
Fall Coeff.				
F	152.39		152.99	
R^2	.676		.613	
n	1039		1268	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-20

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

PPV

	White		Black	
Program	Effect	t	Effect	t
Far West	4.07	3.62	5.51	2.98
Arizona	2.69	2.37	2.44	1.68
Bank Street	.52	.37	.270	.24
Oregon	1.12	.87	3.72	2.64
Kansas	2.25	1.61	1.90	1.37
High/Scope	.983	.84	-.840	-.58
Florida	1.79	1.25	2.02	1.58
EDC	2.05	1.43	1.27	1.04
Pittsburgh	4.42	3.89		
REC	.679	.34	3.89	2.08
Enablers	1.63	1.39	-.309	-.23
NPV	2.37	2.24	1.69	1.54
Control				
Constant	20.9		15.2	
Fall PSI Coeff.	.311		.408	
Fall PPV Coeff.	.509		.558	
Fall Coeff.				
F	111.00		120.60	
R^2	.609		.564	
n	1011		1225	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-21

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

WRTC

Program	White		Black	
	Effect	t	Effect	t
Far West	1.74	3.10	2.29	3.35
Arizona	1.83	3.25	2.49	4.86
Bank Street	1.56	2.29	1.59	3.54
Oregon	4.40	4.40	2.57	4.89
Kansas	3.59	5.24	4.45	8.45
High/Scope	2.51	4.33	.764	1.41
Florida	2.15	3.08	1.28	2.60
EDC/	1.15	1.62	3.36	7.11
Pittsburgh	1.83	3.24		
REC	.636	.71	1.70	2.49
Enablers	1.58	2.72	1.75	3.41
NPV	1.87	3.50	1.53	3.52
Control				
Constant	-2.12		-1.82	
Fall PSI Coeff.	.201		.108	
Fall PPV Coeff.	.037		.050	
Fall WRTC Coeff.	.555		.724	
F	67.52		96.08	
R^2	.497		.512	
n	1033		1295	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-22

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

WRTR

Program	White		Black	
	Effect	t	Effect	t
Far West	1.90	5.68	1.92	3.43
Arizona	2.26	6.77	2.56	6.08
Bank Street	1.33	3.28	1.79	4.86
Oregon	2.11	3.57	1.98	4.59
Kansas	2.46	6.06	3.10	7.17
High/Scope	1.74	5.04	1.16	2.62
Florida	2.16	5.20	2.39	5.92
EDC	2.38	5.67	2.77	7.13
Pittsburgh	2.78	8.25		
REC	2.06	3.87	2.62	4.70
Enablers	1.35	3.90	1.64	3.91
NPV	1.52	4.80	1.65	4.64
Control				
Constant	3.72		2.89	
Fall PSI Coeff.	.023		.069	
Fall PPV Coeff.	.033		.023	
Fall WRTR Coeff.	.221		.275	
F	34.18		44.87	
R^2	.333		.329	
n	1043		1295	

* $t > 1.96$ is equivalent to $p < .05$
  $t > 2.58$ is equivalent to $p < .01$
  $t > 3.27$ is equivalent to $p < .001$

Table VI-23

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

WRTN

Program	White		Black	
	Effect	t	Effect	t
Far West	.754	1.32	2.61	3.53
Arizona	3.23	5.63	2.47	4.44
Bank Street	.841	1.22	1.21	2.49
Oregon	1.01	.99	2.58	4.54
Kansas	2.27	3.26	2.08	3.65
High/Scope	1.68	2.85	.206	.35
Florida	1.09	1.54	1.13	2.12
EDC	4.87	6.78	3.16	6.17
Pittsburgh	1.47	2.56		
REC	1.47	1.61	1.62	2.19
Enablers	1.07	1.81	.596	1.08
NPV	.864	1.59	1.57	3.33
Control				
Constant	-2.90		-2.86	
Fall PSI Coeff.	.186		.181	
Fall PPV Coeff.	.043		.042	
Fall WRTN Coeff.	.723		.764	
F	87.57		92.66	
R^2	.561		.503	
n	1043		1295	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-24

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO CONTROLS)*

WRTD

Program	White		Black	
	Effect	t	Effect	t
Far West	.642	3.25	1.12	3.98
Arizona	.877	4.43	1.04	4.95
Bank Street	.342	1.43	.563	3.14
Oregon	1.85	5.28	2.33	10.81
Kansas	1.32	5.48	1.96	9.05
High/Scope	.384	1.89	.345	1.55
Florida	.341	1.39	.676	3.35
EDC	1.20	4.86	1.16	5.99
Pittsburgh	1.10	5.52		
REC	.525	1.67	.656	2.35
Enablers	.385	1.89	.593	2.82
NPV	.463	2.47	.891	4.99
Control				
Constant	-1.05		-1.03	
Fall PSI Coeff.	.085		.079	
Fall PPV Coeff.	.023		.016	
Fall WRTD Coeff.	.302		.464	
F	71.37		90.85	
R^2	.510		.498	
n	1043		1295	

* t > 1.96 is equivalent to p < .05
t > 2.58 is equivalent to p < .01
t > 3.27 is equivalent to p < .001

Table VI-25

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

PSI

Program	White		Black	
	Effect	t	Effect	t
Far West	.539	1.42	.890	1.13
Arizona	.240	.62	.071	.15
Bank Street	-1.23	-2.10	-1.55	-4.85
Oregon	1.95	2.05	2.54	5.28
Kansas	1.03	1.76	-.578	-1.13
High/Scope	-.664	-1.60	-1.13	-2.08
Florida	-1.10	-1.83	-.069	- .17
EDC	1.24	2.08	.488	1.28
Pittsburgh	.661	1.68		
REC	-1.87	-2.05	-.461	- .58
Enablers	.089	.21	-1.38	-2.90
NPV				
Constant	7.68		6.73	
Fall PSI Coeff.	.531		.548	
Fall PPV Coeff.	.122		.123	
Fall Coeff.				
F	136.13		148.69	
R^2	.646		.597	
n	985		1217	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-26

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

PPV

Program	White		Black	
	Effect	t	Effect	t
Far West	1.69	2.29	3.86	2.44
Arizona	.350	.47	.591	.63
Bank Street	-1.90	-1.70	-1.40	-2.22
Oregon	-.452	-.24	2.05	1.99
Kansas	-.094	-.09	.238	.23
High/Scope	-1.31	-1.63	-2.52	-2.32
Florida	-.535	-.47	.349	.41
EDC	-2.56	-.22	-.399	-.52
Pittsburgh	2.00	2.63		
REC	-1.76	-.98	2.22	1.38
Enablers	-.705	-.88	-1.96	-2.07
NPV				
Constant	23.60		16.90	
Fall PSI Coeff.	.296		.414	
Fall PPV Coeff.	.507		.554	
Fall Coeff.				
F	109.22		124.36	
R^2	.599		.563	
n	963		1173	

* $t > 1.96$ is equivalent to $p < .05$
  $t > 2.58$ is equivalent to $p < .01$
  $t > 3.27$ is equivalent to $p < .001$

Table VI-27

RESULTS OF ANLAYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

WRTC

Program	White		Black	
	Effect	t	Effect	t
Far West	-.142	-.41	.753	1.33
Arizona	-.058	-.16	.963	2.88
Bank Street	-.337	-.64	.068	.30
Oregon	2.48	2.74	1.04	2.96
Kansas	1.71	3.22	2.93	8.22
High/Scope	.615	1.64	-7.57	-1.99
Florida	.257	.47	-.246	-.82
EDC	-.755	-1.35	1.84	6.86
Pittsburgh	-.054	-.15		
REC	-1.25	-1.58	.174	.31
Enablers	-.313	-.83	.217	.65
NPV				
Constant	-.238		-.351	
Fall PSI Coeff.	.202		.108	
Fall PPV Coeff.	.037		.052	
Fall WRTC Coeff.	.562		.724	
F	66.16		97.32	
R^2	.484		.505	
n	1003		1254	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-28

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

WRTR

Program	White		Black	
	Effect	t	Effect	t
Far West	.399	2.01	.274	.60
Arizona	.752	3.78	.903	3.35
Bank Street	-.202	-.68	.141	.78
Oregon	.609	1.19	.328	1.16
Kansas	.945	3.15	1.44	5.03
High/Scope	.242	1.14	-.498	-1.62
Florida	.669	2.17	.740	3.05
EDC	.881	2.80	1.13	5.20
Pittsburgh	1.24	6.07		
REC	.513	1.14	.972	2.14
Enablers	-.145	-.68	-.002	-.01
NPV				
Constant	5.42		4.59	
Fall PSI Coeff.	.026		.071	
Fall PPV Coeff.	.029		.022	
Fall WRTR Coeff.	.210		.268	
F	27.31		42.78	
R^2	.279		.310	
n	1003		1254	

* $t > 1.96$ is equivalent to $p < .05$
$t > 2.58$ is equivalent to $p < .01$
$t > 3.27$ is equivalent to $p < .001$

Table VI-29

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

WRTN

Program	White		Black	
	Effect	t	Effect	t
Far West	-.138	-.39	1.04	1.71
Arizona	2.34	6.58	.901	2.50
Bank Street	-.034	-.06	-.339	-1.40
Oregon	.107	.12	1.01	2.68
Kansas	1.38	2.58	.524	1.36
High/Scope	.781	2.06	-1.35	-3.28
Florida	.190	.34	-.428	-1.32
EDC	3.97	7.07	1.60	5.54
Pittsburgh	.599	1.64		
REC	.603	.75	.055	.09
Enablers	.170	.45	-.969	-2.70**
NPV				
Constant	-2.14		-1.38	
Fall PSI Coeff.	.187		.181	
Fall PPV Coeff.	.047		.045	
Fall WRTN Coeff.	.719		.756	
F	88.44		93.75	
R^2	.556		.496	
n	1003		1254	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-30

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

WRTD

Program	White		Black	
	Effect	t	Effect	t
Far West	.177	1.45	.228	.98
Arizona	.411	3.35	1.52	1.11
Bank Street	-.125	-.68	-.323	-3.53
Oregon	1.38	4.38	1.44	10.07
Kansas	.852	4.61	1.07	7.33
High/Scope	-.083	-.64	-.544	-3.49
Florida	-.123	-.65	-.211	-1.71
EDC	.736	3.81	.278	2.53
Pittsburgh	.628	4.99		
REC	.055	.20	-.235	-1.02
Enablers	-.081	-.61	-.298	-2.18
NPV				
Constant	-.58		-.157	
Fall PSI Coeff.	.086		.079	
Fall PPV Coeff.	.023		.017	
Fall WRTD Coeff.	.299		.452	
F	69.78		90.62	
R^2	.497		.487	
n	1003		1254	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-31

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

ITPA

Program	White		Black	
	Effect	t	Effect	t
Far West	-.901	-1.09	-.437	-.30
Arizona	-.022	-.02	-1.86	-1.81
Bank Street	.028	.02	1.90	-2.85
Oregon	4.91	1.70	-.833	-.77
Kansas	-2.04	-1.56	-.508	-.44
High/Scope	-2.64	-2.83	-1.71	-1.50
Florida	-1.54	-1.12	-.216	-.25
EDC	1.81	1.35	-.147	-.19
Pittsburgh	1.28	1.48		
REC	-4.06	-2.23	-3.45	-1.94
Enablers	-1.61	-1.86	-3.61	-4.23
NPV				
Constant	6.31		6.89	
Fall PSI Coeff.	.148		.186	
Fall PPV Coeff.	.052		.061	
Fall ITPA Coeff.	.476		.406	
F	18.29		15.82	
R^2	.388		.304	
n	418		476	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Table VI-32

RESULTS OF ANALYSIS OF COVARIANCE

(EFFECTS RELATIVE TO NPV)*

ETS

	White		Black	
Program	Effect	t	Effect	t
Far West	.659	1.97	1.20	1.74
Arizona	1.30	3.78	1.38	3.39
Bank Street	.388	.74	.244	.87
Oregon	2.18	2.64	2.34	5.67
Kansas	1.64	3.26	2.22	5.04
High/Scope	-.300	-.83	-1.14	-2.44*
Florida	.733	1.39	.266	.74
EDC	1.41	2.67	.647	1.95
Pittsburgh	.735	2.13		
REC	-2.06	-2.51	-1.81	-2.65
Enablers	-.735	-1.98	.016	.04
NPV				
Constant	4.11		3.76	
Fall PSI Coeff.	.180		.207	
Fall PPV Coeff.	.045		.064	
Fall ETS Coeff.	.404		.358	
F	73.56		75.39	
R^2	.518		.451	
n	974		1207	

* t > 1.96 is equivalent to p < .05
  t > 2.58 is equivalent to p < .01
  t > 3.27 is equivalent to p < .001

Preschool Inventory

For Whites the largest effects were achieved by Oregon(5.94), EDC (5.28), and Kansas (5.10). The Oregon and EDC effects were significantly above NPV. Smallest effects were for REC (2.27) and Bank Street (2.90). These were significantly below NPV. For Blacks, Oregon (5.86) and Far West (4.24) had the largest effects. Only Oregon was significantly above NPV. Smallest effects were for Bank Street (1.80), Enablers (.197), and High/Scope (2.25). All three were significantly below NPV. For both Blacks and Whites, all models performed significantly better than the Controls.

Peabody Picture Vocabulary Test

For Whites, Pittsburgh (4.42) and Far West (4.07) had the largest effects, both significantly above NPV. Smallest effects were for Bank Street (.52), REC (.68), and High/Scope (.98). None of these were significantly below NPV. For Blacks, Far West (5.51), REC (3.89), and Oregon (3.72) had the largest effects, with Far West and Oregon significantly above NPV. High/Scope (-.84), Enablers (-.31) and Bank Street (.27) were lowest, all three significantly below NPV. While all model effects for Whites and 8 of 10 for Blacks were positive, most effects were not significant.

WRAT Copying Marks

Largest effects for Whites were achieved by Kansas (5.24), Oregon (4.40), and High/Scope (4.33). Kansas and Oregon were significantly above NPV. REC (.71), EDC (1.62), and Bank Street (2.29) were low, although none were significantly below NPV. For Blacks, Kansas (4.45) and EDC (3.36) were high and significantly above NPV, while High/Scope (.76) was low and significantly below NPV. For Whites 9 of 11 model effects were significant and for Blacks 9 of 10.

WRAT Recognizing Letters

For Whites, Pittsburgh (2.78), Kansas (2.46), and EDC (2.38) were high, all significantly above NPV. Enablers (1.27) and Bank Street (1.33) were low, but not significantly below NPV. For Blacks, Kansas (3.10), EDC (2.77), REC (2.62), and Arizona (2.56) were high, all significantly above NPV. High/Scope (1.16) was low, but not significantly below NPV. All model effects were significant.

WRAT Naming Letters

For Whites, EDC (4.87) and Arizona (3.23) were high, both significantly above NPV. Bank Street (.84) was low, but not significantly below NPV. For Blacks, EDC (3.16) Far West (2.61), Oregon (2.58), and Arizona (2.47) were high, with all except Far West significantly above NPV.

For Whites 5 of 11 model effects were significant, and for Blacks 8 of 10.

WRAT Reading Numbers

For Whites, Oregon (1.85) and Kansas (1.32) were high, both significantly above NPV. Bank Street (.34), Florida (.34), High/Scope (.38), and Enablers (.39) were low, although none was significantly below NPV. For Blacks, Oregon (2.33) and Kansas (1.96) were high, and both were significantly above NPV, while High/Scope (.35) and Bank Street (.56) were low, and significantly below NPV. For Whites 6 of 11 model effects were significant and for Blacks 9 of 10.

ITPA Verbal Expression

Since the Control children did not take the ITPA, no comparisons with them were possible. From the comparisons with NPV, however, we find for Whites that Oregon (4.91) was highest, though not significantly above NPV. REC (-4.06) and High/Scope (-2.64) were lowest, both significantly below NPV. For Blacks we find that all had smaller estimated effects than NPV, with the Enablers (-3.61) significantly lower.

ETS Enumeration

As for the ITPA, no comparisons with Controls were

possible. From the comparisons with NPV, for the Whites Oregon (2.18) and Kansas (1.64) were high, although neither was significantly above NPV. REC (-2.06) and Enablers (-.74) were both significantly below NPV. For Blacks, Oregon (2.34) and Kansas (2.22) were high and both significantly above NPV, while REC (-1.81) was significantly below NPV.

Summary of ANCOVA Results

As for the previous analyses, we will present in this section evidence furnished by the ANCOVA bearing on our three major questions.

1. To what extend does a Head Start experience accelerate the rate at which disadvantaged preschoolers acquire cognitive skills?

Our evidence here come from the comparisons between the Control and Head Start children. For three of the six tests taken by the Controls (PSI, WRTC, WRTR), nearly all the PV and the NPV children do significantly better than the Controls for both Blacks and Whites. For two tests (WRTN, WRTD) most of the models perform better. Only for the PPV do the Control and Head Start children perform comparably.

2. Are the Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?

On the whole it appears that PV and NPV programs are similar in effectiveness. As a very rough measure of over-all NPV performance, we observe that for White children, of a total of 88 model effects on the 8 tests, 51 were above that of NPV and 37 were below. For Blacks, of 80 model effects, 46 were above and 34 below.

3. Are some PV models particularly effective at imparting certain skills?

Table VI-33 presents a summary of inter-model comparisons. In declaring a model particularly effective or ineffective for a given test, we have considered the size of the estimated difference in effects between the model and the Controls, the significance of the difference between the model and NPV, and the consistency across racial groups. The effects noted in Table VI-33 seem fairly evenly spread across the 8 tests. It is interesting that no model has both positive and negative effects. Oregon and Kansas are overall most impressive, each with 5 positive effects out of 8 tests.

Table VI-33

Summary of Relative Model Effectiveness

Based on Analysis of Covariance*

++ Indicates model appears to be highly effective.
+ Indicates evidence for above average effectiveness.
- Indicates evidence for below average effectiveness.
-- Indicates model appears to be highly ineffective.

Model	PSI	PPV	WRTC	WRTR	WRTN	WRTD	ITPA	ETS
Far West		+						
Arizona					++			++
Bank Street	--	-						
Oregon	++		++		+	++		++
Kansas			++	+	+	++		++
High/Scope	-	-					-	
Florida								
EDC				+	++	+		
Pittsburgh		+		+				
Enablers		-					-	

*REC not included because with only one site we felt it unfair to draw any conclusions.

Chapter VII

Resistant Analysis

Introduction and Theory

It was clear to us that the pre-test was the most important variable to control in making post-test comparisons. We thought it would be worthwhile to do some exploratory analysis to determine the relationship between fall and spring test scores, broken down by program and possibly background characteristics. All our previous analyses used means as summary measures of distributions of effects, and relationships were fitted via ordinary least-squares techniques. While we have confidence in these analyses, we felt it would be nice to have at least one analysis using other summary statistics and fitting methods which would be particularly robust, or resistant to departures from the usual assumptions underlying the standard procedures. We were thus led naturally to the recent work of John Tukey (1970). Tukey's exploratory data analysis techniques enable the analyst to comb a set of data for useful information without unwieldy computations and formal tests based on stringent assumptions. We found the resistant fitting technique particularly appropriate. With it, we could conveniently fit a model of the form

$$Y_{ij}' = \alpha_j + \beta_j f_j(Y_{ij}) + e_{ij} \qquad (7.1)$$

where Y_{ij}' and Y_{ij} are pre- and post-test scores for individuals i in group j, e_{ij} is the error term, and f_j is a transformation or re-expression of Y_{ij} consisting of any power or the logarithm. For example, on the PSI, for White children with no prior preschool in Far West, we fit the model:

$$Y' = -.32 + 19.13 \log Y$$

For Black children with no prior preschool in Far West, we obtained

$$Y' = 26.73 - \frac{80.89}{Y}$$

Note that the class of models described by (7.1) can be characterized as linear in terms of the re-expressed pre-test score. The details of how the appropriate re-expression is selected, and the slope and intercept estimated can be found in Appendix E, written by Sharon Hauck. In ordinary least-squares fitting, outlying observations (those with very large values of e_{ij}) exercise a strong influence in determining the fitted curve. Resistant fitting is much less sensitive to such outliers. Thus, it provides a fairer representation of the data

in a situation where nearly all observations reflect a systematic relationship, but there are a few "wild" observations. Since these wild observations are not given much weight in the curve-fitting, they will also stand out more strongly than in a least-squares regression when we look at the residuals.*

In trying to apply the resistant fitting technique to the various tests in our battery, we found that the floor and ceiling effects of the WRAT subtests made it virtually impossible to implement the fitting algorithm. We decided to take a different tack with three of the WRAT subtests. For the WRAT Recognizing Letters, Naming Letters, and Reading Numbers, it seemed that many children were achieving an effective maximum, so that their potential gain was strongly dependent on where they started out. It seemed reasonable to consider these as criterion-referenced tests. We, therefore, set a level for each test which we felt corresponded to reasonable mastery of the subject matter. For each child we could then note simply whether or not he reached this criterion. Looking at all children with a given pre-test score (or narrow range of scores) we could then see for each program the proportion reaching criterion.

---

*Residual = Observed Value - Fitted Value

It would, of course, be desirable to control for other variables as well. Since ethnicity seemed to have a strong relationship to outcomes, we considered Blacks and Whites separately.* Sample sizes were not adequate for further breakdown by other background characteristics.

The other five tests did not seem to us suitable for the criterion-reference analysis, since the did not involve such clear-cut, concrete skills and it was not clear how to set a criterion for subject mastery. For each of these, we performed a resistant fitting analysis. We broke the children in each program out according to ethnicity and whether or not they had any prior preschool experience. For each sub-class in which there were at least 20 children, we then fit a model of the form described by equation (7.1). We studied the resulting functions, but no strong patterns became obvious. We decided to attempt to obtain simple comparisons among programs by developing a resistant analog to the usual least-squares analysis of covariance.

Suppose that for any ehtnicity by prior preschool experience sub-class, we can represent the relationship between fall and spring tests by a model of the form:

$$Y_{ij}' = \alpha_j + \beta f(Y_{ij}) + e_{ij} \qquad (7.2)$$

---

*There were not enough Spanish-Americans to make the analysis for them worthwhile.

That is, we assume that the re-expression f and the slope β are the same for all programs. Thus, as in the ANCOVA, α_j becomes a measure of relative program effect.

From the various re-expressions found in the model-fitting described above, we selected a compromise re-expression reasonably acceptable for all programs, though perhaps not optimal for any particular program. For each program, a model was then fit using the common re-expression. This resulted in a set of up to 13 slopes,* one for each program. From this set we determined a compromise slope, hopefully reasonable for all programs. Having decided on both β and f, we could now estimate α_j by taking a kind of weighted average of the deviations of the Y_{ij}'s in group j from the function $\beta f(Y_{ij})$. The details of the steps described above can be found in Appendix E.

To calculate a program "effect" we took the median of the estimated α_j's and subtracted this from each of the α_j's individually. The result is analogous to that of a standard one-way ANCOVA with effects computed around a grand mean of all programs. To see how our results compared with those of the more traditional approach, we carried out such an ANCOVA.

---

*The actual number was the number of programs with at least 20 children.

Results of the Criterion-Reference Analysis

Results of the criterion-reference analysis for the WRTR, WRTN, and WRTD appear in Tables VII-1 through VII-5. For the WRTR the maximum score was 10, and we decided that to reach criterion a child must achieve a score of at least 9. For the WRTN our criterion was 10 out of 13 correct. The WRTD requires the child to read the numbers "3, 5, 6, 17, 41." It seems that 17 proved quite difficult and 41 much too difficult for our sample. We, therefore, decided that 3 of 5 correct seemed a reasonable criterion.

For Blacks and Whites separately, we looked at all children with fall scores in certain narrow ranges, and recorded the number reaching criterion and the number failing to do so. We also calculated the proportions of PV children, NPV children, and Control children reaching criterion. Following are summaries of the results. Note that we elected not to perform significance tests for differences between proportions. There were so many possible inter-dependent tests that significance levels would be severely compromised. The reader has, of course, from Tables VII-1 through VII-5 all the information necessary to carry out any tests he may deem useful.

Table VII-1

RESULTS OF CRITERION REFERENCE ANALYSIS FOR WRTR

Black

Fall Score = 2

Far West	Arizona	Bank St.	Oregon	Kansas	H/S	Florida	EDC	Pittsburgh	REC	Enablers	NPV	Control	Total PV	
1	9	12	0	9	3	3	5	0	1	1	12	0	44	Spring = 9,10
1	12	37	0	2	15	7	8	0	3	4	42	8	89	Spring < 9
											.222		.330	

Fall Score = 3, 4

2	5	5	4	7	0	1	10	0	1	4	8	2	39	Spring = 9,10
1	1	11	2	4	6	2	5	0	0	6	20	5	37	Spring < 9
											.286		.513	

Fall Score = 5, 6

2	5	13	4	9	2	6	12	0	3	5	28	1	61	Spring = 9,10
2	3	9	3	3	7	6	6	0	2	8	37	3	49	Spring < 9
											.431		.556	

Fall Score = 7, 8

3	16	37	11	14	3	8	28	0	7	9	28	3	136	Spring = 9,10
2	2	12	3	4	4	5	5	0	2	6	29	7	47	Spring < 9
											.491		.743	

Fall Score = 9, 10

7	19	74	29	13	7	18	51	0	4	24	112	5	246	Spring = 9,10
2	2	10	5	0	6	3	3	0	1	6	12	1	38	Spring < 9
											.903		.866	

Table VII-2

RESULTS OF CRITERION REFERENCE ANALYSIS FOR

WRTR

White

Fall Score ≤ 2

Far West	Arizona	Bank St.	Oregon	Kansas	H/S	Florida	EDC	Pittsburgh	REC	Enablers	NPV	Control	Total PV	
7	9	0	3	3	2	0	2	12	0	0	4	4	38	Spring = 9,10
7	7	5	7	2	7	0	1	8	3	5	11	14	52	Spring < 9
											.267	.222	.422	

Fall Score = 3, 4

7	10	1	1	3	2	3	1	8	1	3	15	0	40	Spring = 9,10
7	6	3	0	1	3	0	0	1	0	8	18	1	29	Spring < 9
											.455		.580	

Fall Score = 5, 6

10	12	5	1	8	6	3	3	24	1	8	19	0	81	Spring = 9,10
8	4	3	0	0	4	1	1	3	2	5	10	4	31	Spring < 9
											.655		.723	

Fall Score = 7, 8

20	23	6	2	10	23	6	7	22	2	20	36	2	141	Spring = 9,10
8	5	2	0	0	6	0	0	2	1	5	12	2	27	Spring < 9
											.750		.839	

Fall Score = 9, 10

47	49	12	6	14	48	14	21	36	5	41	57	7	293	Spring = 9,10
7	2	2	1	0	3	2	1	1	1	7	9	1	27	Spring < 9
											.864		.902	

Table VII-3

RESULTS OF CRITERION REFERENCE ANALYSIS FOR

WRTN

White

Far West	Arizona	Bank St.	Oregon	Kansas	H/S	Florida	EDC	Pittsburgh	REC	Enablers	NPV	Control	Total PV	
8	26	2	1	6	18	2	16	6	1	10	7	1	96	Spring > 10
104	82	30	7	31	66	22	17	99	13	81	170	31	551	Spring < 10
											.040	.032	.145	

Black

1	9	8	12	3	0	1	22	0	0	2	30	0	58	Spring > 10
18	61	181	52	59	52	53	102	0	23	69	308	31	670	Spring < 10
											.089	0	.080	

Table VII-4

RESULTS OF CRITERION REFERENCE ANALYSIS FOR

WRTD

Black

Fall Score = 0

Far West	Arizona	Bank St.	Oregon	Kansas	H/S	Florida	EDC	Pittsburgh	REC	Enablers	NPV	Control	Total PV	
8	13	22	48	26	3	7	25	0	5	11	59	1	168	Spring 3
7	32	133	9	21	41	33	63	0	17	46	212	27	402	Spring 3
											.218	.036	.295	

Fall Score = 1

2	10	14	3	8	1	5	11	0	0	1	24	0	55	Spring 3
0	9	11	0	0	5	3	8	0	0	7	23	1	48	Spring 3
											.511	.000	.534	

Fall Score = 2

2	4	12	3	5	2	2	9	0	0	3	13	1	42	Spring 3
1	4	1	0	1	1	2	2	0	0	2	4	2	14	Spring 3
											.765	.333	.750	

Fall Score = 3

3	2	20	6	4	0	6	13	0	2	1	17	2	57	Spring 3
0	0	1	0	0	0	1	0	0	0	0	1	0	2	Spring 3
											.944	1.000	.965	

Table VII-5

RESULTS OF CRITERION REFERENCE ANALYSIS FOR

WRTD

White

Fall Score = 0

Far West	Arizona	Bank St.	Oregon	Kansas	H/S	Florida	EDC	Pittsburgh	REC	Enablers	NPV	Control	Total PV		
27	29	5	3	17	28	5	14	37	5	22	51	3	192	Spring	3
54	44	23	0	12	38	10	8	49	9	39	104	25	286	Spring	3
											.329	.107	.402		

Fall Score = 1

10	15	2	2	4	2	2	3	12	0	8	8	1	60	Spring	3
8	4	1	0	0	3	1	0	4	0	5	7	1	26	Spring	3
											.533	.500	.698		

Fall Score = 2

6	11	2	1	3	8	1	1	5	1	5	6	2	44	Spring	3
2	1	1	0	0	2	0	0	0	1	0	2	1	7	Spring	3
											.750	.667	.861		

Fall Score = 3

16	21	4	3	4	17	7	10	6	0	3	15	1	91	Spring	3
2	0	0	0	0	0	0	0	0	0	0	0	0	2	Spring	3
											1.00	1.00	.978		

WRAT Recognizing Letters

For Whites, we noticed that the proportion of PV children reaching criterion was higher than the proportion of NPV children for all 5 fall test score levels. It appears that Kansas, Pittsburgh, and possibly Arizona and EDC are particularly effective. For Blacks PV children also did consistently better than NPV children except for those reaching criterion in the fall. Kansas seems particularly effective, and possibly Arizona and EDC. Note that for most programs and fall scores, the proportion reaching criterion is smaller for Blacks than for Whites.

WRAT Naming Letters

It turned out that nearly all children scored 3 or less in the fall, and very few of these reached criterion in the spring. Thus, it was difficult to make much of the results. There is some evidence that Arizona, Oregon, and EDC may be particularly effective.

WRAT Reading Numbers

For both Whites and Blacks the majority of children scored 0 in the fall. There are enough with other scores to be worth presenting, but too few to draw conclusions. For those with fall scores of 0, the proportion reaching criterion was higher for PV than for NPV for both Blacks and Whites. The proportion for Whites was higher than that for

Blacks for both PV and NPV children. For Whites, Kansas and possibly Oregon seem particularly effective. For Blacks, Oregon is outstanding and Kansas also particularly effective.

Results of the Resistant Analysis of Covariance

As explained above, we initially attempted a resistant fit to describe the relationship between fall and spring scores for each of our tests. As a result of ceiling and floor effects, this proved particularly difficult for the WRAT subtests. For the other tests, it was difficult to summarize the results meaningfully, and we decided to attempt the resistant ANCOVA analog. Unfortunately, for the ITPA and ETS, many of the models did not contain enough children to justify their inclusion, and the choice of a compromise re-expression to be applied to all programs was very difficult. We, therefore, decided to carry out the resistant ANCOVA for the PSI and PPV only.

The results appear in Tables VII-6 and VII-7. For each ethnicity by prior preschool sub-classes, we have calculated an estimated effect for each program containing at least 20 children, relative to the median effect for all such programs. As a comparison, we have also performed

RESULTS FOR RESISTANT ANALYSIS OF COVARIANCE FOR PSI

Program	White No Prior PS	Black No Prior PS	Sp. Amer. No Prior PS	White Prior PS	Black Prior PS
Far West	.9	3.1		.5	
Arizona	0	1.0		-.8	0
Bank Street		-1.2		-1.6	-2.0
Oregon		3.0	3.0		.6
Kansas	.6	- .3			
High/Scope	- .1	- .4			
Florida	-1.5	1.3	.2		
EDC		0		.8	.5
Pittsburgh	.2			0	
REC			- .3		
Enablers	.6	.1	0	0	
Control	-4.1	- .3			
NPV	-2.0	- .1	- .4	0	- .2
Coefficient	71.6	.65	.69	.55	5.1
Median	40.4	8.2	9.6	13.1	.2
n	711	807	250	179	312
Re-expression	- .5	1	1	1	.5

RESULTS OF RESISTANT ANALYSIS OF COVARIANCE FOR PPV

Program	White No Prior PS	Black No Prior PS	Sp. Amer. No Prior PS	White Prior PS	Black Prior PS
Far West	- .2	6.			
Arizona	.2	.5		-2.6	1.2
Bank Street	.1	- .1			-1.3
Oregon		4.1	0		
Kansas		1.7			
High/Scope	- .7	-1.2			
Florida	.8	2.3	-1.7		
EDC		0		0	.3
Pittsburgh	.6			6.4	
REC			.5		
Enablers	- .7	-1.8	0	- .9	
Control	0	0			
NPV	- .6	- .5	4.1	2.1	- .2
Coefficient	36.9	.71	6.11	712.53	7.36
Median	8.7	17.2	11.2	68.2	1.5
n	692	765	248	136	272
Re-expression	log	1	.5	-1	.5

RESULTS OF ANALYSIS OF COVARIANCE FOR PSI

Program	White No Prior PS	Black No Prior PS	Sp. Amer. No Prior PS	White Prior PS	Black Prior PS
Far West	1.1	1.8		1.2	
Arizona	.6	.2		- .6	- .2
Bank Street		-1.6		-1.3	-1.8
Oregon		2.7	2.7		1.1
Kansas	1.0	.1			
High/Scope	0	-1.1			
Florida	- .4	.6	- .6		
EDC		- .4		.9	.9
Pittsburgh	.7			- .6	
REC			-1.0		
Enablers	.9	- .7	- .6	.6	
Control	-3.7	-1.7			
NPV	- .1	0	- .5	- .2	0
Regression Coefficient	.685	.702	.610	.651	.656
Spring Mean	20.52	16.55	19.51	22.65	20.68
Fall Mean	15.58	11.99	13.95	17.95	16.89
F	9.980	5.411	7.457	2.239	4.608
Significance	< .001	< .001	< .001	.04	.002

Table VII-9

RESULTS OF ANALYSIS OF COVARIANCE FOR PPV

Program	White No Prior PS	Black No Prior PS	Sp. Amer. No Prior PS	White Prior PS	Black Prior PS
Far West	.4	3.7			
Arizona	.8	1.2		- .6	.4
Bank Street	- .6	-1.4			-1.4
Oregon		3.0	.1		
Kansas		.9			
High/Scope	- .7	-2.9			
Florida	.2	1.6	-4.4		
EDC		-1.6		.4	1.1
Pittsburgh	1.4			1.2	
REC			.8		
Enablers	-1.3	-2.0	- .5	-1.3	
Control	0	-1.7			
NPV	- .3	- .8	4.0	.2	0
Regression Coefficient	.600	.677	.521	.626	.678
Spring Mean	47.68	36.71	43.28	48.63	41.79
Fall Mean	37.05	25.87	29.24	39.53	31.72
F	.903	3.607	4.780	.525	1.38
Significance	not sign.	< .001	.002	not sign.	not sign.

an ordinary least-squares ANCOVA with effects computed about the grand mean.* These results appear in Tables VII-8 and VII-9. We now present brief summaries of the resistant ANCOVA results.

Preschool Inventory

For White children with no prior pre-school, the Controls have by far the lowest effect (-4.1). For Blacks with no prior preschool, Far West (3.1) and Oregon (3.0) are high, and for Spanish Americans, Oregon (3.0) is outstanding. The effects for both Whites and Blacks with prior preschool are rather homogeneous. Overall, it appears that Far West and Oregon are particularly effective; Bank Street and Control particularly ineffective. The results for the standard ANCOVA seem remarkably consistent with those of our resistant ANCOVA.

Peabody Picture Vocabulary Test

For White children with no prior preschool experience, program effects seem quite homogeneous. For Blacks with no

---

*There were three main differences between these ANCOVA's and those carried out in Chapter VI. First, the children were broken down by prior preschool experience as well as ethnicity. Second, the effects were computed about the grand mean (an unweighted mean of the spring means for all programs) rather than relative to the Controls. Third, the analysis was carried out using an unweighted means approach rather than exact least-squares.

prior preschool, Oregon (4.1) and perhaps Florida (2.3) seem particularly effective. For Spanish-Americans, NPV (4.1) does best. For Whites with preschool, Pittsburgh (6.4) seems highly effective, and Arizona (-2.6) possibly ineffective. For Blacks, program effects are quite homogeneous. The general profile of effects from the standard ANCOVA is similar, although the magnitudes of effects differ.

Summary

We have results for only five tests (PSI, PPV, WRTR, WRTN, WRTD) from the analyses discussed in this chapter. As in previous chapters, we present here the evidence provided by these analyses bearing on our three major questions.

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged preschoolers acquire cognitive skills?

For all tests except the PPV, the Controls appear to do substantially worse than both the PV and NPV children. On the PPV, Head Start and Control results are comparable.

2. Are the Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?

For each of the 5 tests except the PPV, the PV programs as a whole perform slightly better than NPV. For the PPV, their performance is roughly equivalent. We are inclined to attribute the slight superiority of PV to a couple of particularly effective models, so that, except for these, the effects of PV and NPV are comparable.

3. Are some PV models particularly effective at imparting certain skills?

Table VII-8 presents a summary of inter-model comparisons. In deciding whether to declare a model particularly effective, we have considered whether the proportion reaching criterion is consistently higher than the overall PV proportion for all ethnic groups and fall scores. For the PSI and PPV, we have considered the size of effects and their consistency over the ethnicity by prior preschool sub-classes. Note that out of 11 positive effects, 7 are for the academic models (Oregon, Kansas, Pittsburgh). Overall, Oregon is most impressive.

Table VII-10

SUMMARY OF RELATIVE MODEL EFFECTIVENESS

BASED ON RESISTANT ANALYSIS*

++ Indicates model appears to be highly effective.
+ Indicates evidence for above average effectiveness.
- Indicates evidence for below average effectiveness.
-- Indicates model appears to be highly ineffective.

Model	PSI	PPV	WRTR	WRTN	WRTD
Far West	+				
Arizona			+	+	
Bank Street	-				
Oregon	++			+	++
Kansas			+		+
High/Scope			-		
Florida					
EDC				+	
Pittsburgh		+	+		
Enablers					
NPV					

*REC not included because with only one site we felt it unfair to draw any conclusions.

Chapter VIII

Background Characteristic by Program Interactions

Introduction

The previous four chapters have attempted to present a picture of the pattern of overall effects of various programs. In this chapter we explore the question of whether the relative effectiveness of various programs is related to certain child background characteristics. Featherstone (1973) studied this question using the 1969-70 and 1970-71 data. We are in no way trying to replicate her careful and thorough study. Without a carefully designed randomized experiment, the problems involved in estimating interaction effects are much more difficult than the already difficult problems involved in measuring main effects (see Chapter III). Definitive conclusions from our data are virtually impossible. Nonetheless, we felt that a modest effort to see what interactions are suggested by the data, and how they relate to Featherstone's general conclusions, would be valuable.

The outcomes Featherstone used were the Stanford-Binet IQ and the 64-item PSI. Since neither of these tests was given in 1971-72, comparisons with her results are

difficult. The background characteristics she considered were initial (fall) IQ, prior preschool experience, sex, age, socio-economic status, ethnicity (Black and White only) and cognitive style (as measured by the Hertzig-Birch coding of the Stanford-Binet). We have no measure of IQ for the 1971-72 cohort, and, although a version of the Hertzig-Birch scoring system was used with the 32-item PSI, we felt the system was too experimental to use at this time. As for SES, we felt that from the standpoint of reliability and impact on test scores, mother's education was our best variable. We therefore decided to look only at sex, mother's education, ethnicity (Black and White only), age, and prior preschool experience.

Since the interpretation of interaction effects is sometimes confusing, it may be useful to explain exactly what they mean in this context, and why they are so difficult to estimate. For simplicity, suppose we have two programs, A and B, and that sex is the background variable of interest. Assume we have some measure of program effectiveness (e.g., residual, adjusted mean) and that in terms of this measure we obtain the hypothetical results displayed in Figure VIII-1a. In this case there is no interaction between sex and program, since the difference between the effects of the two programs is 4 for both boys

Figure VIII-1

ILLUSTRATION OF PROGRAM BY SEX INTERACTION

Program

Interaction

	A	B
Male	12	8
Female	14	10

(a)

(12-8)-(14-10) = 0

Male	12	8
Female	14	12

(b)

(12-8)-(14-12) = +2

Male	12	8
Female	10	12

(c)

(12-8)-(10-12) = +6

and girls. In situation (b), on the other hand, the difference between program A and B is larger for boys than for girls. Thus, the relative effectiveness of the programs is related to the child's sex. We have a program-by-sex interaction. Finally, in (c) we have a disordinal interaction. Not only is the difference in effects greater for boys, but the direction of relative effectiveness of the two programs is actually reversed. Program A is better than B for boys, while program B is better for girls.

Notice that an interaction is really a difference of differences. Thus, in estimating an interaction effect from a finite sample, a small sample in any of the four cells can lead to imprecise estimates (i.e., large variance). In the extreme, an empty cell makes the estimation impossible. If, for example, there were no boys in program A, no statistical procedure could provide a reasonable estimate of the interaction.

Methodology

We decided that the simplest way to measure interaction effects would be to use the "combined" residuals derived in Chapter V as an outcome measure. Recall that the residual is an estimate of the effect of the program

in which the child is enrolled over and above what we would expect on the basis of natural maturation. We can perform a two-way analysis of variance, with program and a background variable as factors. This ANOVA will provide an F-test for the significance of the overall interaction effect. Looking at the pattern of cell means, we will hopefully be able to interpret any interactions detected. As an added benefit, we will also obtain F-tests for the main effects of program and background characteristics. If we observe a large main effect corresponding to a background variable which is unevenly distributed across the various programs, there may be some bias in the magnitude of the estimated model mean residuals.

Because the design is quite unbalanced (unequal cell sizes) an exact least-squares solution would be quite complex. We therefore carried out an unweighted means analysis. Unfortunately, for some background variables, the design may be so unbalanced that the F-test resulting from the unweighted means analysis may be misleading. Since our primary interest is in the estimation of effects rather than formally testing hypotheses, this does not concern us overly. Moreover, in carrying out ANOVA's on the six tests for which we have computed residuals, for each of the five background characteristics, we perform a

large number of statistical tests. Thus the formal significance level of any individual test might be compromised even with an exact least-squares analysis.

Results of Interaction Analysis

In this section we present the results of the interaction study. Detailed results are presented in Tables VIII-1 through VIII-28. We first present brief summaries for each background characteristic, followed by some concluding comments.

Sex. The only significant main effects for sex occur on the PSI and PPV. There are small differences favoring boys on both. There are no significant program-by-sex interaction effects on any of the tests. The overall pattern of relative model effectiveness is quite similar for boys and girls.

Ethnicity. All tests except the WRTD show significant main effects for ethnicity. The PSI and WRTC effects favor Whites, while the PPV, ITPA, and ETS effects favor Blacks. Only the WRTC and WRTD have significant interaction effects. The WRTC effect ($p<.001$) is largely attributable to High/Scope and EDC. High/Scope was highly

effective for Blacks and below average for Whites. These results may well be attributable to site characteristics other than ethnicity, as ethnicity and site are confounded. From Table II-2, we see that of High/Scope's two sites Fort Walton Beach was 75.3% Black, while Central Ozarks was 100% White. In EDC virtually all the White children were in one of the two sites. High/Scope may also be responsible for the WRTD interaction effect ($p=.05$).

Age. We divided the age range into three categories: under 54 months, 54 to 60, over 60. There were only four tests for which the age distrubution was sufficiently balanced to allow us to carry out the analysis. Even for these four (PSI, PPV, WRTC, WRTD), it was necessary to eliminate the Oregon model, since it contained no children under 54 months of age. All tests except the PSI have significant main effects for age, and the PSI effect is almost significant ($p=.08$). For the PSI and PPV, age is negatively related to residual size (younger children gain more), while for the WRTC and WRTD it is positively related. Interaction effects were significant for the PSI ($p=.003$), WRTC ($p < .001$), and WRTD ($p=.04$). The pattern of interaction effects is difficult to interpret. Moreover, several models contain very few children over 60 months of age. If we look only at those children 60 months old

or younger, the pattern of relative program effectiveness appears fairly consistent across the two other age groups.

Prior Preschool Experience. Three of the six tests show significant main effects favoring children with no prior preschool experience. These are the PSI ($p < .001$), PPV ($p < .001$). and WRTC ($p = .025$). Recall, however, that in Chapter V we noted that the residuals may be less valid measures of program effectivenss for children with prior preschool experience than for those without. Thus an apparent prior preschool effect might really be an artifact of the way in which the residuals were computed. Significant interaction effects occur on the PSI ($p < .001$) and WRTD ($p = .01$). The PSI interaction may well be at least in part a spurious artifact of the unbalanced design. The Kansas model in particular appears to do terribly for children with prior preschool experience, but the mean for this cell is based on a sample of only seven children. Thus, although the data suggest the possibility that relative model effectiveness on the PSI is related to prior preschool experience, we cannot interpret this interaction with much confidence. Note that REC and Enablers are both more effective than average for children with prior preschool and less effective for those without.

Mother's Education

There are significant main effects for mother's education on the PSI ($p < .001$), PPV ($p < .001$), WRTC ($p=.01$), WRTD ($p=.01$), and ETS ($p=.003$). The PSI, PPV, and WRTC effects reflect a negative relationship between mother's education and residual size. There are no significant interaction effects on any of the tests.

Featherstone (1972) found generally that relative model effectiveness tended to be related to variables which describe the child at a particular stage of development rather than to permanent, unalterable characteristics. Our results generally corroborate this finding. The fixed characteristics we studied (sex, ethnicity, and mother's education) showed very few significant interaction effects. Age and preschool experience, on the other hand, yielded a fair number. These effects were not, however, easy to interpret and the unbalanced design severely limited our confidence in their validity.

Table VIII-1

RESULTS OF INTERACTION ANALYSIS

PSI

Program	Male	Female	Row Marginals*
Far West	2.79	2.63	2.71
Arizona	1.66	1.54	1.60
Bank Street	-.07	-.36	-.22
Oregon	3.79	2.26	3.03
Kansas	2.89	2.22	2.55
High/Scope	1.43	.73	1.08
Florida	1.09	1.99	1.54
EDC	2.24	.162	1.93
Pittsburgh	3.28	1.62	2.45
REC	2.59	2.36	2.47
Enablers	2.36	1.40	1.88
Control	.83	-.33	.25
NPV	1.80	.94	1.37
Column Marginals	2.05	1.43	1.74

	F	Significance
Program	8.433	<.001
Sex	11.438	<.001
Program X Sex	1.040	.41

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

PPV

Program	Male	Female		Row Marginals*
Far West	5.28	6.41		5.84
Arizona	5.42	5.00		5.21
Bank Street	5.41	5.64		5.53
Oregon	8.16	6.79		7.47
Kansas	7.15	5.75		6.45
High/Scope	3.99	3.09		3.54
Florida	6.96	4.54		5.75
EDC	5.75	6.49		6.12
Pittsburgh	4.96	6.92		5.94
REC	12.09	7.89		9.99
Enablers	4.38	4.16		4.27
Control	7.22	4.92		6.07
NPV	6.87	5.83		6.35
Column Marginals	6.44	5.65		6.35
	F	Significance		
Program	4.799	<.001		
Sex	4.052	.045		
Program X Sex	1.365	.176		

*Marginals are unweighted averages of cell means.

Table VIII-3

RESULTS OF INTERACTION ANALYSIS

WRTC

Program	Male	Female	Row Marginals*
Far West	1.40	1.65	1.52
Arizona	2.15	2.18	2.17
Bank Street	1.35	1.42	1.39
Oregon	3.86	2.79	3.33
Kansas	4.01	4.80	4.40
High/Scope	2.70	1.88	2.29
Florida	1.36	2.55	1.96
EDC	2.40	2.97	2.67
Pittsburgh	1.61	1.24	1.42
REC	1.32	1.34	1.33
Enablers	1.74	2.13	1.93
Control	.84	.05	.45
NPV	1.83	2.05	1.94
Column Marginals	2.04	2.08	2.06

	F	Significance
Program	13.03	<.001
Sex	.06	>.5
Program X Sex	1.43	.15

*Marginals are unweighted averages of cell means.

TABLE VIII-4

RESULTS OF INTERACTION ANALYSIS

WRTD

Program	Male	Female	Row Marginals*
Far West	.94	.86	.90
Arizona	1.08	1.25	1.16
Bank Street	.42	.57	.50
Oregon	2.71	2.63	2.67
Kansas	1.77	2.03	1.90
High/Scope	.74	.68	.71
Florida	.32	1.05	.69
EDC	1.17	1.35	1.26
Pittsburgh	1.28	.86	1.07
REC	.59	.82	.70
Enablers	.71	.71	.71
Control	.30	.05	.18
NPV	.84	.88	.86
Column Marginals	.99	1.06	1.02

	F	Significance
Program	37.54	<.001
Sex	1.26	.26
Program X Sex	1.72	.06

*Marginals are unweighted averages of cell means.

Table VIII-5

RESULTS OF INTERACTION ANALYSIS

ITPA

Program	Male	Female	Row Marginals*
Far West	2.43	1.36	1.90
Arizona	1.02	2.73	1.88
Bank Street	1.47	2.07	1.77
Oregon	2.15	4.36	3.26
Kansas	3.04	.94	1.99
High/Scope	1.30	0.05	.63
Florida	1.94	3.48	2.71
EDC	3.27	2.94	3.11
Pittsburgh	3.61	2.53	3.07
REC	2.81	.10	1.45
Enablers	.73	.14	.44
Control	-	-	-
NPV	2.43	3.21	2.82
Column Marginals	2.18	1.99	2.08

	F	Significance
Program	1.933	.03
Sex	.259	>.500
Program X Sex	1.341	.20

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

ETS

Program	Male	Female	Row Marginals*
Far West	1.97	1.86	1.92
Arizona	2.66	3.24	2.95
Bank Street	1.83	1.81	1.82
Oregon	4.09	2.04	3.06
Kansas	4.01	3.50	3.75
High/Scope	.21	.24	.22
Florida	1.30	2.60	1.95
EDC	1.66	1.63	1.65
Pittsburgh	3.11	1.88	2.49
REC	.07	-.05	.01
Enablers	-.30	-.06	-.18
Control	-	-	-
NPV	1.42	1.32	1.37
Column Marginals	1.84	1.67	1.75

	F	Significance
Program	8.032	<.001
Sex	.461	>.5
Program X Sex	.913	>.5

*Marginals are unweighted averages of cell means.

Table VIII-7

RESULTS OF INTERACTION ANALYSIS

PSI

Program	White	Black	Row Marginals*
Far West	2.89	1.95	2.42
Arizona	2.03	1.05	1.54
Bank Street	.34	-.34	0.00
Oregon	1.79	2.89	2.34
Kansas	3.58	1.91	2.75
High/Scope	1.50	.31	.91
Florida	1.25	1.41	1.33
EDC	2.64	1.66	2.16
Pittsburgh	-	-	-
REC	2.98	1.79	2.39
Enablers	2.65	.95	1.80
Control	-.13	.39	.13
NPV	2.11	1.38	1.74
Column Marginals	1.97	1.28	1.63

	F	Significance
Program	4.372	<.001
Ethnicity	7.865	.006
Program X Ethnicity	1.009	.43

*Marginals are unweighted averages of cell means.

Table VIII-8

RESULTS OF INTERACTION ANALYSIS

PPV

Program	White	Black	Row Marginals*
Far West	4.45	9.65	7.05
Arizona	4.01	7.38	5.69
Bank Street	3.84	5.82	4.83
Oregon	-.50	8.26	3.88
Kansas	4.64	8.00	6.32
High/Scope	2.08	4.17	3.13
Florida	1.76	6.86	4.31
EDC	4.69	6.82	5.75
Pittsburgh	-	-	-
REC	5.21	8.48	6.84
Enablers	2.13	4.67	3.41
Control	4.89	7.52	6.20
NPV	4.01	6.19	5.10
Column Marginals	3.43	6.99	5.21

	F	Significance
Program	2.495	.005
Ethnicity	54.395	<.001
Program X Ethnicity	1.392	.17

*Marginals are unweighted averages of cell means.

Table VIII-9

RESULTS OF INTERACTION ANALYSIS

WRTC

Program	White	Black	Row Marginals*
Far West	1.44	1.62	1.53
Arizona	2.27	2.07	2.17
Bank Street	1.37	1.39	1.38
Oregon	5.08	2.74	3.91
Kansas	4.29	4.39	4.34
High/Scope	3.62	.39	2.00
Florida	2.26	1.29	1.77
EDC	1.61	3.09	2.35
Pittsburgh	-	-	-
REC	.51	1.19	.85
Enablers	2.25	1.56	1.91
Control	.20	.17	.19
NPV	2.17	1.31	1.74
Column Marginals	2.26	1.77	2.01

	F	Significance
Program	11.485	<.001
Ethnicity	6.166	.01
Program X Ethnicity	3.525	<.001

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

WRTD

Program	White	Black		Row Marginals*
Far West	.87	1.09		.98
Arizona	1.26	1.03		1.14
Bank Street	.30	.54		.42
Oregon	2.18	2.77		2.48
Kansas	1.79	1.94		1.86
High/Scope	.97	.19		.58
Florida	.65	.68		.66
EDC	1.69	1.16		1.43
Pittsburgh	-	-		-
REC	.72	.61		.67
Enablers	.77	.60		.69
Control	.22	.05		.14
NPV	.80	.85		.82
Column Marginals	1.02	.96		.99
	F	Significance		
Program	23.654	<.001		
Ethnicity	.564	.45		
Program X Ethnicity	1.807	.05		

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

ITPA

Program	White	Black	Row Marginals*
Far West	1.04	3.02	2.03
Arizona	1.55	2.20	1.88
Bank Street	1.73	1.82	1.77
Oregon	3.45	3.84	3.65
Kansas	.90	3.19	2.05
High/Scope	-.73	2.38	.83
Florida	-.10	3.83	1.87
EDC	2.28	3.40	2.84
Pittsburgh	-	-	-
REC	.08	1.07	.58
Enablers	.56	.33	.44
Control	-	-	-
NPV	2.14	3.67	2.90
Column Marginals	1.17	2.61	1.89

	F	Significance
Program	1.175	.30
Ethnicity	6.669	.01
Program X Ethnicity	.486	.5

*Marginals are unweighted averages of cell means.

Table VIII-12

RESULTS OF INTERACTION ANALYSIS

ETS

Program	White	Black	Row Marginals*
Far West	1.84	3.74	2.79
Arizona	2.90	3.43	3.16
Bank Street	.50	2.17	1.33
Oregon	1.85	3.84	2.85
Kansas	2.98	4.55	3.77
High/Scope	.36	.33	.35
Florida	.95	2.21	1.58
EDC	1.42	1.82	1.62
Pittsburgh	-	-	-
REC	-1.79	.70	-.55
Enablers	-.62	-.13	-.38
Control	-	-	-
NPV	1.10	1.68	1.39
Column Marginals	1.05	2.21	1.63

	F	Significance
Program	4.023	.001
Ethnicity	7.435	.007
Program X Ethnicity	.327	>.5

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

PSI

Program	Under 54	54-60	Over 60	Row Marginals*
Far West	3.37	2.93	1.07	2.46
Arizona	2.33	1.52	1.25	1.70
Bank Street	.97	-.46	-1.35	-.28
Oregon	-	-	-	-
Kansas	2.35	2.60	4.18	3.04
High/Scope	1.40	.58	1.51	1.16
Florida	2.16	2.68	.78	1.87
EDC	1.23	1.33	2.29	1.62
Pittsburgh	2.96	1.90	4.09	3.00
REC	3.02	2.27	-2.45	.95
Enablers	1.75	1.32	2.54	1.87
Control	.29	.24	-.11	.14
NPV	1.95	1.45	1.08	1.49
Column Marginals	1.98	1.53	1.24	1.58

	F	Significance
Program	4.734	<.001
Age	2.586	.08
Program X Age	2.066	.003

*Marginals are unweighted averages of cell means.

Table VIII-14

RESULTS OF INTERACTION ANALYSIS

PPV

Program	Under 54	54-60	Over 60	Row Marginals*
Far West	5.72	6.84	2.34	4.97
Arizona	4.78	5.51	5.23	5.17
Bank Street	5.75	6.42	4.64	5.60
Oregon	-	-	-	-
Kansas	8.61	5.13	4.22	5.99
High/Scope	3.69	5.46	1.24	3.46
Florida	7.49	7.98	3.97	6.48
EDC	5.31	5.78	6.60	5.90
Pittsburgh	7.71	3.84	2.79	4.78
REC	13.57	6.56	10.97	10.37
Enablers	4.41	5.32	3.05	4.26
Control	6.67	4.18	6.57	5.81
NPV	6.47	5.96	6.71	6.38
Column Marginals	6.68	5.75	4.86	5.76

	F	Significance
Program	2.819	.002
Age	3.249	.04
Program X Age	1.015	.44

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

WRTC

Program	Under 54	54-60	Over 60	Row Marginals*
Far West	2.07	1.29	1.57	1.64
Arizona	1.13	2.28	2.76	2.06
Bank Street	.26	1.12	2.92	1.43
Oregon	-	-	-	-
Kansas	3.75	4.71	5.35	4.61
High/Scope	.30	1.05	5.34	2.23
Florida	1.46	1.95	2.09	1.84
EDC	2.80	3.36	2.48	2.88
Pittsburgh	1.15	1.64	2.99	1.93
REC	1.38	1.33	.81	1.17
Enablers	1.24	1.46	3.28	1.95
Control	.13	1.26	.13	.51
NPV	.42	1.45	3.17	1.68
Column Marginals	1.34	1.90	2.73	1.99

	F	Significance
Program	8.387	<.001
Age	16.284	<.001
Program X Age	2.437	<.001

*Marginals are unweighted averages of cell means.

Table VIII-16

RESULTS OF INTERACTION ANALYSIS

WRTD

Program	Under 54	54-60	Over 60	Row Marginals*
Far West	.86	.93	.87	.89
Arizona	.82	1.24	1.30	1.12
Bank Street	.21	.41	.92	.51
Oregon	-	-	-	-
Kansas	1.96	1.89	1.23	1.69
High/Scope	.25	.62	1.17	.68
Florida	.70	.64	.67	.67
EDC	.75	1.18	1.49	1.14
Pittsburgh	1.07	1.01	2.19	1.42
REC	.73	.68	.49	.64
Enablers	.43	.74	.92	.70
Control	.09	.20	.56	.28
NPV	.39	.69	1.26	.78
Column Marginals	.69	.85	1.09	.88

	F	Significance
Program	8.332	<.001
Age	8.382	<.001
Program X	1.588	.04

*Marginals are unweighted averages of cell means.

Table VIII-17

RESULTS OF INTERACTION ANALYSIS

PSI

Program	No Prior Preschool	Prior Preschool	Row Marginals*
Far West	2.82	2.03	2.43
Arizona	2.18	.40	1.29
Bank Street	.37	-.93	-.28
Oregon	3.63	.87	2.25
Kansas	3.02	-2.77	.12
High/Scope	1.26	.37	.81
Florida	1.77	2.39	2.08
EDC	1.55	2.21	1.88
Pittsburgh	3.06	.77	.192
REC	2.43	2.78	2.60
Enablers	1.77	2.39	2.08
Control	.54	-.58	-.02
NPV	1.80	.54	1.17
Column Marginals	2.01	.64	1.32

	F	Significance
Program	4.699	<.001
PSEXP	31.179	<.001
Program X PSEXP	3.598	<.001

*Marginals are unweighted averages of cell means.

Table VIII-18

RESULTS OF INTERACTION ANALYSIS

PPV

Program	No Prior Preschool	Prior Preschool		Row Marginals*
Far West	6.07	3.84		4.96
Arizona	5.60	4.37		4.99
Bank Street	6.32	4.59		5.46
Oregon	8.28	4.00		6.14
Kansas	6.82	3.73		5.28
High/Scope	3.63	3.28		3.46
Florida	6.28	1.98		4.13
EDC	5.17	7.13		6.15
Pittsburgh	6.39	3.82		5.10
REC	9.71	11.81		10.76
Enablers	4.91	.91		2.91
Control	7.04	3.40		5.22
NPV	6.78	5.18		5.98
Column Marginals	6.38	4.47		5.43
	F	Significance		
Program	3.890	<.001		
PSEXP	13.216	<.001		
Program X PSEXP	1.255	.24		

*Marginals are unweighted averages of cell means.

Table VIII-19

RESULTS OF INTERACTION ANALYSIS

WRTC

Program	No Prior Preschool	Prior Preschool		Row Marginals*
Far West	1.54	1.35		1.45
Arizona	2.03	2.46		2.24
Bank Street	1.25	1.55		1.40
Oregon	3.56	2.45		3.00
Kansas	4.53	2.29		3.41
High/Scope	2.53	1.20		1.86
Florida	1.87	2.41		2.14
EDC	3.14	2.32		2.73
Pittsburgh	1.64	.74		1.19
REC	1.37	1.14		1.26
Enablers	1.78	2.81		2.29
Control	.65	-.56		.05
NPV	1.93	1.54		1.74
Column Marginals	2.14	1.67		1.90
	F	Significance		
Program	5.452	<.001		
PSEXP	5.054	.025		
Program X PSEXP	1.449	.14		

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

WRTD

Program	No Prior Preschool	Prior Preschool	Row Marginals*
Far West	.93	.75	.84
Arizona	1.15	1.17	1.16
Bank Street	.41	.61	.51
Oregon	2.81	2.08	2.44
Kansas	1.93	1.32	1.62
High/Scope	.69	.82	.76
Florida	.74	.05	.40
EDC	1.02	1.53	1.27
Pittsburgh	1.21	.70	.96
REC	.61	1.17	.89
Enablers	.72	.66	.69
Control	18	.11	.15
NPV	.82	.93	.87
Column Marginals	1.02	.91	.97

	F	Significance
Program	16.137	<.001
PSEXP	1.641	.20
Program X PSEXP	2.166	.01

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

ITPA

Program	No Prior Preschool	Prior Preschool	Row Marginals*
Far West	1.96	1.24	1.60
Arizona	1.55	2.43	1.99
Bank Street	1.49	2.20	1.85
Oregon	3.06	3.31	3.19
Kansas	2.10	2.17	2.13
High/Scope	.70	.60	.65
Florida	2.43	3.19	2.81
EDC	3.57	2.67	3.12
Pittsburgh	3.87	.20	2.04
REC	1.69	.22	.96
Enablers	.41	.53	.47
Control	-	-	-
NPV	2.67	3.43	3.05
Column Marginals	2.13	1.85	1.99

	F	Significance
Program	1.192	.29
PSEXP	.304	> .5
Program X PSEXP	.560	> .5

*Marginals are unweighted averages of cell means.

Table VIII-22

RESULTS OF INTERACTION ANALYSIS

ETS

Program	No Prior Preschool	Prior Preschool	Row Marginals*
Far West	1.79	2.53	2.16
Arizona	3.32	1.87	2.60
Bank Street	1.95	1.70	1.82
Oregon	3.19	3.30	3.24
Kansas	3.86	3.40	3.63
High/Scope	.12	.63	.38
Florida	1.61	3.09	2.35
EDC	1.61	1.68	1.65
Pittsburgh	2.42	2.92	2.67
REC	-.10	.56	.23
Enablers	-.01	-1.00	-.51
Control	-	-	-
NPV	1.53	1.20	1.37
Column Marginals	1.77	1.82	1.80

	F	Significance
Program	4.858	$<.001$
PSEXP	.021	$>.5$
Program X PSEXP	.501	$>.5$

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

PSI

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	4.07	2.57	2.43	3.02
Arizona	2.30	1.33	1.36	1.66
Bank Street	-.15	-.25	-.26	-.22
Oregon	4.11	3.07	2.12	3.10
Kansas	3.23	1.63	2.57	2.47
High/Scope	1.82	1.21	.43	1.16
Florida	2.01	2.21	.26	1.49
EDC	2.97	1.40	.67	1.68
Pittsburgh	2.96	2.51	2.50	2.66
REC	2.05	3.03	2.26	2.45
Enablers	3.65	1.46	.96	2.02
Control	.76	-.42	.31	.22
NPV	2.27	1.11	.75	1.37
Column Marginals	2.46	1.60	1.26	1.78

	F	Significance
Program	8.578	$< .001$
Mother's Education	14.160	$< .001$
Program X Mother's Education	.97	$> .5$

*Marginals are unweighted averages of cell means.

RESULTS OF INTERACTION ANALYSIS

PPV

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	9.00	6.00	4.53	6.51
Arizona	6.40	4.04	5.29	5.24
Bank Street	5.33	5.74	5.51	5.53
Oregon	10.88	7.87	4.17	7.64
Kansas	7.78	6.20	5.77	6.58
High/Scope	6.43	3.92	.93	3.76
Florida	5.44	6.79	5.47	5.90
EDC	7.58	5.37	4.70	5.88
Pittsburgh	6.94	9.48	4.14	6.85
REC	13.16	9.93	8.56	10.55
Enablers	8.62	4.00	1.44	4.69
Control	8.83	5.81	4.33	6.34
NPV	7.13	5.80	6.06	6.33
Column Marginals	7.96	6.23	4.69	6.29

	F	Significance
Program	4.864	< .001
Mother's Education	21.626	< .001
Program X Mother's Education	1.354	.12

*Marginals are unweighted averages of cell means.

Table VIII-25

RESULTS OF INTERACTION ANALYSIS

WRTC

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	2.05	1.18	1.61	1.61
Arizona	3.06	1.67	1.91	2.21
Bank Street	1.78	1.02	1.44	1.41
Oregon	3.76	3.36	2.97	3.36
Kansas	4.37	3.68	4.63	4.23
High/Scope	2.38	2.04	2.54	2.32
Florida	2.41	2.12	.80	1.78
EDC	3.13	2.52	2.31	2.65
Pittsburgh	2.31	1.16	1.37	1.62
REC	1.13	1.89	.96	1.33
Enablers	2.42	1.81	1.71	1.98
Control	.48	.88	.16	.49
NPV	2.26	1.78	1.91	1.92
Column Marginals	2.43	1.93	1.85	2.07

	F	Significance
Program	10.662	<.001
Mother's Education	5.027	.01
Program X Mother's Education	.664	>.5

*Marginals are unweighted averages of cell means.

Table VIII-26

RESULTS OF INTERACTION ANALYSIS

WRTD

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	.79	.68	1.11	.86
Arizona	1.38	1.09	1.05	1.18
Bank Street	.78	.37	.44	.53
Oregon	2.76	2.95	2.43	2.71
Kansas	2.02	1.49	1.94	1.82
High/Scope	.79	.59	.75	.71
Florida	.71	.71	.56	.66
EDC	1.57	1.15	.90	1.21
Pittsburgh	.92	.90	1.24	1.02
REC	.78	.85	.53	.72
Enablers	.81	.63	.71	.72
Control	.13	-.13	.39	.13
NPV	.85	.78	.93	.85
Column Marginals	1.10	.93	1.00	1.01

	F	Significance
Program	10.662	<.001
Mother's Education	5.027	.01
Program X Mother's Education	.664	>.5

*Marginals are unweighted averages of cell means.

Table VIII-27

RESULTS OF INTERACTION ANALYSIS

ITPA

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	3.57	.60	2.10	2.09
Arizona	3.01	.91	1.84	1.92
Bank Street	2.33	1.92	1.38	1.87
Oregon	2.38	2.57	3.89	2.95
Kansas	3.21	3.71	.48	2.47
High/Scope	.58	1.70	.11	.79
Florida	1.03	3.16	3.56	2.58
EDC	3.23	1.98	4.08	3.10
Pittsburgh	3.90	2.28	3.05	3.08
REC	-.96	.87	2.66	.86
Enablers	1.65	-.45	.55	.58
Control	-	-	-	-
NPV	1.84	3.38	3.32	2.85
Column Marginals	2.15	1.89	2.25	2.10

	F	Significance
Program	1.644	.08
Mother's Education	.278	> .5
Program X Mother's Education	.966	> .5

*Marginals are unweighted averages of cell means.

Table VIII-28

RESULTS OF INTERACTION ANALYSIS

ETS

Program	Under 10	10 or 11	Over 11	Row Marginals*
Far West	1.55	3.04	1.34	1.97
Arizona	3.22	3.63	2.27	3.04
Bank Street	2.84	1.38	1.43	1.88
Oregon	3.81	4.34	2.05	3.40
Kansas	3.03	5.59	3.39	4.00
High/Scope	.99	.61	-.37	.41
Florida	2.03	3.20	.54	1.92
EDC	2.10	1.88	.57	1.51
Pittsburgh	2.54	2.88	2.40	2.61
REC	-.65	.24	.13	-.09
Enablers	.31	-.54	-.19	-.14
Control	-	-	-	-
NPV	1.52	1.68	.89	1.36
Column Marginals	1.94	2.83	1.21	1.82

	F	Significance
Program	8.304	$< .001$
Mother's Education	6.184	.003
Program X Mother's Education	.732	$> .50$

*Marginals are unweighted averages of cell means.

Chapter IX

MAJOR CONCLUSIONS

Throughout this report we have focused on three major questions. Each of our four analytical approaches has provided evidence bearing on these questions. In this chapter we summarize the evidence and present conclusions.

1. To what extent does a Head Start experience accelerate the rate at which disadvantaged pre-schoolers acquire cognitive skills?

Our evidence here is of two types. Each analysis provides a comparison between the performance of Head Start and Control children for the six tests taken by both. The residual analysis provides a direct estimate of the amount of growth attributable to Head Start over and above what the child would otherwise have achieved. On the Preschool Inventory (PSI), WRAT Copying Marks (WRTC), WRAT Recognizing Letters (WRTR), WRAT Naming Letters (WRTN), and WRAT Reading Numbers (WRTD), the Head Start children (both PV and NPV) did substantially better than the Control

children. On the Peabody Picture Vocabulary Test (PPV), Head Start and Control performances were comparable.

From the residual analysis, we found that the growth rates for Head Start children on all six tests considered increased substantially. For the PSI, the growth rate increased by about 50%, for the PPV by about 100%, for the WRTC 200%, for the WRTD 300%. For the ITPA Verbal Expression the gain was approximately 100%, and for the ETS Enumeration about 75%. Moreover, except for the PPV, the average residuals for the Controls were near zero. For the PPV, the average residual for the Controls was close to that for both PV and NPV Children. In conclusion it seems fair to say that:

> In terms of a wide variety of cognitive skills, Head Start is effective in accelerating the growth rate of disadvantaged preschoolers.

We do not know, of course, whether the changes wrought are permanent and can be built upon. Has Head Start simply made the child a bit more aware of certain specific things at a particular point in his life, or has it altered him more profoundly and increased his capacity to learn. The answer is probably to some extent unique to each child.

There is perhaps a certain pessimism at present about our ability to effect desirable social change. Head Start is only part of a child's life experience over a short period of his life. If even this relatively minor effort to alter the child's environment can have substantial, measurable impact, then there is reason to hope that more extensive societal efforts may have profound and lasting effects.

2. Are Planned Variation models, simply by virtue of sponsorship, more effective than ordinary non-sponsored Head Start programs?

There are two reasons why we might expect PV programs to be generally more effective than NPV programs. First, they involve the expenditure of substantially more money per child. The nature of these expenditures is detailed by McMeekin (1973). Second, we might expect a great deal of effort on the part of sponsors to ensure that their approaches perform optimally. In fact, if we were to find PV programs generally superior to NPV programs, we might be concerned that this was the result of special effort expended in the competitive experimental situation which might disappear when the programs were routinely implemented.

Smith (1973) found no overall difference in performance between PV and NPV programs for the 1970-71 cohort. Our analyses strongly support this finding. There are no clear differences between the 28 PV and the 12 NPV sites on any test. The general picture which emerges is that:

> Relative to the condition of no preschool program, the effects of Head Start programs are quite homogeneous, with no systematic differences between *sponsored and non-sponsored programs.*

3. Are some PV models particularly effective at imparting certain skills.

We have results for each of eight tests in our battery on at least three of the four analyses. Table IX-1 presents an overall summary of inter-model comparisons. Each analysis has its own assumptions and implicit or explicit measure of program effectiveness. Thus we would not expect the different analyses to yield identical results. We would, however, be concerned about large discrepancies. An effect which shows up consistently over several analyses is more likely to be real, and not simply an artifact of some mathematical manipulations. In Table IX-1 we have given a model + on a particular

test if it achieves + on at least two of three analyses. We have given a ++ only for models with at least + on all analyses and at least one ++. The same standards apply to negative effects. These standards are of course arbitrary, and the reader is free to apply his own standards to summarize the results in Chapters IV through VII.

Smith (1973) found a rather small number of examples of programs which were especially effective at promoting skills. He also reached the tentative conclusion that "differential model effects are more easily discerned if the outcome measure taps specific rather than general cognitive growth." Our results tend to corroborate these findings. Only 22 "effects" are cited for the eight tests. Three of our tests (WRTR, WRTN, WRTD) measure very specific academic skills. Two others (WRTC, ETS) measure skills which are somewhat more general but relatively easily taught. Three tests (PSI, PPV, ITPA) measure general skill relatively difficult to teach in a preschool program. These three account for only 6 of the 22 effects. Moreover, 3 of these effects are on the ITPA, which in terms of reliability and validity is the most questionable test in our battery. The PPV, which is probably our most general measure, shows no effects. The 32-item PSI we have used does appear to be possibly more sensitive to

program differences than the 64-item version used in 1970-71. The clearest and most dramatic examples of special program effectiveness are Kansas on the WRTC and Oregon and Kansas on the WRTD.

We mentioned in Chapter I the hypothesis that the "academic" models (Oregon, Kansas, Pittsburgh), which consciously emphasize the acquisition of academic skills, would be overall more effective than the other models. From Table IX-1, we see that of the 17 positive effects noted, 12 are for these three models. Moreover, none of the three received a - for any test on any of the analyses. Kansas has four ++'s and Oregon three. The only other model which can lay claim to better than average overall performance is Arizona, with three positive effects and no negatives. Our conclusions in terms of inter-model comparisons can be summarized in the following statements.

a) Head Start programs are quite homogeneous in their ability to promote general cognitive development.

b) No Head Start program is of above average effectiveness for all of our measures.

c) Oregon and Kansas appear to be overall particularly effective in imparting specific academic skills.

d) Arizona and Pittsburgh may be overall particularly effective in imparting specific academic skills.

e) No program appears to be overall particularly ineffective.

Table IX-1

OVERALL SUMMARY OF RELATIVE MODEL EFFECTIVENESS*

++ Indicates model appears to be highly effective.
+ Indicates evidence for above average effectiveness.
- Indicates evidence for below average effectiveness.
-- Indicates model appears to be highly ineffective.

Model	PSI	PPV	WRTC	WRTR	WRTN	WRTD	ITPA	ETS
Far West	+							
Arizona				+	+			++
Bank Street	-							
Oregon	++		+		+	++		++
Kansas			++	++		++		++
High/Scope							-	-
Florida								
EDC					+			
Pittsburgh				+			+	+
Enablers							-	-

*REC not included because with only one site we felt it unfair to draw any conclusions.

REFERENCES

Abt Associates (1973). Project Follow Through. Interim Report, An Analysis of Selected Data 1969-72. Cambridge, Massachusetts.

Bereiter, C., Engelmann, S., Osborn, J. and Redford, P. (1965). An Academically Oriented Pre-School for Culturally Deprived Children, Paper presented at the American Educational Research Convention in Chicago.

Campbell, Donald T. and Erlebacher, A. (1970). How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful. In J. Hellmuth (Ed.), Compensatory Education: A National Debate. Vol. III. Disadvantaged Child. New York: Brunner-Mazel.

Cicerelli, G., et al. (1969). 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.

Cochran, W.G. (1968). Errors of Measurement in Statistics. Technometrics, 10, 637-660.

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

Datta, L. (1969). A Report on Evaluation Studies of Project Head Start. Paper presented at the 1969 American Psychological Association Convention, Washington, D.C.

De Gracie, J.S. and Fuller, W.A. (1972). Estimation of the Slope and Analysis of Covariance When the Concomitant Variables is Measured with Error. JASA 67, 930-37.

Draper, N.R., Smith, H. (1966): Applied Regression Analysis. New York; Wiley,.

Featherstone, H. (1972). Cognitive Effects of Pre-School Programs on Different Types of Children. Huron Institute, Cambridge, Massachusetts. (Preliminary version).

Gray, S.W. and Klaus, R.A. (1963). The Early Training Project: Interim Report, Murfreesboro City School and George Peabody College for Teachers, Murfreesboro, Tennessee.

Holmes, D. and Holmes, M.B. (1966). Evaluation of Two Associated YM-YWCA Head Start Programs of New York City. Associated YM-YWCA's of New York City. Final Report, New York.

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

Lord, F.M. (1960). Large - Sample Covariance Analysis When The Control Variable Is Fallible. Journal of the American Statistical Association, 55, 309-321.

Lord, F.M. (1967). A Paradox in the Interpretation of Group Comparisons. Psychological Bulletin, 68, 304-305.

Lord, F.M. and Novick, M.R. (1968). Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley.

Lukas, C. and Wollheb, C. (1972). Implementation of Head Start Planned Variation. Huron Institute, Cambridge, Mass. (Preliminary version).

Maccoby, E. E. and Zellner, M. (1970). Experiments in Primary Education: Aspects of Project Follow-Through. New York: Harcourt Brace.

McMeekin, R.W. (1973). Synthesized Estimates of the Costs of Head Start Planned Variation Models. Huron Institute, Cambridge, Mass.

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

Porter, Andrew G. (1968). The Effects of Using Fallible Variables in the Analysis of Covariance. Ph. D. Dissertation, University of Wisconsin, June, 1967. (University Microfilms, Ann Arbor, Michigan).

Porter, Andrew G. (1971). How Errors of Measurement Affect ANOVA, Regression Analyses. Paper presented at the 1971 AERA Convention.

Rubin, D.B. (1973). Matching to remove bias in observational studies. Biometrics 29, 159-83.

Searle, S.R. (1971). Linear Models. New York: Wiley.

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

Smith, M.S. and Bissell, J.S. (1970). Report analysis: the impact of Head Start. Harvard Educational Review, Vol. 40, No. 1, pp. 51-104.

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

Snedecor, G.W. and Cochran, W.G. (1972). Statistical Methods. Iowa State University Press, Ames.

Stanford Research Institute (1972). Implementation of Head Start Planned Variation Testing and Data Collection Effort. Menlo Park, California.

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

Stroud, T.W.F. (1972). Comparing Conditional Means and Variances in a Regression Model with Measurement Errors of Known Variances. Journal of American Statistical Association, 67, 407-412.

Tukey, J.W. (1970). Exploratory Data Analysis. Addison-Wesley, Reading.

University of Kansas Support and Development Center for Behavior Analysis Follow Through (1972). Final Report on Behavior Analysis Planned Variation Head Start.

Walker, D. (1972). Socio-Emotional Measures for Preschool and Kindergarten Children: A Summary and Critique. Unpublished qualifying paper. Harvard University Graduate School of Education. Cambridge, Mass.

Walker, D., Bane, M., and Bryk, T. (1973). The Quality of the Head Start Planned Variation Data, Huron Institute, Cambridge, Mass.

Weikart, D.P., Kamii, C.K., and Radin, N. (1964). Perry Preschool Progress Report, Ypsilanti Public Schools, Ypsilanti, Michigan.

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

Wolff, M. and Stein, A. (1966). Six Months Later. Study I. A Comparison of Children Who Had Head Start, Summer 1965, with Their Classmates in Kindergarten, A Case Study of the Kindergartens in Four Public Elementary Schools, O.C.D. Project 141-61, Yeshiva University, New York.

DESCRIPTION OF VARIABLES

This appendix describes the child, classroom, teacher, and outcome variables used in the preceding analyses. Where multiple forms of essentially the same variable existed, they are noted. Variable names as referred to elsewhere are capitalized. Where categories are given as unindexed lists (White/Black/Mexican-American), the codes were the ascending integers 1, 2, 3, etc. Related variables are grouped together.

I. Child characteristics, demographic and background. (Source: Classroom Information Form)

AGE	child's age in months as of October 1, 1971
AGE1	for each test, child's age in months as of fall test date.
AGED	for each test, interval in months between fall and spring testing. Note: a separate value of AGE1 and AGED was computed for each test, e.g. AGE1(6) for the WRAT-D.
DAYSAB	days absent during the Head Start year
ETHWHITE	White ethnicity. 1=white, 0=not White
ETHBL	Black ethnicity. 1=Black, 0=not Black Note: for the residual analysis sample, children other than White, Black, Mexican-American and Puerto Rican were deleted.
FLANG	first language spoken in the child's home. 0=English, 1=not English

PSEXP	preschool experience. some non-Head Start/ none/some Head Start
PS	preschool experience recode. 1=some, 0=none
PSMNTHS	months of Head Start preschool experience if PSEXP was 3
SEX	child's sex. 1=female, 0=male

II. Child Household Characteristics
(Source: Classroom Information Form -- MOMED, FAMINC, SEXHH, HHSIZE
Parent Information Form

FAMINC	annual family income in $100's. 0-98, or 99 if $9,900 or more
HHSIZE	total number of persons resident in household
MOMED	mother's education in school years. 0-16, or 17 if any graduate work
PIF1	child watches Sesame Street. 1=yes, 0=no
PIF2	how often each week. 5+ times/4 or 5/2 or 3/ 1 or less
PIF3	parents also watch. always/usually/sometimes/ hardly
PIF4 to PIF17	materials available for child at home. 1=yes, 0=no. blackboard, chalk, colored paper, scissors, crayons, coloring books, paints, clay, other arts and crafts, music equipment, alphabet/number cards, games, puzzles, children's records respectively.
PIF18	parents read to child. 1=yes, 0=no
PIF19	how often per week. less than once/once/several times/daily

PIF20	how far parents wish child to go in school. High School/College/Graduate School
PIF21	how far parents expect child to go in school. Same
PIF22 to PIF30	household attributes. 1=present, 0=not present auto, black and white TV, color TV, encyclopedia, dictionary, clothes washer, vacuum, hifi, telephone respectively
PIF40	child likes Head Start. very much/some/not at all
PIF41	parents satisfied with Head Start. very/fairly/not
SEXHH	sex of household head, father if present. 1=female, 0=male

III. Classroom and Teacher Characteristics (non-control children)
(Source: Teacher Information Form
Rating Forms

CYTPE	classroom type. PV/non-PV/control
SITE1	SRI site code
SITE2	Huron site code
SPONSOR	site sponsor. 2-27=PV sponsors, 28=control, 30=non-PV
TIF4	fall implementation rating of PV classroom by sponsor. 0-9, 9 highest
TIF5	spring same
TIF6	fall rating by site director. 0-9, 9 highest
TIF7	spring same

TIF8	class watches Sesame Street. 1=yes, 0=no
TIF9	how often. days per week
TIF10	number of field trips taken. 0-10, or 11=more than 10
TIF11	teacher sex. 1=male, 0=female
TIF12	teacher ethnicity. AM. Indian/Black/Oriental/White/Mexican-American/Puerto Rican/Cuban/other Spanish/Portuguese
TIF13	teacher marital status. single/married/widowed-divorced-separated
TIF14	teacher has children. 1=yes, 0=no
TIF15	teacher neighborhoood similar to center. 1=yes, 0=no
TIF16	teacher age. years
TIF17	teacher education. 0-16 years, or 17=more than 16
TIF18	teacher certification. none/temporary/regular
TIF19	teacher rating of classroom differs from sponsor's model. much/some/none
TIF20	sponsor changed teacher's ways. very much/much
TIF21	teacher use of model given choice. use/change some/change most/not use

IV. Outcome Measures.

These are more fully described in the body of the report and in an earlier report, "The Quality of the Head Start Planned Variations Data". Tester assigned validity was used as a criterion for accepting a score.

PSI	Preschool Inventory
PPVT	Peabody Picture Vocabulary Test
WRAT C/R/N/D	Wide Range Achievement Test - Copying marks/Recognizing Letters/Naming Letters/Reading Numbers subtests respectively

ITPA	Illinois Test of Psycholinguistic Abilities Verbal Expression Subtest
ETS	Educational Testing Service Enumeration Test

Appendix B

SITE MEAN RELIABILITIES

In this appendix we explain the basis for the estimates of site-mean reliabilities quoted in Chapter IV. Shaycroft (1962) provides a formula for the reliability of the means of groups of size n which may be written as

$$r_G = 1 - \frac{(1-r_I)}{nB}$$

where r_G and r_I are the reliabilities for the groups and for individuals, and B is the proportion of individual test variance which lies between groups. We have carried out an unweighted-means analysis of variance with sites as factors for each of our eight tests. The proportion of variance which is between sites and the harmonic mean n* of the site sample sizes have been recorded in Table C-1.

A useful and convenient way of summarizing the information provided by Shaycroft's formula for our tests is in terms of the quantity

$$\gamma = \frac{r_G - r_I}{1 - r_I} = \frac{nB - 1}{nB}$$

We can think of γ as the proportion of the gap between r_I and 1 which is closed by aggregation to the group level. Let γ^* be the value of γ for sites of size n*. Then γ^* provides a rough idea of the improvement afforded by aggregation to the site level for each of our tests. These values are also displayed

Information Used to Estimate Site Mean Reliabilities

Test	B	n*	Y*
PSI	22.67	77	.94
PPV	19.93	80	.94
WRTC	15.63	78	.92
WRTR	9.43	78	.87
WRTN	4.37	78	.71
WRTD	6.79	78	.81
ITPA	26.55	38	.90
ETS	28.78	36	.90

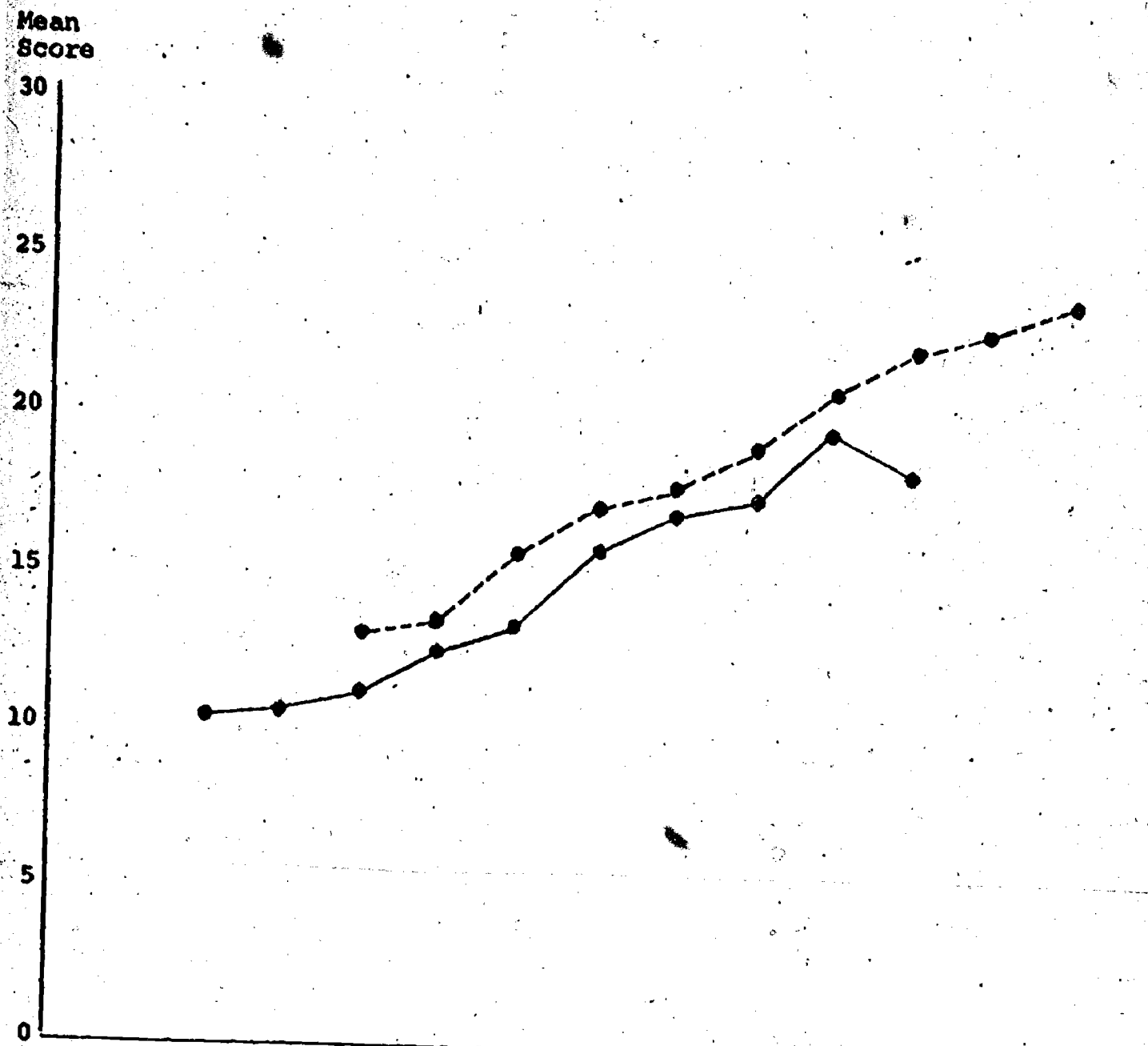
Appendix C

RESULTS OF GRAPHICAL ANALYSIS

This appendix presents graphs of fall and spring test scores versus age for each of the eight tests in our battery. The sample is broken out into children with no prior preschool experience and those with some prior preschool experience.

PSI - FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE

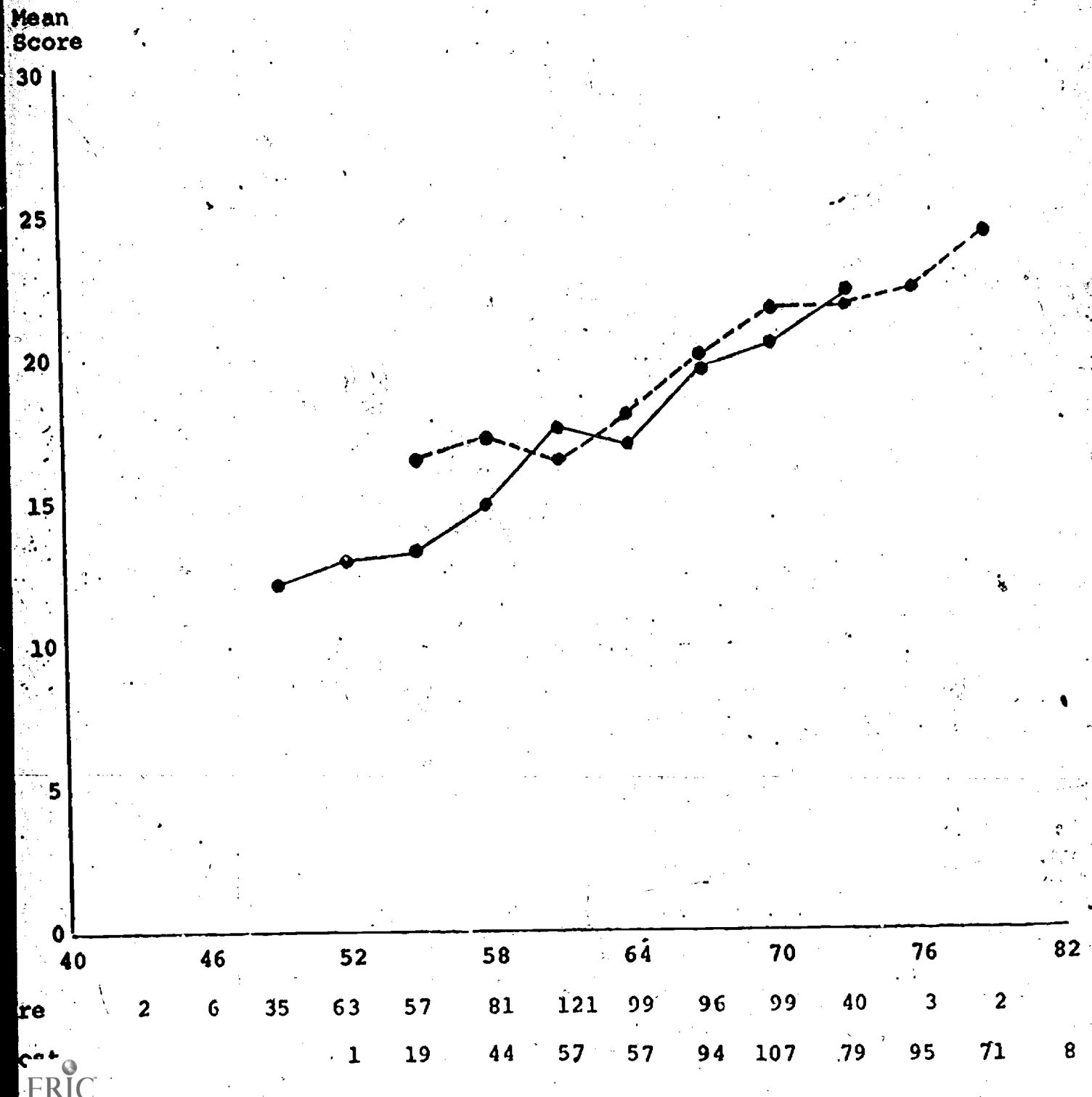
Fall ———
Spring - - - - -



re	4	16	63	207	374	397	368	257	162	165	119	52	2	1
ost		3	5	6	19	94	284	311	313	300	186	146	124	98	10

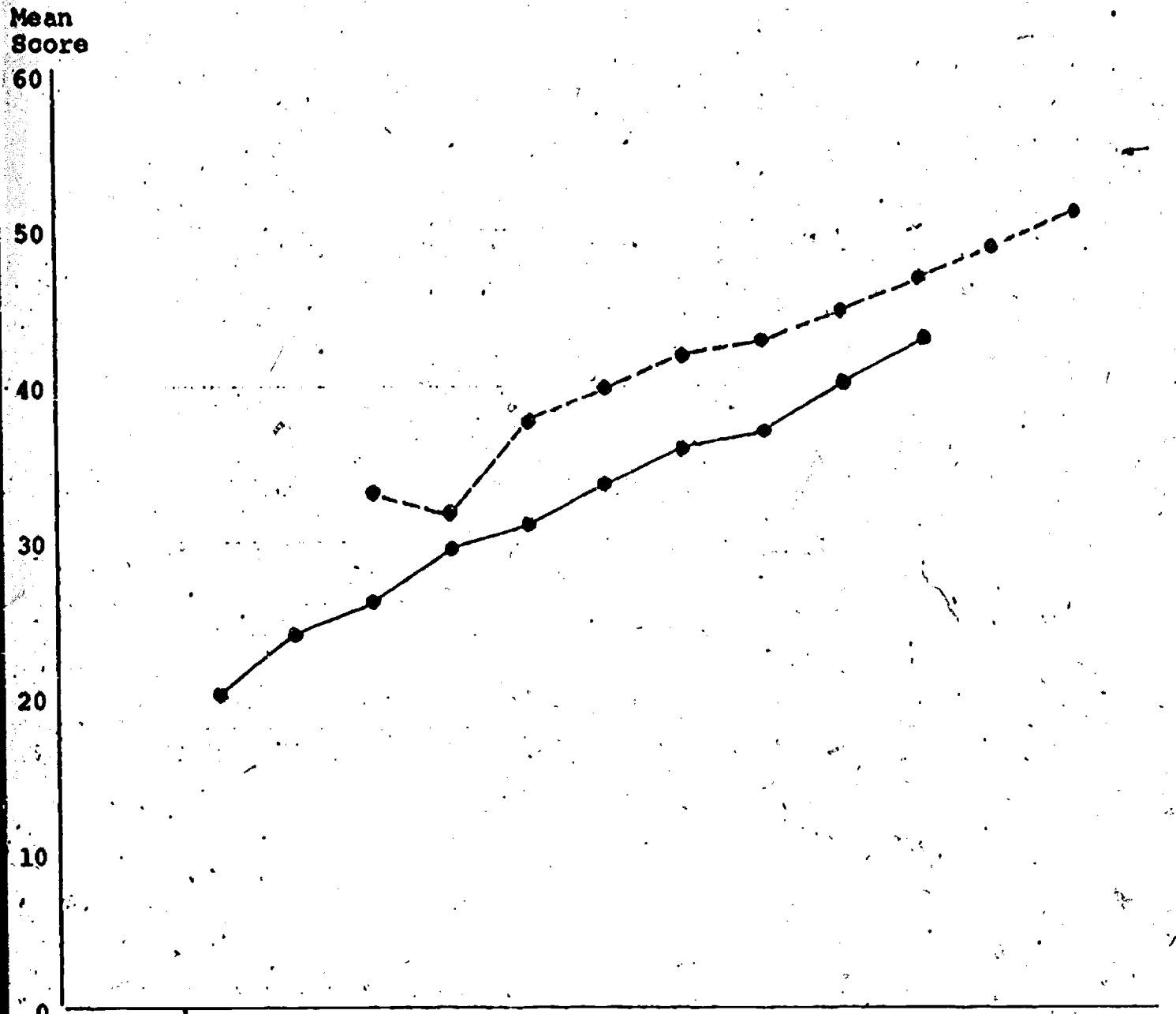
PSI - FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE

Fall ——
Spring - - - - -



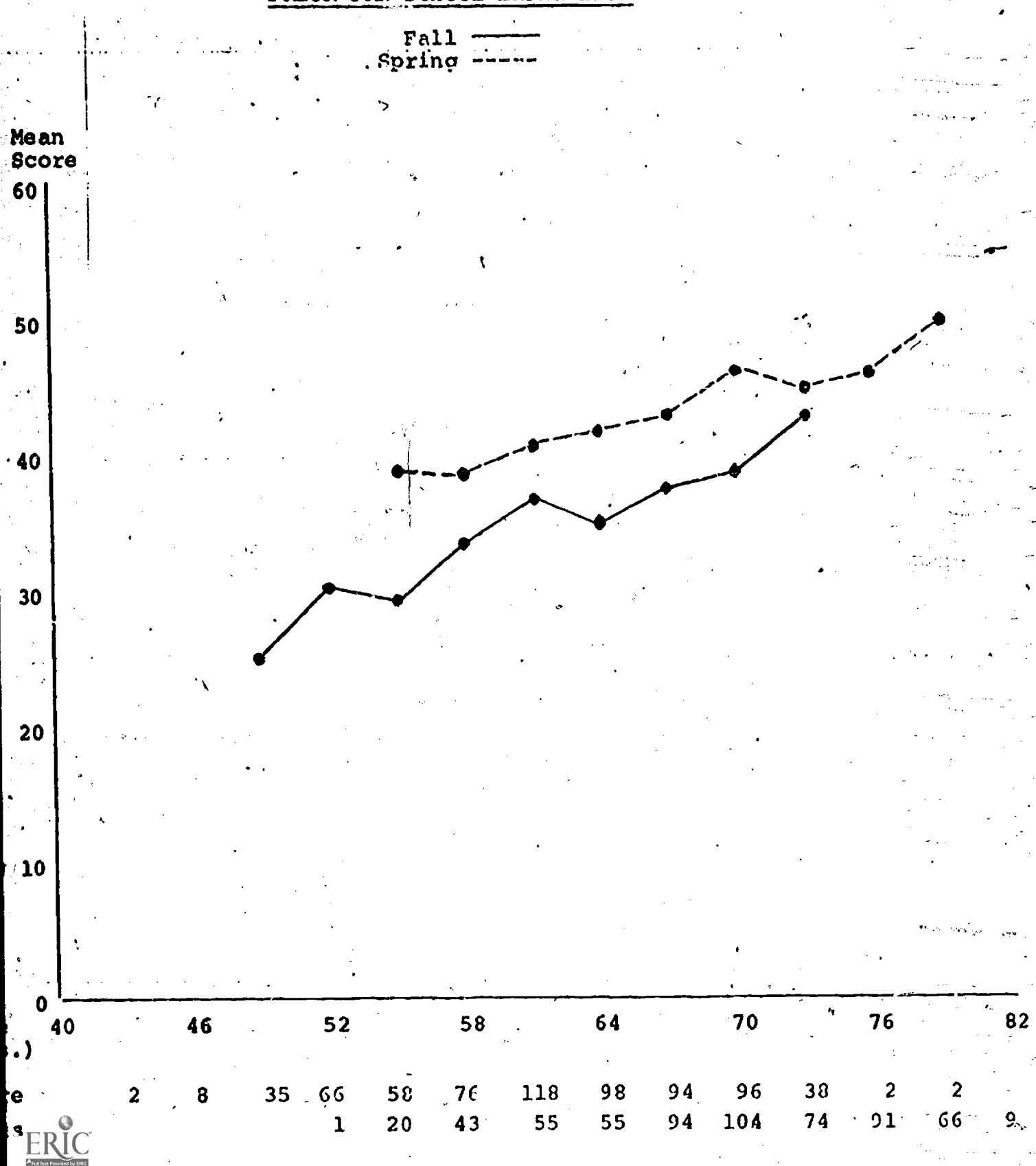
PPV - FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE

Fall ———
Spring - - - - -



)														
e	4	19	67	222	391	404	379	253	153	161	112	51	2	1
		3	8	6	20	95	272	308	307	286	177	134	114	93	11

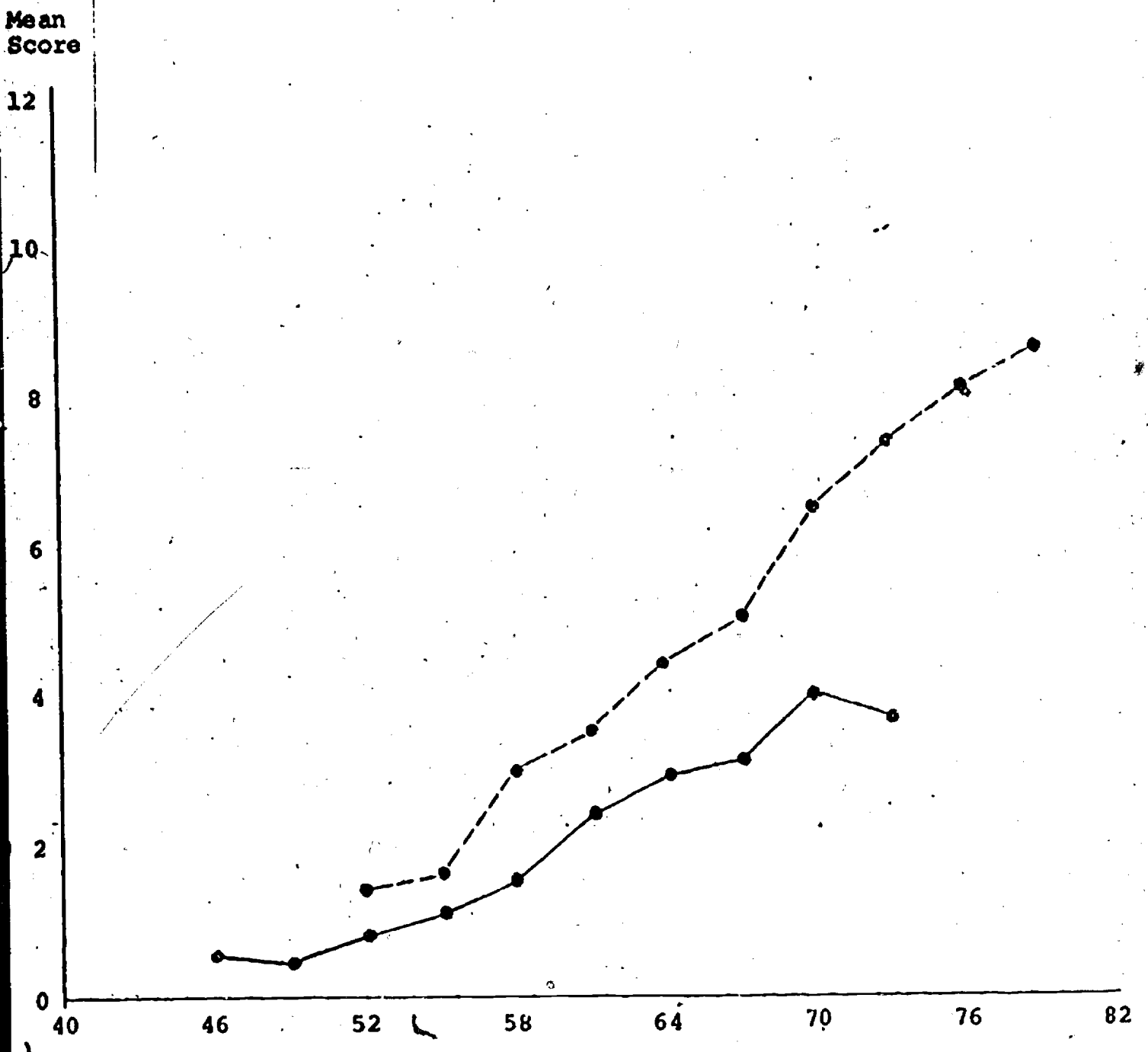
PPV - FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE



.)													
e	2	8	35	66	58	76	118	98	94	96	38	2	2
s				1	20	43	55	55	94	104	74	91	66	9

WRAT COPYING MARKS SUBTEST FOR ALL CHI
WITH NO PRIOR PRE-SCHOOL EXPERIENCE

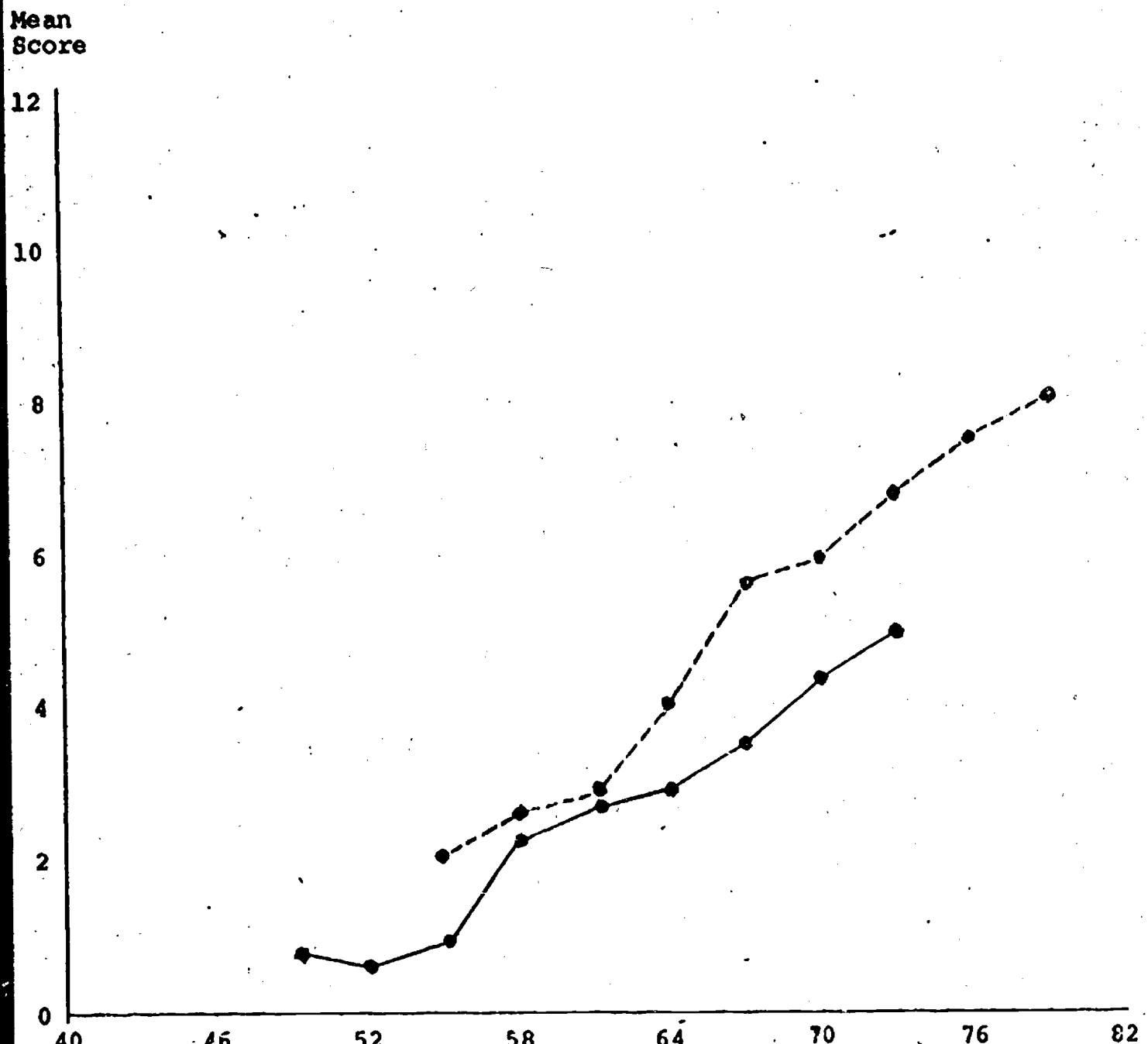
Fall ———
Spring - - - - -



.)													
re	4	19	68	221	397	405	373	249	146	157	107	50	2
o			8	6	22	94	302	323	322	308	187	141	125	95	12

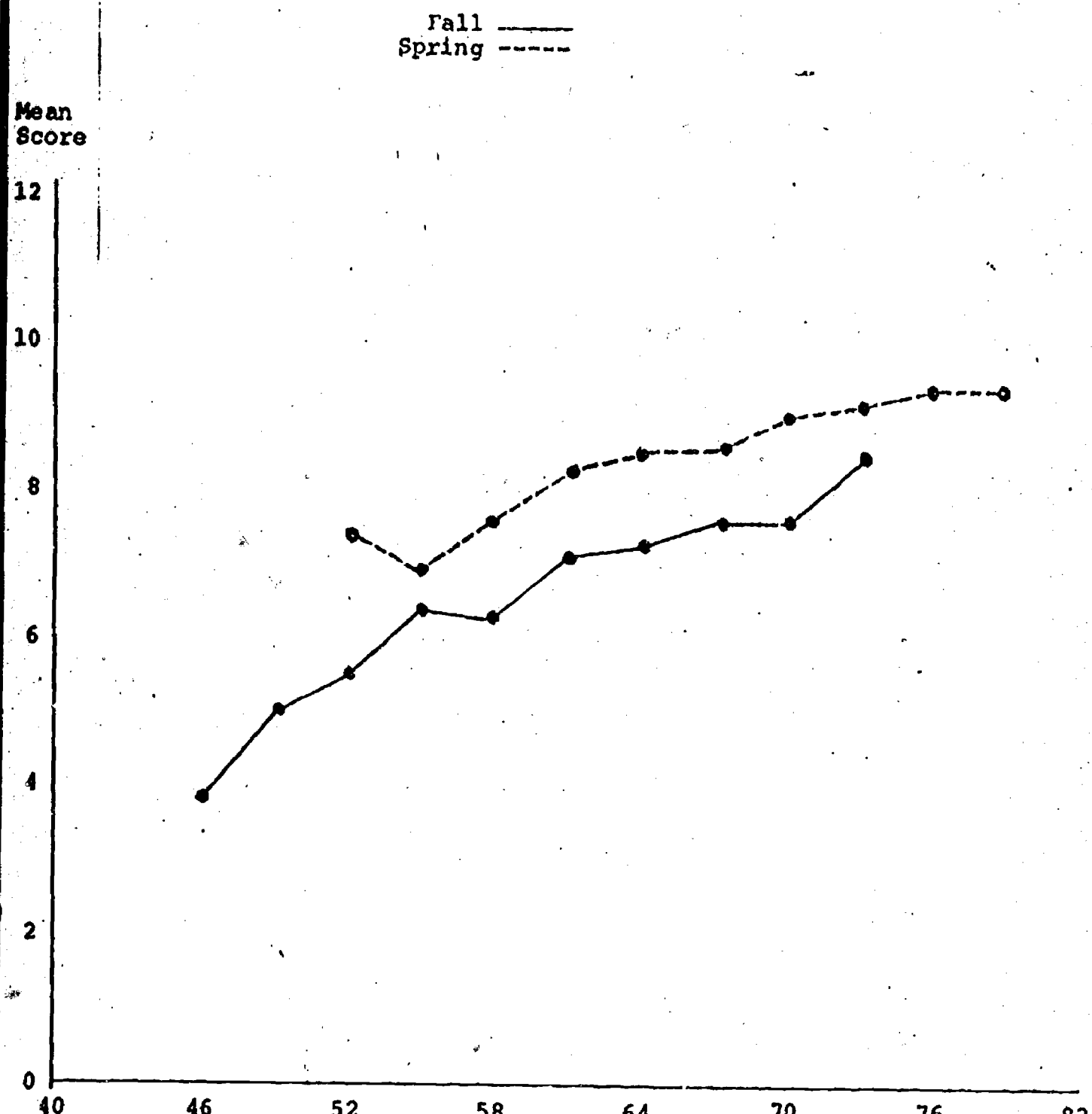
WRAT COPYING MARKS SUBTEST FOR ALL CHILDREN WITH PRIOR PRE SCHOOL EXPERIENCE

Fall ———
Spring - - - - -



)													
re	1	8	34	64	62	79	119	98	92	95	37	3	
				1	20	44	60	60	96	112	77	97	73	10

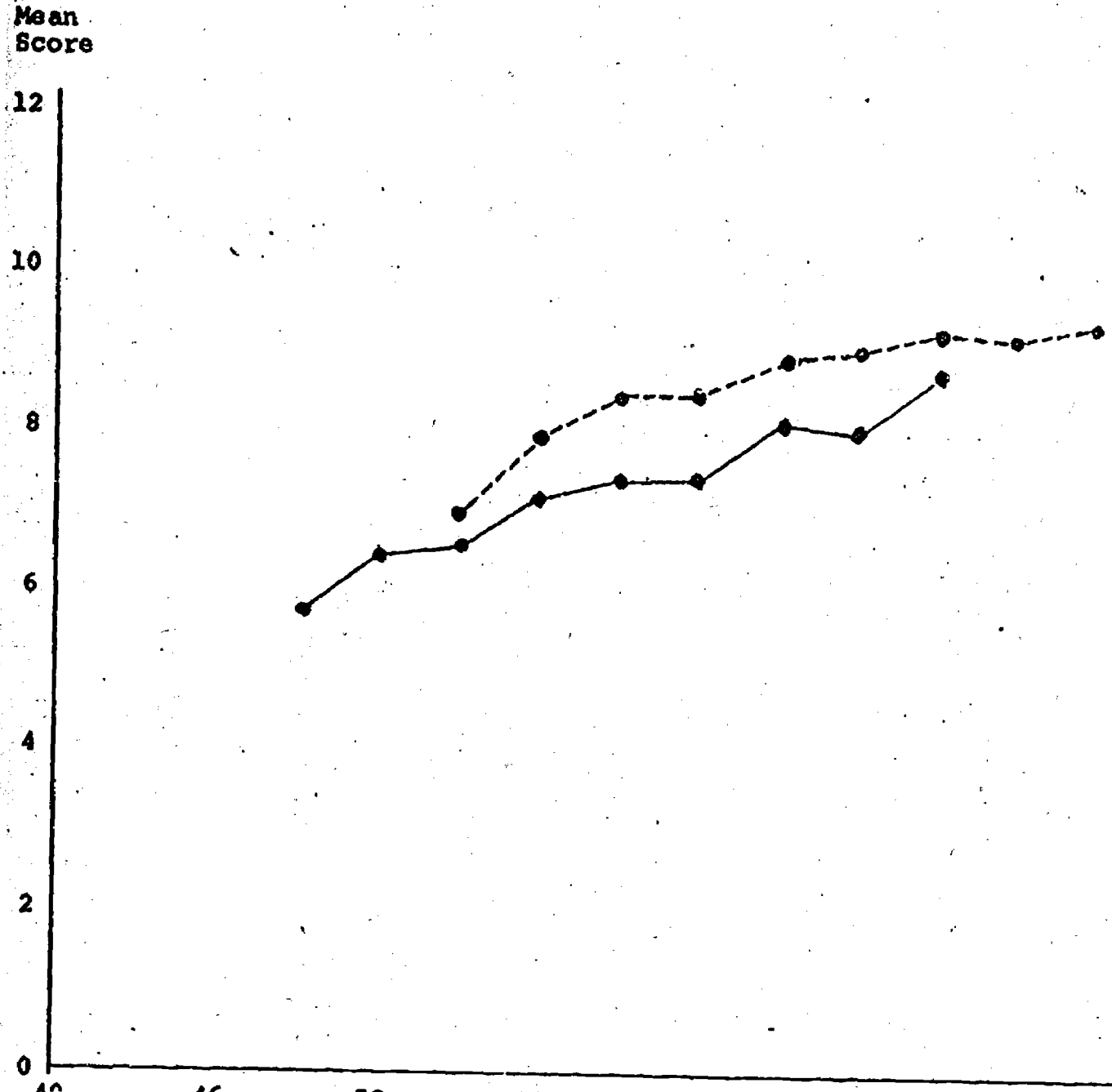
WRAT RECOGNIZING LETTERS SUBTEST FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE



.)														
re	4	19	68	221	397	405	373	249	146	157	107	50		
		3	8	6	22	94	302	323	322	308	187	141	125	95	12

WRAT RECOGNIZING LETTERS SUBTEST FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE

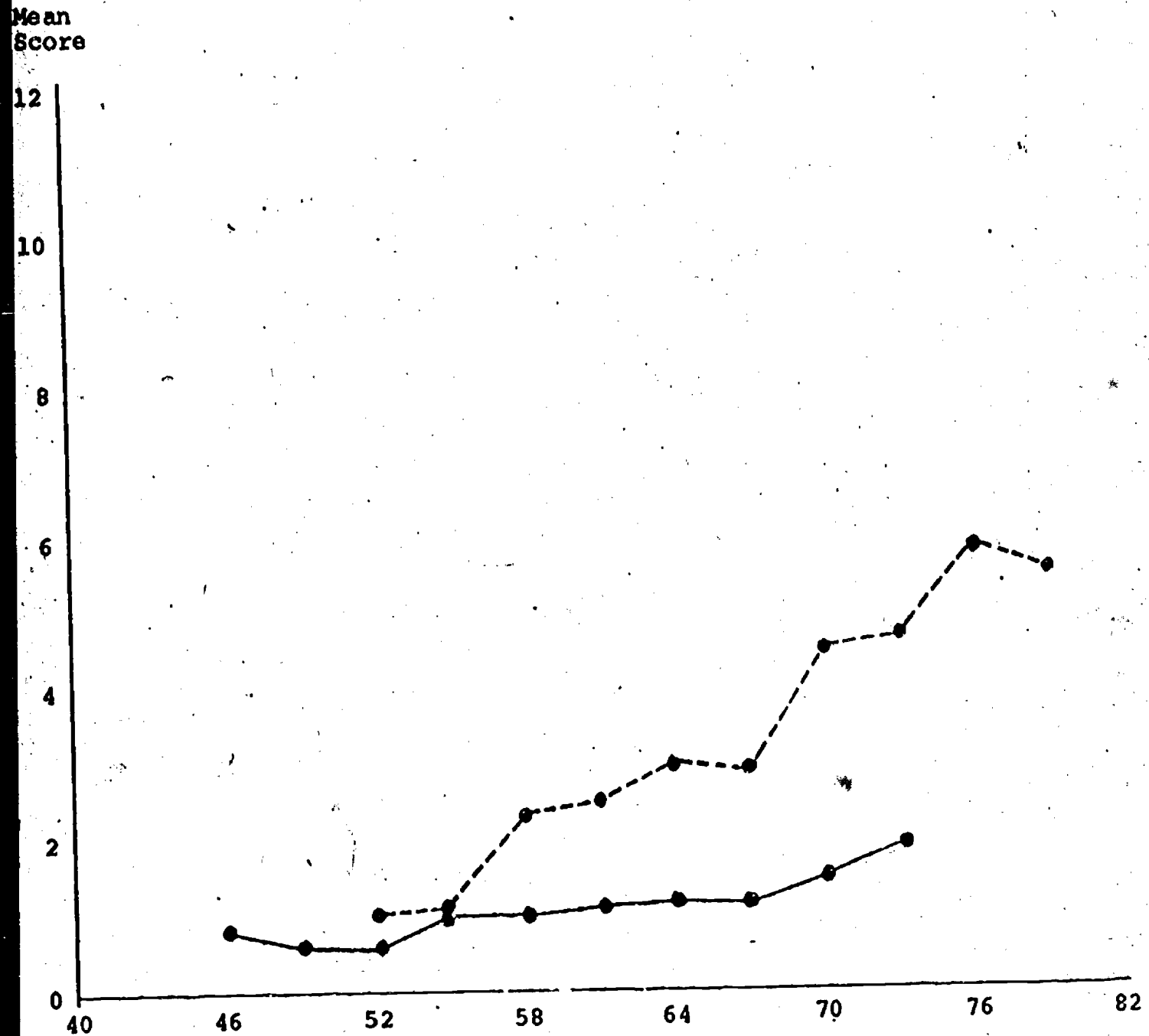
Fall ———
Spring - - - - -



.)													
re	2	8	34	64	62	79	119	98	92	95	37	3	
ost				1	20	44	60	60	96	112	77	97	73	10

WRAT NAMING LETTERS SUBTEST FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE

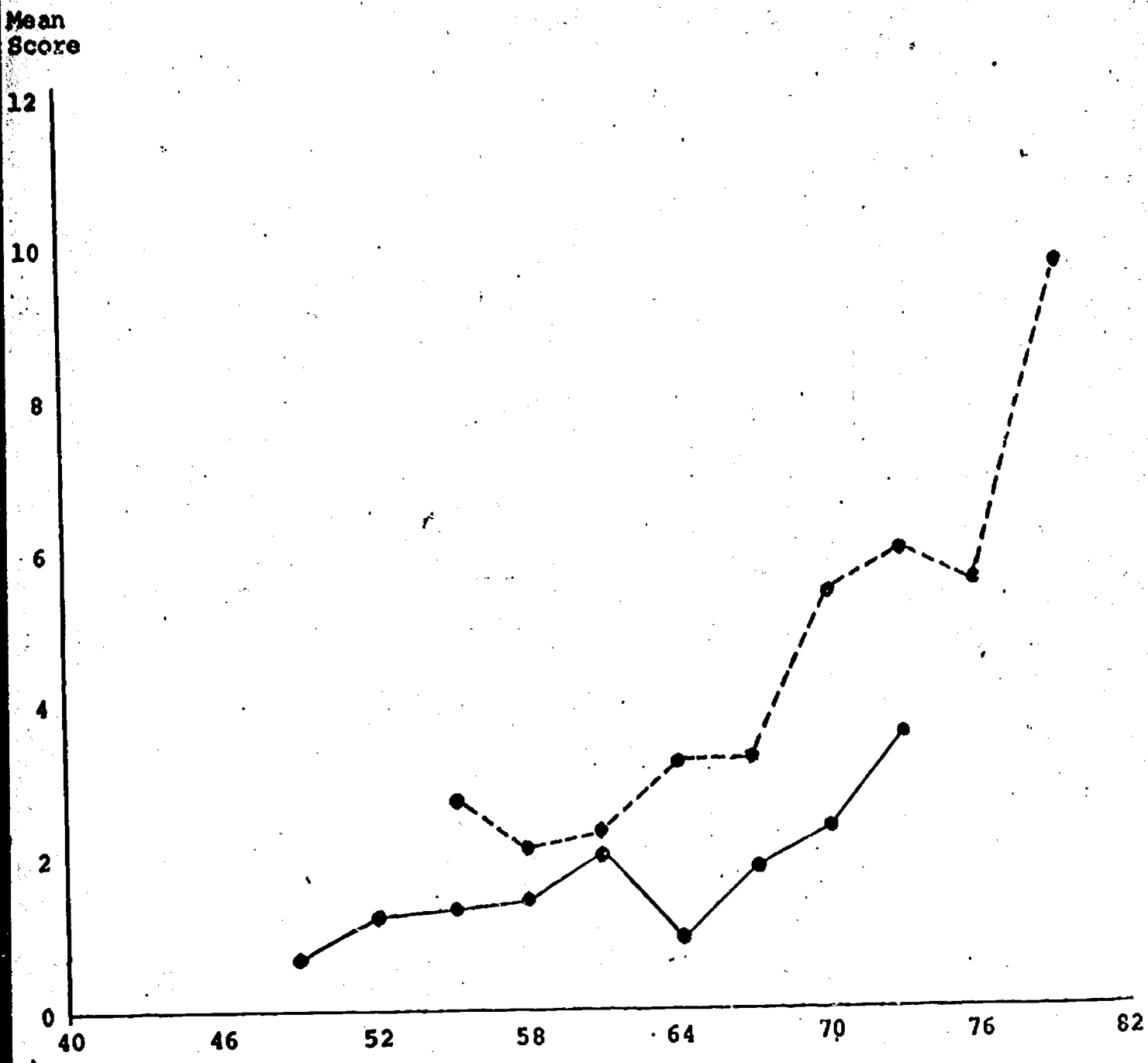
Fall ———
Spring - - - - -



.)													
re	19	68	221	307	405	373	249	146	157	107	50	2	
	3	8	6	22	94	302	323	322	308	107	141	125	95	12

WRAT NAMING LETTERS SUBTEST FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE

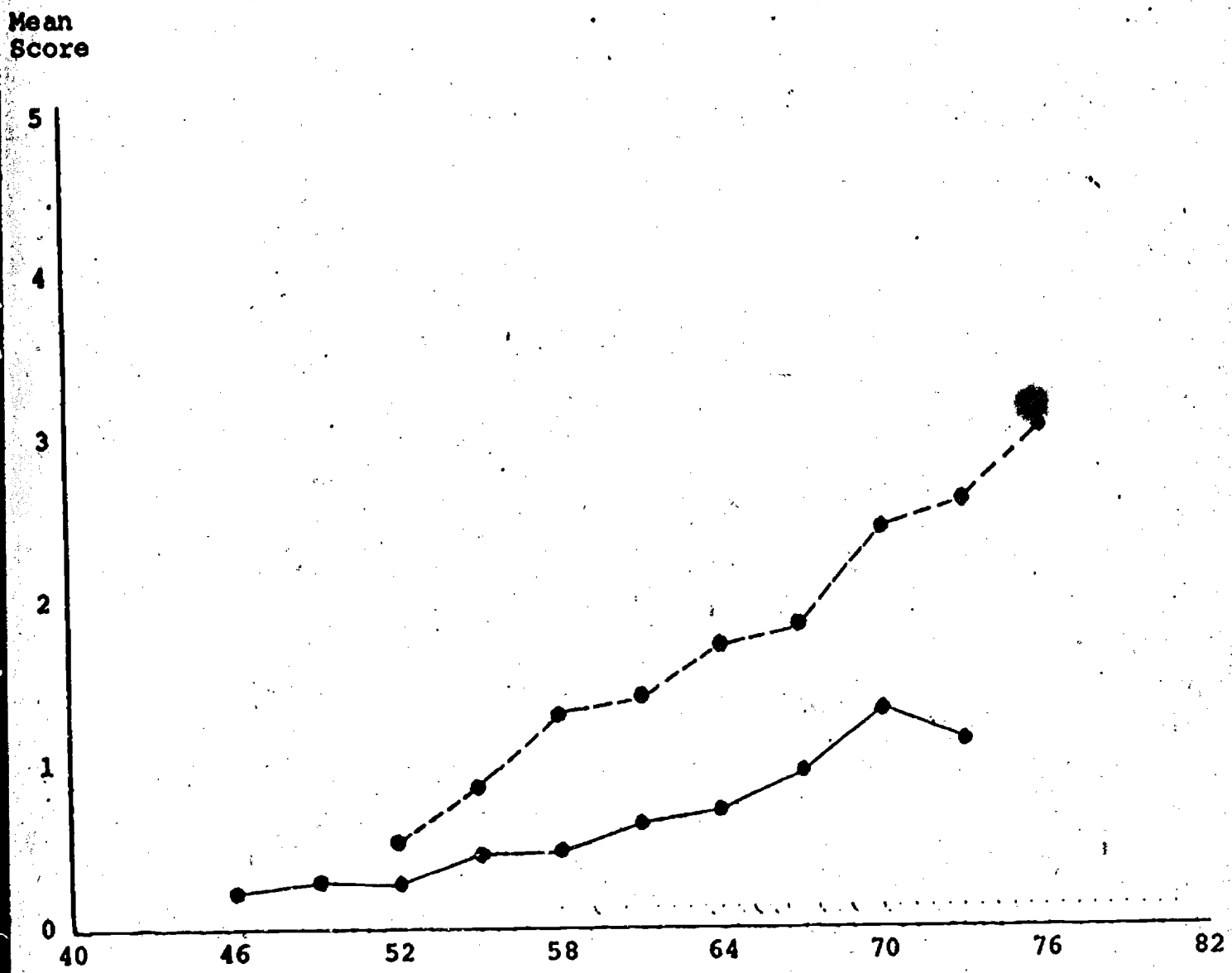
Fall ———
Spring - - - - - -



.)													
re	2	8	34	64	62	79	119	98	92	95	37	3	
					20	44	60	60	96	112	77	97	73	10

WRAT READING NUMBERS SUBTEST FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE

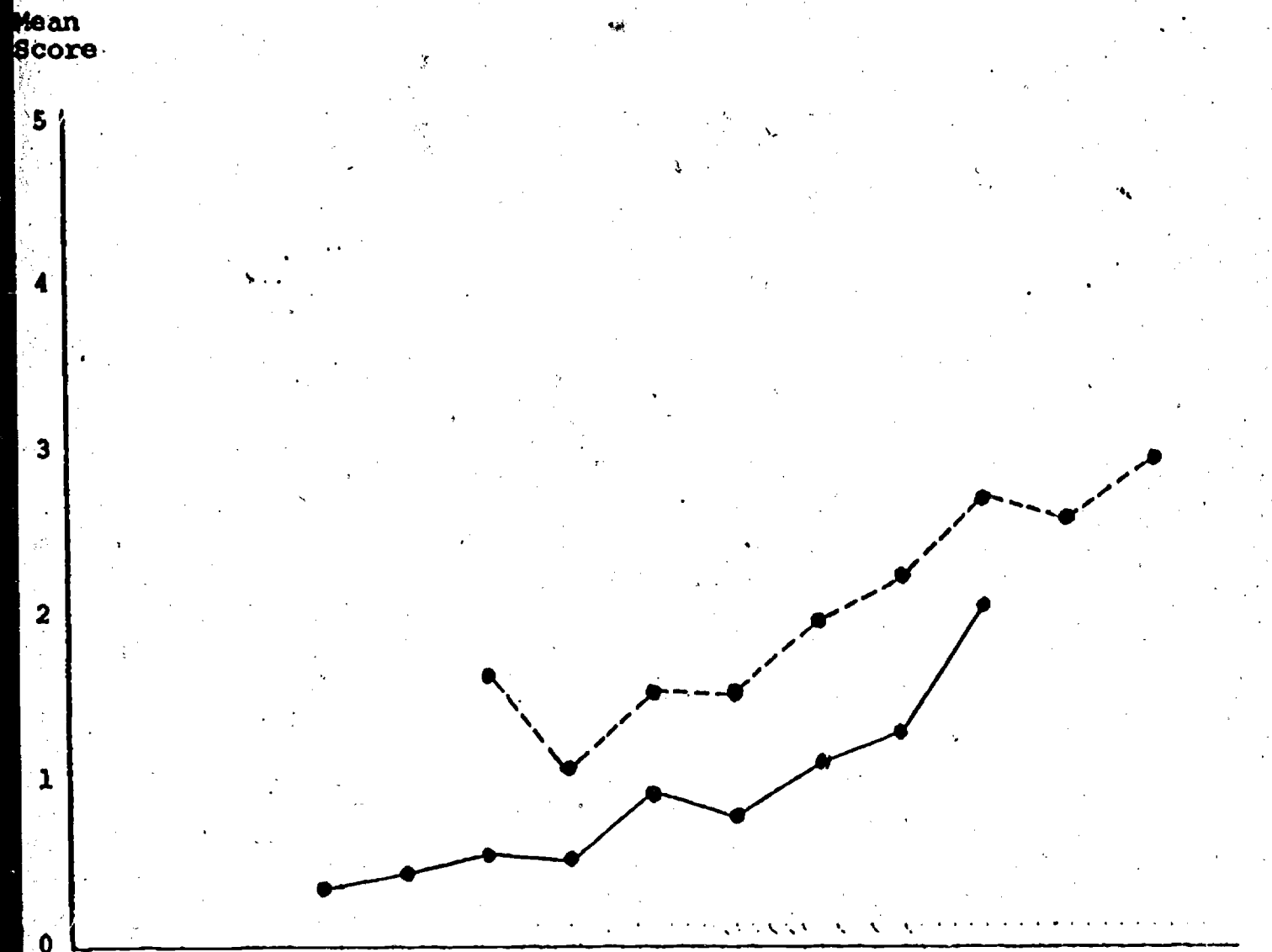
Fall ———
Spring - - - - -



re	4	19	68	221	397	405	373	249	146	157	107	50		
ost			8	6	22	94	302	323	322	308	187	141	125	95	12

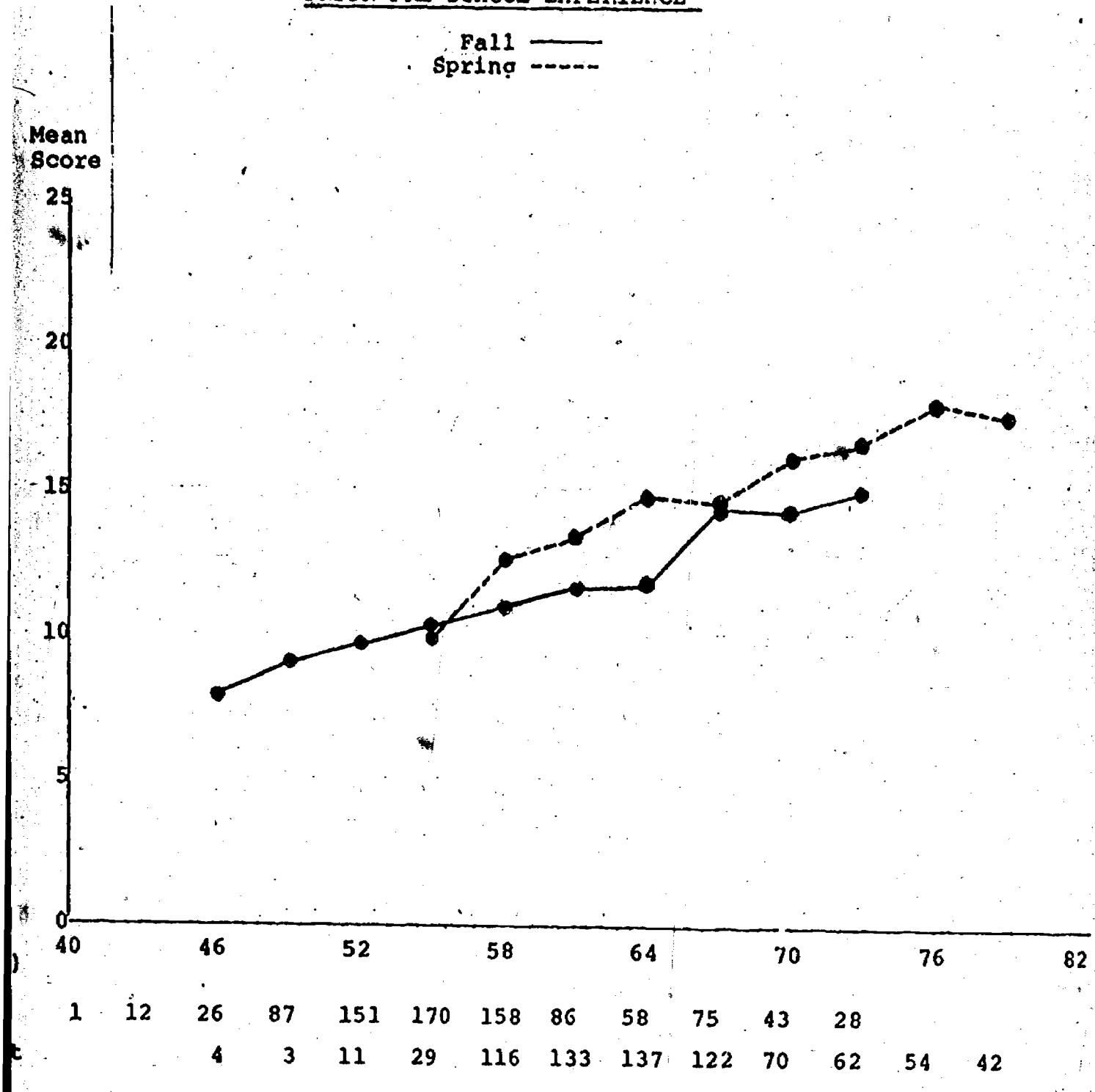
WRAT READING NUMBERS SUBTEST FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE

Fall ———
Spring - - - - -



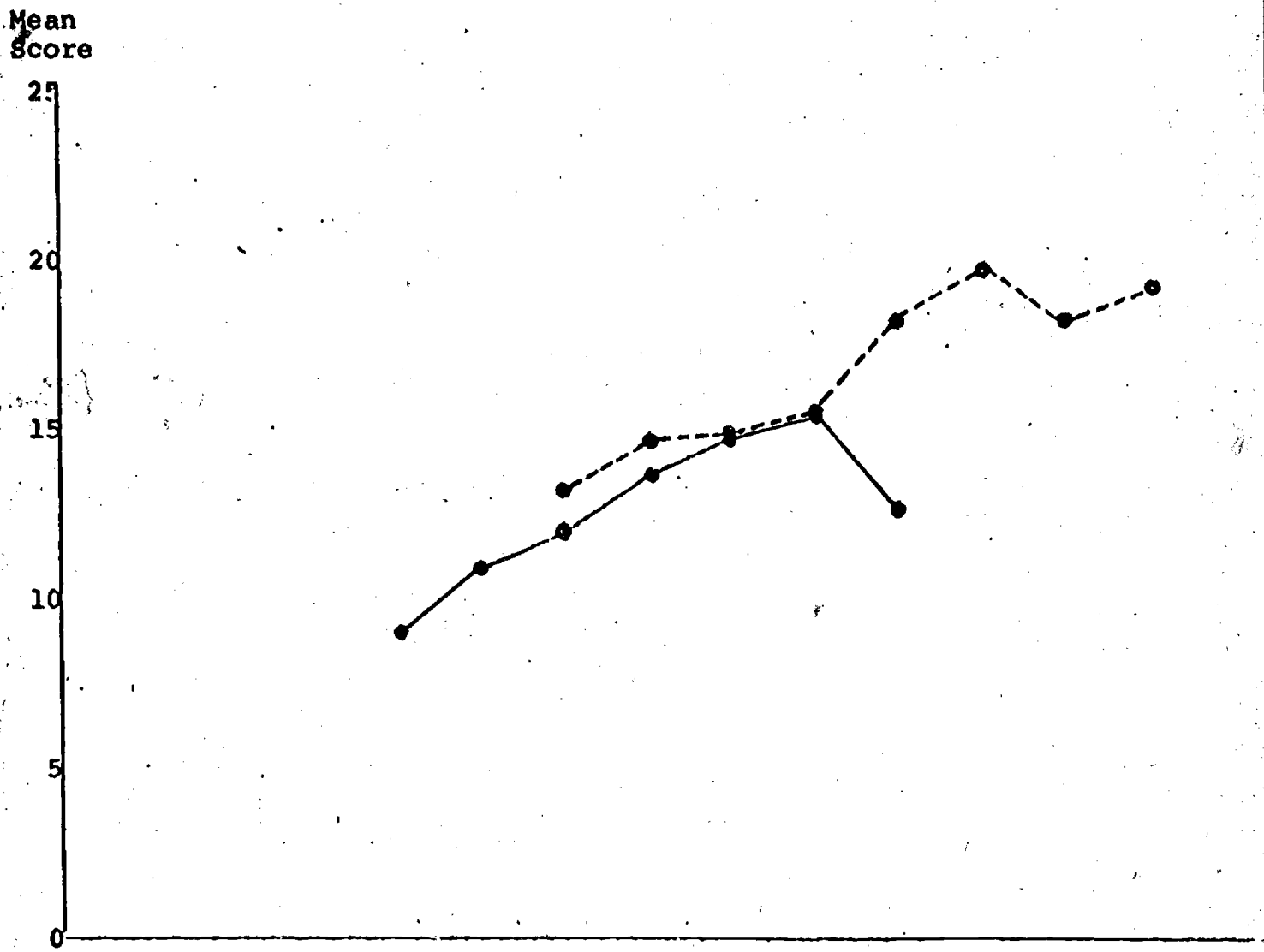
e		2	8	34	64	62	79	119	98	92	95	37	3	
st						20	44	60	60	96	112	77	97	73	10

ITPA SCORES -- FOR ALL CHILDREN WITH NO PRIOR PRE-SCHOOL EXPERIENCE



ITPA SCORES -- FOR ALL CHILDREN WITH PRIOR PRE-SCHOOL EXPERIENCE

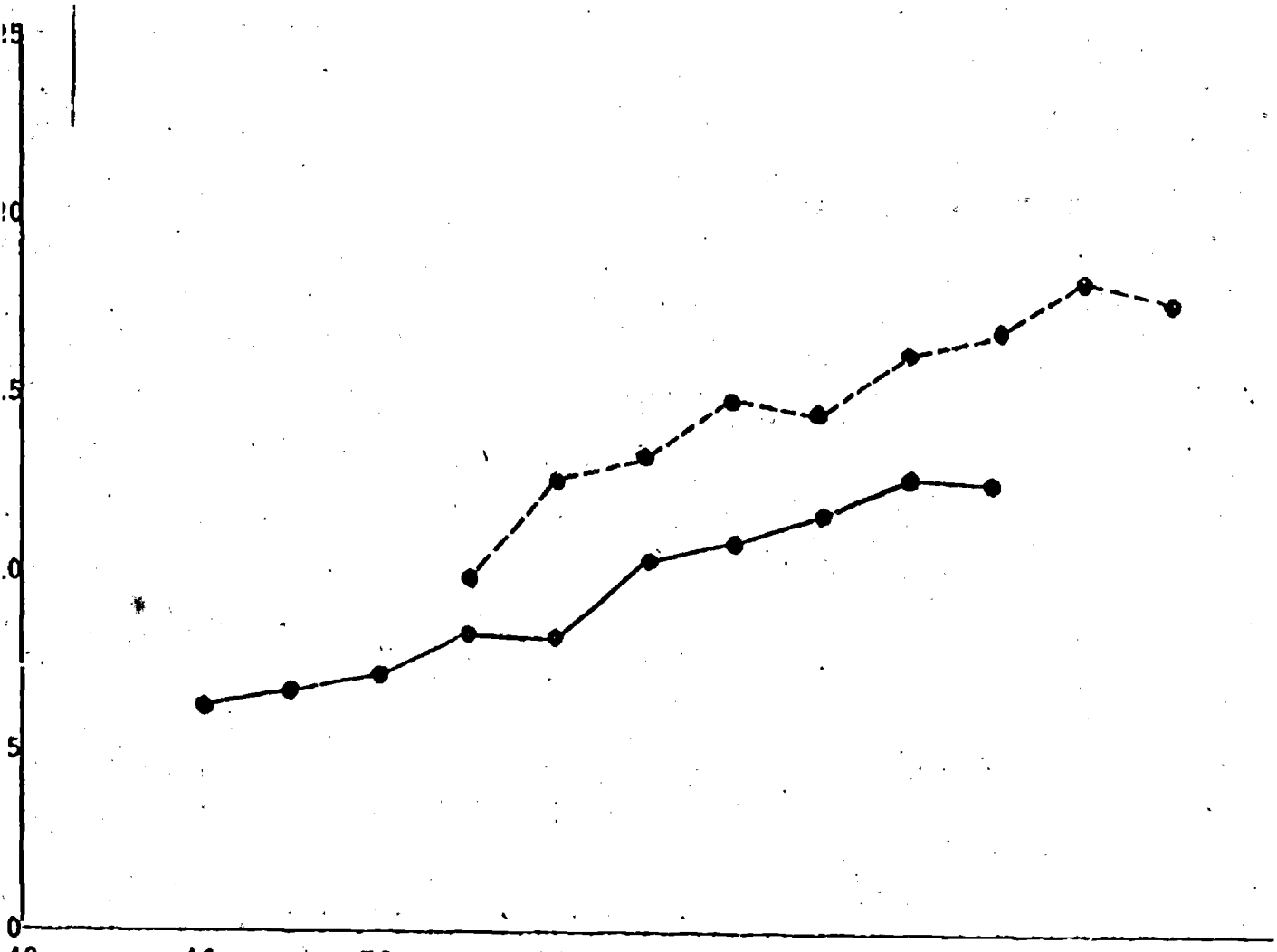
Fall ———
Spring - - - - -



.)											
e	7	18	37	22	26	46	41	39	29	14	1
st				11	25	28	20	35	42	31	37

ETS SCORES -- FOR ALL WITH NO PRIOR PRE-SCHOOL EXPERIENCE

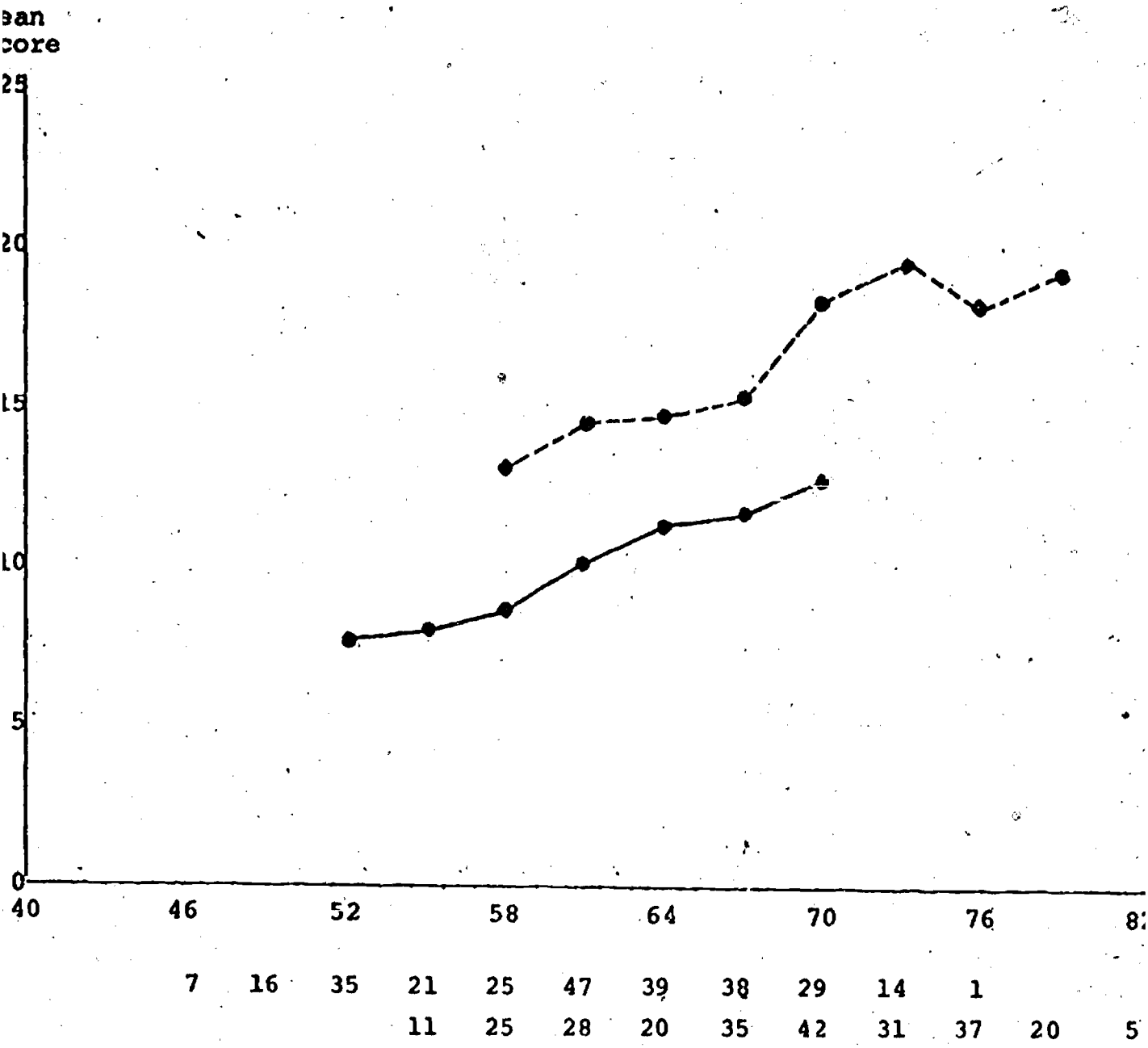
Fall ———
Spring - - - - -



1	8	21	77	132	162	145	86	54	73	43	27			
		4	3	11	29	116	133	137	122	70	62	54	42	6

x ETS SCORES -- FOR ALL
WITH PRIOR PRE-SCHOOL EXPERIENCE

Fall — ——
Spring -----



7 16 35 21 25 47 39 38 29 14 1
11 25 28 20 35 42 31 37 20 5

Appendix D

THEORY OF RESIDUAL ANALYSIS

The purpose of this appendix is to give more theoretical underpinning to the residual analysis described in Chapter V by describing explicitly the mathematical model on which it is based.

Let Y_{ij} and Y_{ij}' represent the observed pre and post test scores for individual i in group j. Let T_{ij} and T_{ij}' be the corresponding true scores. We assume that

$$Y_{ij} = T_{ij} + e_{ij}$$

$$Y_{ij}' = T_{ij}' + e_{ij}'$$

$$V(e_{ij}) = V(e_{ij}') = \sigma^2_e \qquad \forall\ i,j$$

Let a_{ij} and a_{ij}' be the age of individual i in group j at pre and post testing times respectively. Let M_{ij} be the component of true score representable as a linear function of measurable variables other than age. Let S_{ij} be the component of true score which is independent of both age and other measurable variables.

We assume $E(S_{ij})$ is constant for all individuals in the same treatment group. Let

$$E(S_{ij}) \equiv \bar{S}_j$$

If treatment groups do not differ in terms of average true score unrelated to measured variables (as would happen, for example, if individuals were assigned randomly to treatment groups), then $\bar{S}_j$ is 0 for all groups. Otherwise $\bar{S}_j$ may differ from 0 for some or all groups.

Let

$$V(S_{ij}) = \sigma^2_S$$

Our basic model can be represented by the following equations:

$$T_{ij} = \alpha + \beta a_{ij} + M_{ij} + S_{ij}$$

$$T_{ij}' = \alpha + \beta a_{ij}' + M_{ij} + S_{ij} + r_j$$

where r_j represents a residual effect attributable to the program to which the child is exposed between the pre and post tests. Then

$$r_j = T_{ij}' - T_{ij} - \Delta_{ij}$$

where* $$\Delta_{ij} = \beta(a_{ij}' - a_{ij})$$

For any individual, a sample residual can be computed in either of two reasonable ways:

$$\hat{r}_{ij1} = Y_{ij}' - Y_{ij} - \Delta_{ij}$$

$$\hat{r}_{ij2} = Y_{ij}' - \{\alpha + \beta a_{ij}' + M_{ij}\}$$

$$= Y_{ij}' - \{\alpha + \beta a_{ij} + M_{ij}\} - \Delta_{ij}$$

We can interpret $\hat{r}_{ij1}$ as the observed gain minus the expected gain, and $\hat{r}_{ij2}$ as the observed post-test score minus the predicted post-test score. Note that

$$\hat{r}_{ij1} = r_j + e_{ij}' - e_{ij}$$

$$\hat{r}_{ij2} = r_j + S_j + e_{ij}'$$

So that

$$E(\hat{r}_{ij1}) = r_j \qquad V(\hat{r}_{iji}) = 2\sigma_e^2$$

$$E(\hat{r}_{ij2}) = r_j + \bar{S}_j \qquad V(\hat{r}_{ij2}) = \sigma_e^2 + \sigma_s^2$$

Thus for any individual, $\hat{r}_{ij1}$ is an unbiased estimate of r_j with variance $2\sigma^2_e$, and $\hat{r}_{ij2}$ has bias $\bar{S}_j$ and variance $\sigma^2_e + \sigma^2_S$. A useful measure of the accuracy of an estimator is the mean squared-error (MSE), which equals the sum of the variance and the square of the bias. Thus, we have

$$MSE(\hat{r}_{ij1}) = 2\sigma^2_e$$

$$MSE(\hat{r}_{ij_2}) = \sigma^2_e + \sigma^2_S + (\bar{S}_j)^2$$

Suppose now that we consider combined estimates of the form

$$\hat{r}_{ij} = w\hat{r}_{ij1} + (1 - w)\hat{r}_{ij2} \qquad 0 \leq w \leq 1$$

Then

$$E(\hat{r}_{ij}) = wr_j + (1 - w)(r_j + \bar{S}j)$$

so that the bias is $(1 - w)\bar{S}_j$.

$$V(\hat{r}_{ij}) = w^2V(\hat{r}_{ij1}) + (1 - w)^2V(\hat{r}_{ij2}) +$$

But

$$\text{cov}(\hat{r}_{ij1}, \hat{r}_{ij2}) = \text{cov}(Y_{ij}' - Y_{ij}, Y_{ij}')$$

$$= \text{cov}(e_{ij}' - e_{ij}, e_{ij}') = \sigma^2_e$$

Thus

$$V(\hat{r}_{ij}) = \sigma^2_e\{2w^2 + w^2 + 2w(1-w)\} + (1-w)^2\sigma^2_S$$

$$= \sigma^2_e\{1+w^2\} + (1-w)^2\sigma^2_S$$

and

$$MSE(\hat{r}_{ij}) = \sigma^2_e\{1+w^2\} + (1-w)^2\{(\bar{S}_j)^2 + \sigma^2_S\}$$

Minimizing this with respect to w, we find

$$w^* = \frac{\sigma^2_S + (\bar{S}_j)^2}{\sigma^2_S + (\bar{S}_j)^2 + \sigma^2_e}$$

Thus, the theoretically optimal weight to place on $\hat{r}_{ij1}$ increases as σ^2_S and $\bar{S}_j$ increase. As mentioned above, if as the result of randomization or by luck $\bar{S}_j = 0$, then both estimates are unbiased and w* yields the minimum variance estimator and becomes simply

$$w^* = \frac{\sigma^2_S}{\sigma^2_S + \sigma^2_e}$$

It is interesting to note that this expression represents a kind of residual reliability of the test after variation related to age and other measurable background variables has been removed. Thus, the weight to be placed on the

method which uses the observed pre-test score is simply the residual reliability of this score.

Now suppose we wish to estimate r_j on the basis of the entire treatment group of size n_j. Let

$$\hat{r}_{j1} = \frac{\Sigma \hat{r}_{ij1}}{n_j}$$

$$\hat{r}_{j2} = \frac{\Sigma \hat{r}_{ij2}}{n_j}$$

Then

$$E(\hat{r}_{j1}) = r_j$$

$$E(\hat{r}_{j2}) = \frac{2\sigma^2_e}{n_j}$$

$$E(\hat{r}_{j2}) = r_j + E\left(\frac{\Sigma S_{ij}}{n_j}\right) = r_j + \bar{S}_j$$

$$V(\hat{r}_{j2}) = \frac{\sigma^2_e + \sigma^2_S}{n_j}$$

After some calculation, we find that w* which minimizes MSE is given by

$$w^* = \frac{\sigma^2_S + n(\bar{S}_j)^2}{\sigma^2_S + n(\bar{S}_j)^2 + \sigma^2_e}$$

Note that the relative advantage of Method 1 increases with sample size as well as σ^2_S and $\bar{S}_j$.

Suppose now that $\bar{S}_j = 0$ for all groups. To compute w^* we need σ^2_S and σ^2_e which are not directly available. We can estimate w^*, however, in two ways. If we have some estimate ρ of reliability, then

$$\rho = \frac{V(T_{ij})}{V(Y_{ij})} = \frac{V\{(\alpha + \beta a_{ij} + M_{ij})\} + \sigma^2_S}{V(Y_{ij})}$$

From the regression equations used to produce the residuals, we can obtain an estimate R^2 of

$$\frac{V\{(\alpha + \beta a_{ij} + M_{ij})\}}{V(Y_{ij})}$$

A natural estimator of w^* is then

$$\frac{\rho - R^2}{1 - R^2}$$

Alternatively, one can adopt an empirical approach. Let σ^2_w be the variance of the combined residual with weight w. From our previous discussion

$$\sigma^2_w = \sigma^2_e (1 + w^2) + (1 - w)^2 \sigma^2_S.$$

If we use these residuals as outcomes, and perform a one-way ANOVA with treatments as factors, the mean square error provides an estimate of σ^2_w. Carrying out the ANOVA

for different values of w, and choosing that value which minimizes the mean square error yields a reasonable estimate of w*. For each test, we carried out such a procedure, calculating our estimate of w* to the nearest tenth. Actually, the minimum can be found analytically. Since the mean square error is a quadratic function of w we need only calculate its value for any two distinct values of w to determine the entire function and hence the minimum. It is doubtful, however, that our estimation procedure is reliable enough to justify the exact calculation.

THEORY UNDERLYING RESISTANT ANALYSIS

by Sharon Hauck

RESISTANT FITTING TECHNIQUE

Because of the use of means, the usual least squares regression estimates will be affected by any extreme observations. Therefore, if one does not wish to throw away these outliers, but does wish a less sensitive estimate, he should look for another method of estimation. Tukey's resistant fitting technique by its use of medians serves this purpose.

The resistant fitting technique of Tukey may be used to fit a model of the form

$$Y_i' = \alpha + \beta f(Y_i) + e_i \qquad \text{(E.1)}$$

where Y_i and Y_i' are the pre and post test scores for individual i, e_i is the error term and f is a transformation of Y_i. Assume there are n individuals.

We will first find three pairs of representative values called summary values. Let us denote these values by $(\tilde{Y}_j, \tilde{Y}_j')$, where j = 1, 2 and 3. These values will be used to find the best transformation f and to estimate the slope and the intercept. The slope is estimated by

$$\hat{\beta} = \frac{\tilde{Y}_3' - \tilde{Y}_1'}{f(\tilde{Y}_3) - f(\tilde{Y}_1)} \qquad \text{(E.2)}$$

and the intercept is estimated by

$$\hat{\alpha} = \frac{1}{3}\left(\sum_{j=1}^{3} \tilde{Y}_j' - \beta f(\tilde{Y}_j) \right) \qquad \text{(E.3)}$$

The first step of the procedure involves sorting the pair (Y_i, Y_i') in ascending order on the pre-test score. The ordered observations are then divided into thirds of approximate size $n/3$, again according to the pre-test score (Y_{ij}). No ties are broken and, if necessary, the middle third will contain the most values.

The next step is to determine the summary values. If n is less than thirty, the median pre-test score and the median post-test score in each third serve as the summary values. (Note, these median scores need not correspond to the same individual).

If n is at least equal to thirty, each third is again divided into thirds, forming "ninths". Within each ninth, the median pre and post-test scores are found. For each third, its three pairs of ninth median values are then averaged to give a pair of summary values for that third.

These three pairs of values are used to help determine whether the linear model may be fit on the raw scale (i.e., $f(Y_i) = Y_i$) and, if not, the best re-expression to induce linearity. This is done by comparing the upper and lower slopes, denoted by S_U and S_L respectively.* A total

---

* $S_U = \tilde{Y}_3' - \tilde{Y}_2' \,/\, f(\tilde{Y}_3) - f(\tilde{Y}_2)$ and $S_L = \tilde{Y}_2' - \tilde{Y}_1' \,/\, f(\tilde{Y}_2) - f(\tilde{Y}_1)$
Also, initially $f(\tilde{Y}_i) = \tilde{Y}_i$

slope, S_T, is also used.*

If the data is linear, $S_U - S_L$ should be approximately equal to zero. If this difference is not close to zero, various re-expressions or transformations are then tried. The transformations which are used are of the form $f(Y_i) = kY_i^p$ where $k = -1$ if p is less than zero and $k = 1$ if p is greater than or equal to zero. The negation preserves the order of the scores. The case of $p = 0$ corresponds to a natural log re-expression. For the PSI and PPV tests, the following powers were considered: -1.5, -1, -.5, 0, .5, 1, 2, 3, 4 and 5. In Tukey's terminology these values form a ladder of powers.

The configuration of the three summary values or the relation between the upper and lower slopes determines in what direction one goes on the ladder for trial re-expressions. There are four possible configurations as shown in Figure 1. If either example A or B is the case, one would go down the ladder powers starting at $p = .5$. The configurations as shown in C and D call for powers of re-expression greater than one.**

---

*$S_T = \tilde{Y}_3' - \hat{Y}_1'/f(\tilde{Y}_3) - f(\hat{Y}_1)$

**In the computer program, the direction was determined by comparing $\tilde{Y}_2$ to

$$X = \hat{Y}_1 + \frac{\tilde{Y}_3 - \tilde{Y}_1}{\hat{Y}_3' - \hat{Y}_1'} (\tilde{Y}_2' - \tilde{Y}_1').$$

If $\tilde{Y}_2$ is greater than X one should go up the ladder of powers starting at $p = 2$. If Y_2 was less than X, one should go down the ladder starting at $p = .5$.

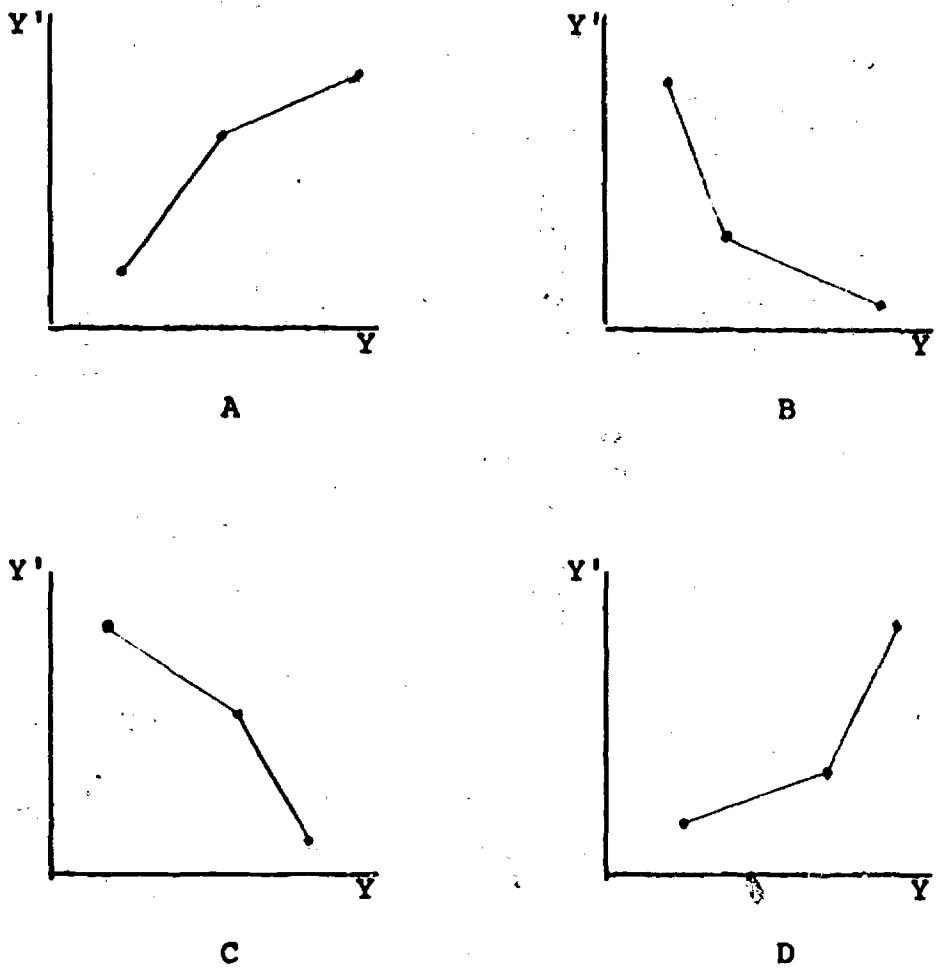


FIGURE 1

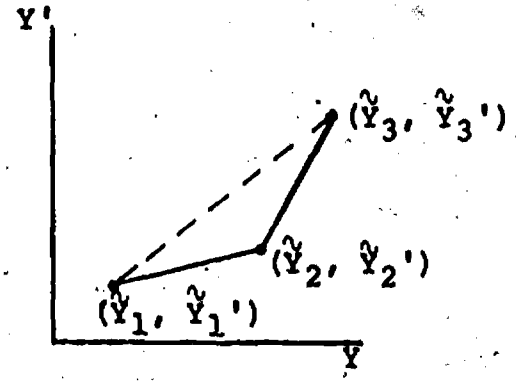
Once the first trial power has been determined, the pre-test summary values are re-expressed and the three slopes, S_U, S_L and S_T are recalculated.* As was indicated previously, the transformation which makes $S_U - S_L$ the closest to zero is the one to use. However, because the various re-expressions cause differences in the magnitudes of the slopes, the difference $S_U - S_L$, is divided by the total slope, S_T, to give a comparable relative difference. Therefore, one must look for the smallest relative difference (in absolute value).

Another consideration in determining the best transformation is the sign of $S_U - S_L$. If this difference changes sign, it indicates one has gone too far on the ladder of powers (over-expressed). (See Figure 2). However, if a power caused over-expression but still resulted in the smallest relative difference it would be used. This is because of the .5 or 1 difference between the trial powers.

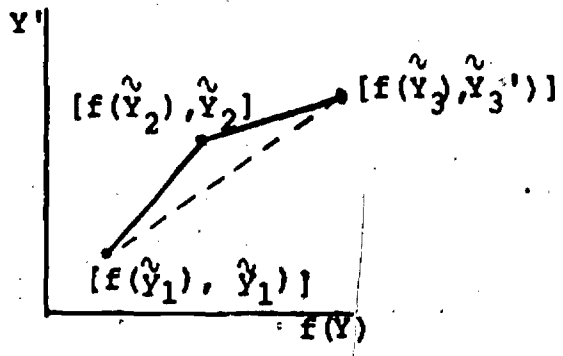
After the best re-expression has been found and the slope and intercept have been calculated from the re-expressed summary values, the fitting procedure is completed with an examination of the residuals. The quality of the fit can be judged by examining a plot of the residuals vs $f(Y_i)$. The residuals should lie in a band centered around zero.

---

*If the number of observations in a third or ninth is even, the median is the average of two values. In that case, it is necessary to re-express the two values used in calculating the median. The re-expressed summary value is then the average of these two values.



Initial Configuration



Over-expressed Configuration

FIGURE 2

If a linear trend is apparent from the plot, one may repeat the fitting procedure described above, treating [$f(Y_i)$, residual i] as the data, and omitting the search for a re-expression. The new found intercept and slope are then added to the original values for a final fit.

One might also calculate a five number summary of the residual values. This summary gives the two extreme values, the median value and the two hinge values. The hinge values are the quartiles and are found from the ordered residuals by taking those two values whose ranks lie half way between the rank of the median and either extreme. For example, if there are five observations, the hinge values would correspond to those values with ranks of two and four. The difference of the hinge values approximates the interquartile range of the residual distribution. Also, if one believes the residuals to be normally distributed, .7 times the difference of the hinge values serves as an approximation of the standard deviation.*

IMPLEMENTATION OF THE RESISTANT FITTING TECHNIQUE

Tables of the results of using this technique for the PSI and PPV tests may be found at the end of this section. All program subclasses with more than 5 children were fit.

---

* The difference of the hinge values is called the hinge spread.

However, in the final comparison only fits based on 20 or more children were considered.

In setting up the computer program to perform the resistant fitting procedure, it was necessary to decide which powers of re-expression would be considered. The core of the ladder, powers from -1 to 3, was chosen because it is the most commonly used. To allow more freedom, however, the ladder was extended from -1.5 to 5.

In this context, there is a problem of interpretation when the fitted power is greater than 1 -- for this reason all such fits were not included in the final comparison. Strict interpretations of such fits indicate that there is no ceiling effect, i.e., a child who did well on the spring test could be expected to do better than 100% in the fall. The problem is due to the fact that in some cases the model is not complex enough to adequately fit the data. We would expect the fitted curve to go to the asymtote as in Figure 3A. However, if we fit a curve as shown in Figure 3B with a simple polynomial, the resulting fit will look like that in Figure 3C.

In general, the resistant fitting technique allows either Y_i or Y_i' or both to be transformed in order to induce linearity. We chose only to re-express the pre-test scores for sake of interpretation.

As was mentioned in Chapter VII, this resistant fitting technique could not be applied to the WRAT subtests. The difficulty was due to the very large number of zero pre-test

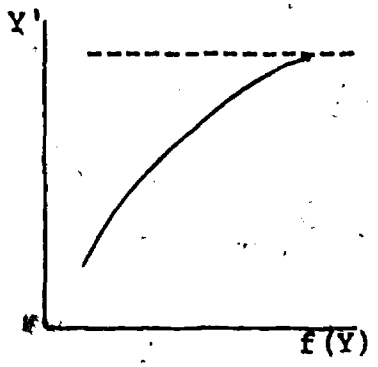


FIGURE 3A

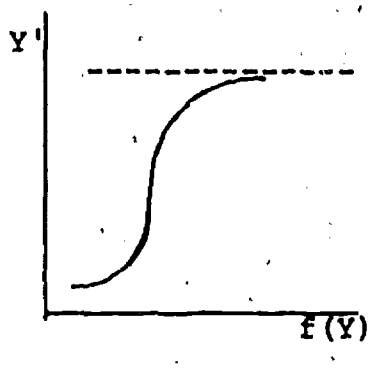


FIGURE 3B

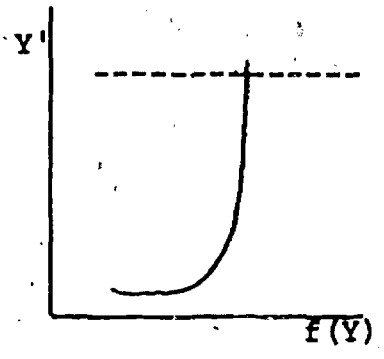


FIGURE 3C

scores and the very small range of test scores. The algorithm that divides the data into thirds does not break ties unless there are more than n/3 scores which are the same. In that case, only one value is left in the third. For the WRAT tests this meant the first third only contained one observation with $Y_i = 0$. And, because so many of the pre-test scores were the same, $\tilde{Y}_2$ was equal to $\tilde{Y}_1$ for many of the programs. This implies S_L was infinite and could not be used to determine a re-expression.

In addition, the small range of test scores led to $\tilde{Y}_3'$ being equal to $\tilde{Y}_2'$ in some cases. This resulted in an infinite relative difference which made it impossible to determine a proper re-expression.*

Besides, the obvious difficulties discussed above, the over-abundance of zero pre-test scores also led to large differences between the upper and lower slopes. The following serves an an example:

WRAT Naming Test
Sponsor: Far West
White children with no preschool experience
Sample size is 104

Summary values: (1, 11) (1.33, 2.66) (6.66, 10.66)***

Power = 1
S_L: -25.00 S_U: 1.50 S_T: -.06 RD: -450.50**

---

*If we had been finding fits, we would then have fit a model of the form $Y_{ij}' = \alpha_i + e_{ij}$.

**RD is the relative difference.

*** In order to try the various re-expressions, it was necessary to add 1 to all the pre-test scores.

Power = .5
S_L: -53.87 S_U: 5.61 S_T: -.21 RD: -282.25

Power = 0
S_L: -66.70 S_U: 11.45 S_T: .24 RD: -193.15

Power = -.5

Such large differences between the slopes and very large relative differences make any sensible chioice of a re-expression out of the question.*

---

*Part of this difficulty is due to the fact that the program which breaks the data into thirds does hot order the Y' values. Therefore, if there is only one value in a third, the post-test score may not be the lowest.

RESISTANT FITS

The six sub-classes are coded in following manner: the first digit represents ethnicity where "1" is white, "2" is Black, and "3" is Mexican-American; and the second digit represents prior preschool experience where "0" indicates none and "1" indicates some. For example code 21 represents Black children with some prior preschool experience.

RESISTANT FITS - PSI TEST

Model		# of children
FAR WEST	$Y' = 37.93 + 61.75(-1/\sqrt{y})$	167
(10)	$Y' = -.32 + 19.13 \log Y$	104
(20)	$Y' = 26.73 + 80.89(-1/Y)$	20
(30)	$Y' = 28.50 + 106.36(-1/Y)$	19
(11)	$Y' = 17.81 + .00\ Y^3$	21
(21)		3
(31)		0
ARIZONA	$Y' = -7.44 + 23.95 \log Y$	200
(10)	$Y' = 16.01 + .02\ Y^2$	86
(20)	$Y' = 11.83 + .00\ Y^3$	40
(30)	$Y' = 32.67 + 52.00\ (-1/\sqrt{Y})$	8
(11)	$Y' = 3.69 + 20.99 \log Y$	34
(21)	$Y' = .10 + 5.06\ \sqrt{Y}$	32
(31)	$Y' =$	0
BANK STREET	$Y' = -7.47 + 22.11 \log Y$	239
(10)	$Y' = 6.47 + .81\ Y$	19
(20)	$Y' = -3.34 + 5.34\ \sqrt{Y}$	116
(30)		0
(11)	$Y' = 16.67 + .00\ Y^3$	20
(21)	$Y' = 8.60 + .61\ Y$	84
(31)		0

OREGON	$Y' = -.96 + 20.24 \log Y$	155
(10)	$Y' = 21.35 + .00\ Y^3$	9
(20)	$Y' = 43.02 + 80.42\ (-1/\sqrt{Y})$	51
(30)	$Y' = 9.58 + 3.50\ \sqrt{Y}$	66
(11)		3
(21)	$Y' = -16.84 + 31.37 \log Y$	22
(31)		4
KANSAS	$Y' = 11.10 + .03\ Y^2$	101
(10)	$Y' = 33.09 + 151.29\ (-1/Y)$	39
(20)	$Y' = -4.85 + 20.61 \log Y$	54
(30)		0
(11)		1
(21)	$Y' = 39.39 + 382.60\ (-1/Y)$	7
(31)		
HIGH/SCOPE	$Y' = 13.05 + .02\ Y^2$	179
(10)	$Y' = 15.54 + .02\ Y^2$	85
(20)	$Y' = 13.55 + .01\ Y^2$	41
(30)	$Y' = 14.30 + .00\ Y^5$	18
(11)	$Y' = 56.10 + 144.50\ (-1/\sqrt{Y})$	14
(21)	$Y' = 11.71 + .02\ Y^2$	8
(31)	$Y' = 17.00$	13
FLORIDA	$Y' = 8.49 + .72\ Y$	153
(10)	$Y' = 31.48 + 159.85\ (-1/Y)$	32
(20)	$Y' = 8.29 + .71\ Y$	76
(30)	$Y' = 13.65 + .00\ Y^3$	21
(11)	$Y' = 16.67 + .00\ Y^5$	6
(21)	$Y' = 31.29 + 189.66\ (-1/Y)$	18
(31)		0
EDC	$Y' = -7.80 + 23.73 \log Y$	162
(10)	$Y' = 4.00 + 1.00\ Y$	7
(20)	$Y' = 13.48 + .00\ Y^5$	72
(30)		0
(11)	$Y' = 20.72 + .00\ Y^3$	30
(21)	$Y' = 19.50 + .00\ Y^5$	53
(31)		0
PITTSBURGH	$Y' = -1.86 + 5.89\ \sqrt{Y}$	114
(10)	$Y' = 38.96 + 67.79\ (-1/\sqrt{Y})$	89
(20)		0
(30)		0
(11)	$Y' = -11.43 + 29.34 \log Y$	25
(21)		0
(31)		0

REC	$Y' = 15.06 + .00\ Y^5$	72
(10)	$Y' = 14.92 + .42\ Y$	11
(20)	$Y' = 13.44 + .00\ Y^5$	18
(30)	$Y' = 12.00 + .41\ Y$	32
(11)		0
(21)		4
(31)	$Y' = 26.96 + 29.87\ (-1/\sqrt{Y})$	7
ENABLERS	$Y' = -4.03 + 6.27\ Y$	208
(10)	$Y' = 33.76 + 167.57\ (-1/Y)$	73
(20)	$Y' = 12.49 + .02\ Y^2$	63
(30)	$Y' = -3.43 + 19.94\ \log Y$	39
(11)	$Y' = 54.78 + 133.67\ (-1/\sqrt{Y})$	21
(21)	$Y' = 17.01 + .00\ Y^3$	6
(31)	$Y' = 10.68 + .62\ Y$	6
CONTROL	$Y' = -13.24 + 27.61\ \log Y$	105
(10)	$Y' = 28.42 + 160.86\ (-1/Y)$	33
(20)	$Y' = 10.50 + .00\ Y^3$	27
(30)	$Y' = 4.42 + .92\ Y$	16
(11)	$Y' = 11.04 + .63\ Y$	11
(21)	$Y' = 6.09 + .01\ Y^3$	14
(31)		4
NPV	$Y' = -.97 + 5.38\ \sqrt{Y}$	669
(10)	$Y' = -6.73 + 23.26\ \log Y$	170
(20)	$Y' = 8.65 + .69\ Y$	247
(30)	$Y' = 9.56 + .65\ Y$	92
(11)	$Y' = 15.48 + .42\ Y$	28
(21)	$Y' = -11.86 + 26.70\ \log Y$	121
(31)	$Y' = 4.02 + .91\ Y$	11

RESISTANT FITS - PPV TEST

Model		# of children
FAR WEST	$Y' = -16.60 + 41.84\ \log Y$	161
(10)	$Y' = -19.84 + 43.71\ \log Y$	100
(20)	$Y' = 39.71 + .00\ Y^2$	20
(30)	$Y' = -9.92 + 9.26\ \sqrt{Y}$	19
(11)	$Y' = 47.33 + .00\ Y^3$	19
(21)		3
(31)		0

ARIZONA	$Y' = -1.56 + 8.07\ \sqrt{Y}$	191
(10)	$Y' = -20.06 + 44.21\ \log Y$	85
(20)	$Y' = 31.39 + .00\ Y^3$	39
(30)	$Y' = 28.61 + .03\ Y$	6
(11)	$Y' = 81.10 + 1303.90\ (-1/Y)$	31
(21)	$Y' = 13.08 + .94\ Y$	30
(31)		0
BANK STREET	$Y' = -3.97 + 8.09\ \sqrt{Y}$	243
(10)	$Y' = 34.62 + .01\ Y^2$	17
(20)	$Y' = 72.46 + 173.94\ (-1/\sqrt{Y})$	121
(30)		0
(11)	$Y' = 82.03 + 1358.47\ (-1/Y)$	19
(21)	$Y' = 77.91 + 197.49\ (-1/\sqrt{Y})$	86
(31)		0
OREGON	$Y' = 10.26 + 6.12\ \sqrt{Y}$	136
(10)	$Y' = 14.84 + .87\ Y$	7
(20)	$Y' = 31.52 + .36\ Y$	41
(30)	$Y' = 12.30 + 5.86\ \sqrt{Y}$	65
(11)		3
(21)	$Y' = -17.17 + 42.31\ \log Y$	16
(31)		4
KANSAS	$Y' = 1.12 + 7.28\ \sqrt{Y}$	98
(10)	$Y' = 96.99 + 303.36\ (-1/\sqrt{Y})$	40
(20)	$Y' = .64 + 26.50\ \log Y$	49
(30)		0
(11)		1
(21)	$Y' = -21.34 + 10.70\ \sqrt{Y}$	8
(31)		0
HIGH/SCOPE	$Y' = -44.95 + 58.33\ \log Y$	168
(10)	$Y' = 21.58 + .68\ Y$	82
(20)	$Y' = 10.98 + .94\ Y$	37
(30)	$Y' = 39.52 + .00\ Y^5$	15
(11)	$Y' = -5.58 + 1.25\ Y$	15
(21)	$Y' = 25.38 + .00\ Y^5$	7
(31)	$Y' = 43.81 + .00\ Y^4$	12
FLORIDA	$Y' = 20.02 + .70\ Y$	143
(10)	$Y' = 97.43 + 290.63\ (-1/\sqrt{Y})$	32
(20)	$Y' = -7.82 + 34.05\ \log Y$	70
(30)	$Y' = 28.38 + .02\ Y^2$	24
(11)		4
(21)	$Y' = 29.29 + .00\ Y^3$	13
(31)		0
EDC	$Y' = -34.21 + 51.85\ \log Y$	160
(10)	$Y' = 64.57 + 554.72\ (-1/Y)$	7
(20)	$Y' = 21.34 + .02\ Y^2$	70
(30)		2
(11)	$Y' = 63.14 + 533.94\ (-1/Y)$	31
(21)	$Y' = 62.07 + 560.87\ (-1/Y)$	50
(11)		0

PITTSBURGH	$Y' = 34.83 + .01\ Y^2$		111
(10)	$Y' = 35.36 + .01\ Y^2$		85
(20)	Y		0
(30)			0
(11)	$Y' = 35.36 + .01\ Y^2$		26
(21)			0
(31)			0
REC	$Y' = 24.72 + .63\ Y$		71
(10)	$Y' = 38.12 + .00\ Y^5$		12
(20)	$Y' = 23.57 + .59\ Y$		18
(30)	$Y' = 78.27 + 173.17\ (-1/\sqrt{Y})$		31
(11)			0
(21)			4
(31)	$Y' = 34.27 + .43\ Y$		6
ENABLERS	$Y' = -3.30 + 8.08\ \sqrt{Y}$		202
(10)	$Y' = 32.85 + .01\ Y^2$		72
(20)	$Y' = 69.23 + 170.87\ (-1/\sqrt{Y})$		63
(30)	$Y' = 24.74 + .63\ Y$		32
(11)	$Y' = 75.21 + 1068.17\ (-1/\sqrt{Y})$		23
(21)	$Y' = 9.52 + .78\ Y$		6
(31)	$Y' = 10.64 + .73\ Y$		6
CONTROL	$Y' = -21.37 + 43.48\ \log Y$		106
(10)	$Y' = 32.79 + .01\ Y^2$		30
(20)	$Y' = 10.20 + .97\ Y$		32
(30)	$Y' = 35.18 + .00\ Y^4$		14
(11)	$Y' = -39.03 + 54.85\ \log Y$		11
(21)	$Y' = 28.28 + .00\ Y^5$		15
(31)			4
NPV	$Y' = 22.80 + .63\ Y$		629
(10)	$Y' = 11.95 + 5.93\ \sqrt{Y}$		166
(20)	$Y' = -.37 + 7.24\ \sqrt{Y}$		223
(30)	$Y' = 56.37 + 220.68\ (-1/Y)$		96
(11)	$Y' = 71.26 + 742.24\ (-1/Y)$		25
(21)	$Y' = 35.62 + .00\ Y^4$		106
(31)	$Y' = 27.67 + .00\ Y^4$		13

RESISTANT ANALYSIS OF COVARIANCE

Because the resistant analysis of covariance technique is new, it is advisable to begin with a brief discussion of the motivation behind it.* This is followed by a detailed description of the procedure.

This resistant procedure is analogous to the usual or classical one way analysis of covariance. However, RANCOVA is designed to be resistant to two kinds of error. The first of these involves certain observations which may be either wrong or wild, i.e., the observation though correct, is not a representative member of the population. To protect against this, medians are used in finding slopes and effects. (See the discussion of the resistant fitting technique.)

Another problem is that the assumption of equal slopes may not be satisfied. Recall in the classical situation, it is assumed that all the treatment means lie on parallel lines (as a function of the covariate), and there are tests to judge whether this assumption holds. There are no such tests for a resistant analysis. In addition, the use of a non-interactive computer to perform the analysis precludes any use of human judgment concerning the exclusion of any treatment. To overcome this second type of possible error, a weighted midmean is used instead of a

*RANCOVA will be used to denote resistant analysis of covariance.

weighted mean when combining slopes across treatments.*

In our usual notation, the model to be fit is of the form

$$Y_{ij}' = \alpha_j + \beta f(Y_{ij}) + e_{ij} \qquad \text{(E.4)}$$

where Y_{ij} is the covariate and Y_{ij}' is the response for individual i in treatment j, e_{ij} is the error term, f is a predetermined re-expression, and α_j is the j^{th} intercept or treatment fit.** Assume there are k treatments and n_j is the number of observations in treatment j.

In the classical method, assuming zero means, the least squares estimate of β is

$$\hat{\beta} = \frac{\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij} Y_{ij}'}{\sum_{j=1}^{k} \sum_{i=1}^{n_j} Y_{ij}^2} \qquad \text{(E.5)}$$

This estimate may be viewed as a weighted mean of the classical individual slope estimates for each treatment, i.e.,

$$\hat{\beta} = \frac{\sum_{j=1}^{k} w.j \hat{\beta}_j}{\sum_{j=1}^{k} w.j} \qquad \text{(E.6)}$$

---

*The choice of a midmean instead of a median was made because of results from the Princeton Study on Robust Estimation: Andrews, D.F., et al., Robust Estimates of Location. Princeton, N.J.: Princeton University Press, 1972.

**We determine f by finding the summary values and going through the search procedure as outlined in the first section for each program. A compromise re-expression is then used.

where $w.j = \sum_{i=1}^{n_j} Y_{ij}^2$ and $\hat{\beta}_j = \frac{\sum_{i=1}^{n_j} Y_{ij}Y_{ij}'}{w.j}$. (E.7)

Under the standard assumption of equal variance, the weighting is inversely proportional to the variance of the $\hat{\beta}_j$'s.

RANCOVA proceeds analogously by first calculating the slopes for each treatment according to Tukey's resistant fitting technique, using the specified re-expression. Then, the weighted mean of the $\hat{\beta}_j$'s is replaced by a weighted midmean.

Recall the Resistant slope estimate for treatment j is given by $(\tilde{Y}_3' - \tilde{Y}_1')_j / \{f(\tilde{Y}_3) - f(\tilde{Y}_1)\}_j$. Because the exact variance of this estimate is not known to this writer, an approximation is necessary. Asymptotically, the variance of an order statistic, and therefore of $\tilde{Y}_1'$ and $\tilde{Y}_3'$, is proportional to 1/n.* Because of this, the variance of $\hat{\beta}_j$ will be approximately proportional to $\frac{1}{n_j\{f(\tilde{Y}_3) - f(\tilde{Y}_1)\}_j^2}$.

Therefore, weighting inversely proportional to the variance, a weighted sum of slopes,

$$\frac{\sum_{j=r}^{k-r+1} n_j\{f(\tilde{Y}_3) - f(\tilde{Y}_1)\}_j^2 \frac{(\tilde{Y}_3' - \tilde{Y}_1')_j}{\{f(\tilde{Y}_3) - f(\tilde{Y}_1)\}_j}}{\sum_{j=r}^{k-r+1} n_j\{f(\tilde{Y}_3) - f(\tilde{Y}_1)\}_j^2} \quad \text{(E.8)}$$

is used. The sum is over those treatments whose slopes lie

*Wilks, S. S., Mathematical Statistics, New York: Wiley & Sons, 1962, pp. 273-74.

between or are the hinge values.*

In order to determine the treatment effects, the intercepts must first be calculated. These values are found using Tukey's estimate of

$$\hat{\alpha}_j = 1/3\{\sum_{m=1}^{3} (\tilde{Y}_m')_j - \hat{\beta} f(\tilde{Y}_m)_j\}. \qquad (E.9)$$

Now, rewrite the model as

$$Y_{ij}' = \mu + \gamma_j + \beta\{f(Y)_{ij} - \overline{f(Y)}..\} + e_{ij}^{**} \qquad (E.10)$$

where μ is the overall mean, α_j is the treatment effect, and $\overline{f(Y)}..$ is the grand mean of the re-expressed covariate values. Comparing this to the original model (E.1), we see that

$$\alpha_j = \mu + \gamma_j - \beta\overline{f(Y)..} \qquad (E.11)$$

and therefore,

$$\gamma_j = \alpha_j - \{\mu - \beta\overline{f(Y)..}\}. \qquad (E.12)$$

Thus, to estimate γ_j, one should subtract a quantity like $\{\mu - \beta\overline{f(Y)..}\}$, which is constant over all treatments. Analogous to the classical situation where $\Sigma n_j\gamma_j = 0$, we will require the median of the γ_j's to be zero, and, therefore, will use the median of the α_j's as an estimate of this quantity. The j^{th} treatment effect is then found by subtracting the median intercept value from $\hat{\alpha}_j$.

---

*The subscripts refer to the ordered treatments which have been ranked by magnitude of their slopes.

** $f(Y)_{ij} = f(Y_{ij})$

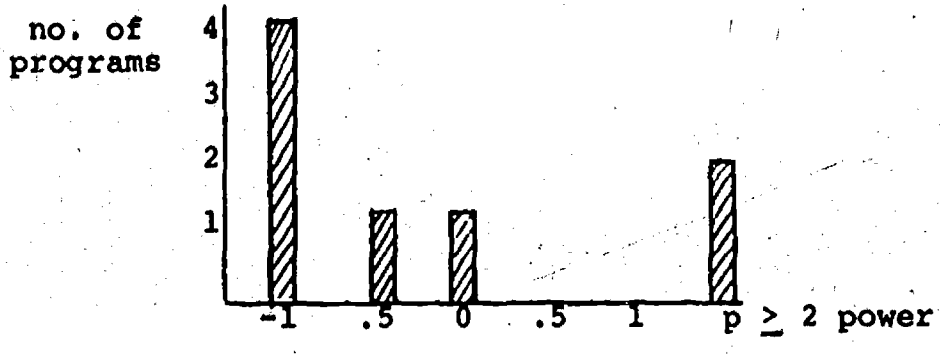
As with the resistant fitting technique, the analysis concludes with an examination of the residuals. In this case a plot of the residuals vs. re-expressed covariates for each program is useful to judge the goodness of fit. The same is true of a five number summary of the residuals for each program.

IMPLEMENTATION OF RANCOVA.

As was mentioned in Chapter VII, only those programs for which there were at least 20 children in the sub-class were used in the analysis of the PSI and PPV tests. This minimum sample size allowed enough programs for comparison while excluding fits based on too few observations.

RANCOVA could not be used on the WRAT subtests, ETS and ITPA tests because of the difficulty in determining re-expressions. The reasons for this were discussed in a previous section.

The choice of a good re-expression f is left up to the judgment of the analyst. Because of the interpretation difficulties related to powers greater than one, we constrained the choice of a compromise re-expression to lie between -1.5 and 1 although programs for which p > 1 were allowed to influence the choice. For example, the following displays the number of programs with powers ranging from -1 to 4 for White children with no prior preschool experience for the PSI test.



In this case, the choice of power was -.5, i.e., $f(Y_{ij}) = -Y_{ij}^{-.5}$. All eight programs were re-expressed accordingly and the analysis continued.

Left without the classical tests to aid his judgment, the user of resistant methods must rely on plots, displays and summary values. To judge the goodness of fit, we used plots of the raw and fitted post-test scores together vs. the covariate scores, five number summaries, and hinge-spreads of the residuals. By eye, the fitted line should "explain" the data very well, i.e., it should follow the general trend of the data. The five number summary should indicate symmetric residuals, and the hingespread should be small relative to the range of test scores.

In a covariance analysis, one expects the treatment fits to be parallel or at least very similar. Here again, no F test may be performed on the resistantly determined slopes. A plot of the fits for several programs together is helpful in examining the parallelness of the fits. Any program whose slope is wildly different from the others will be quickly pointed out and one can get an idea of the similarity between programs.

Another assumption of the classical ANCOVA is the homogeneity of the variances. To examine this in RANCOVA, one may look at a stem and leaf display of the hingespreads of the residuals from the initial fit.* Ideally the range of

*See Tukey (1970) for an explanation of stem and leaf displays.

these hingespreads should be small. The following is such a display for the PSI test, sub-class White children with no prior preschool experience.

```
6 | 34
5 | 02, 29, 09
4 | 25, 02, 25
3 | 85, 50          unit = .01
```

The values in the last line are 3.85 and 3.50. The range of these hingespreads is not very large.

As has been emphasized, plots of the residuals from the final fits were also examined. The results of RANCOVA for the PSI and PPV tests may be found in Tables.

Appendix F

INTERPRETATION OF PPV RESULTS

From the results of our various analyses, it seems clear that the performance of the Control children was comparable to that of the Head Start children, and that no particular Head Start program was outstanding in raising PPV scores. We are tempted to infer simply that passive vocabulary and whatever other skills are measured by the PPV cannot be influenced very much by preschool curricula. If this were indeed the case, we would expect the residual analysis to show small residuals for both Head Start and Control children. On the contrary, both groups approximately double their rates of growth. Thus we are left to conclude either that there is something misleading about the size of the residuals, or that the growth rate for both groups actually did increase over the period studied.

One possibility is that the observed gains can be attributed to some sort of test-sensitization or practice effect. Suppose that for some reason, children find the PPV particularly difficult the first time they take it. Then if they took it a second time soon after, they might

do a bit better even if their true ability had not changed, simply by virtue of being more used to the test.

A possible "pseudo-gain" might also have resulted from the fact that the test procedure in the spring was slightly different from that in the fall. Children began the test with item 25 instead of one. If they answered eight in a row correctly, they were allowed to proceed and given credit for the first 24. If not they essentially moved back until they were able to correctly answer eight consecutively. Thus it is conceivable that a child could be given credit for an item he would have answered incorrectly if given the opportunity. By looking at the number answered incorrectly in the fall of those given credit for in the spring, we obtained a rough upper bound on the pseudo-gain attributable to the scoring system. Our best guess is that on average one to two points of the observed gain may be explained in this way. This still leaves us with an average residual of about four or five points to explain.

The fact that the Control children tended to be younger than Head Start children might also be a factor. Perhaps younger children are growing at a faster rate, so that we are underestimating their expected increments. A breakdown of residuals by age for the Control children revealed no clear relationship between residual size and age.

Unfortunately, we see no way of determining whether the apparent increase in growth rate is real. The fact that there is so little difference among the PV models leads us to suspect that the gains are not related to program characteristics. If programs were effective agents, we would expect that, as in the case of all other tests, at least one or two would be particularly effective. With the possible exception of REC this does not appear to be the case. Thus, although we encourage the reader to draw his own conclusion from the evidence, we are inclined to believe that Head Start programs are ineffective in raising PPV scores.