DOCUMENT RESUME

ED 160 237 PS 010 165

AUTHOR Granville, Arthur C.; And Others
TITLE A Process Evaluation of Project Developmental Continuity. Interim Report III, Executive Summary: Preliminary Recommendations for the Study of Child Impact.
INSTITUTION High/Scope Educational Research Foundation, Ypsilanti, Mich.
SPONS AGENCY Office of Child Development (DHEW), Washington, D.C. Early Childhood Research and Evaluation Branch.
PUB DATE Mar 76
CONTRACT HEW-105-75-1114
NOTE 63p.; For related documents, see ED 144 715, PS 010 163-164, and PS 0010 166-176; This series includes all the public reports generated by this study

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage.
DESCRIPTORS Attrition (Research Studies); Data Collection; *Demonstration Programs; *Early Childhood Education; *Evaluation Methods; Longitudinal Studies; Matched Groups; *Measurement Instruments; *Program Evaluation; Reliability; Research Design; Sampling; Site Analysis; Validity
IDENTIFIERS *Developmental Continuity; *Project Developmental Continuity; Project Head Start

ABSTRACT

This executive summary presents the major findings of Interim Report III, which reports preliminary evaluation of Project Developmental Continuity (PDC). A Head Start demonstration program, PDC is aimed at promoting greater educational and developmental continuity as children make the transition from preschool to school. The report addresses three fundamental questions: the appropriateness of the measuring instruments; the potential comparability of children in PDC schools and those in comparison schools; and the number of children in PDC and comparison schools at each site available for a longitudinal study of program effects. Evaluation methods (instrument selection, data collection, and data analysis procedures) are discussed briefly and findings are summarized in terms of descriptive characteristics of the samples; reliability and validity of the instruments; comparability of PDC and comparison groups site-level and aggregate-level; and sample size requirements and availability. Conclusions and recommendations are presented. Numerous charts and tables illustrating the findings are included and descriptions of the selected measures of evaluation are appended. (CM)

Project Developmental Continuity Evaluation

Preliminary Recommendations for the Study of Child Impact

Interim Report III

March 1976

This report was prepared for the Early Childhood Research and Evaluation Branch, the Office of Child Development, Office of Human Development Services, Department of Health, Education and Welfare under Contract No. HEW-105-75-1114, Dr. Esther Kresh, Project Officer. Views or conclusions contained herein should not be interpreted as reflecting the official opinion of the sponsoring agency.

A PROCESS EVALUATION OF PROJECT DEVELOPMENTAL CONTINUITY
INTERIM REPORT III, EXECUTIVE SUMMARY:

PRELIMINARY RECOMMENDATIONS FOR THE STUDY OF CHILD IMPACT

March 1976

Prepared by:

Arthur C. Granville
Judy T. McNeil
John M. Love

With the Assistance of:

Mary Morris
Mel Shelly
Sally Wacker

High/Scope Educational Research Foundation
600 North River Street
Ypsilanti, Michigan 48197

Acknowledgments

A large number of people have contributed to the work that resulted in this report. The staff of eight trainers and 45 testers were directed by Mary Morris, who also was responsible for preparing the test protocols and designing the training and field procedures. Marrit Nauta of Abt Associates provided valuable consultation on the conduct of the tester training session.

Robert Hanvey coordinated the data processing effort, and, along with Helen Kiddon and Ann Hale, wrote computer programs, built the computer files, and prepared and monitored the data verification procedures. Data processing staff--Jane Oden, Barbara Bruemmer, Jodi Bruemmer, Dorothy Kelly and Daniel Muraider--were supervised by Nancy Naylor. Most of the analyses, including preliminary cross-tabulations, constructing scale scores, and the psychometric analyses, were completed by Mel Shelly. All of the typing and retyping was carried out by Lynne Dermody, Cathy Peterson, and coordinated by Jana von Fange.

We are grateful to other High/Scope Foundation staff who contributed suggestions, ideas and helpful criticism, and would like especially to thank James T. Bond, Stewart Gordon, Salvador Lopez, Judy Meece, Juanita Mendez, Allen Smith, and David Weikart.

Table of Contents

List of Figures

List of Tables

List of Tables
(continued)

I

INTRODUCTION

An Overview of Project Developmental Continuity (PDC)

The Office of Child Development originated Project Developmental Continuity (PDC) in 1974 as a Head Start demonstration program "aimed at promoting greater continuity of education and comprehensive child development services for children as they make the transition from preschool to school." The single most important effect of this undertaking, it is hoped, will be to enhance the social competence of the children served--that is, to increase their everyday effectiveness in dealing with their environment (at school, at home, in the community, and in society). Additional effects are expected in the areas of parent involvement, teacher attitudes, and institutional change.

As part of the overall Head Start Improvement and Innovation effort, PDC emphasizes the involvement of administrators, classroom staff, and parents in formulating educational goals and developing a comprehensive curriculum. The object of this effort is to ensure that children receive continuous individualized attention as they progress from Head Start through the early primary grades. Existing discontinuities between Head Start and elementary school experiences will be reduced, if the program is successful, by PDC mechanisms which encourage communication and mutual decision-making among preschool and elementary school teachers, administrators, and parents.

School organizations at fifteen sites around the country received OCD funding during 1974-1975 (Program Year I) to design and plan implementation of the seven prescribed components of PDC. The components focus respectively on:

- coordination of curriculum approaches and educational goals;
- parent participation in policy-making, home-school activities, and classroom visits or volunteering;

- comprehensive services (medical, nutritional and social) to children and families;
- preservice and inservice teacher training and child-rearing training for parents;
- programs for bilingual/bicultural or multicultural children;
- services for handicapped children and children with learning disabilities;
- administrative coordination between and within Head Start and elementary school.

Purposes of the PDC Evaluation

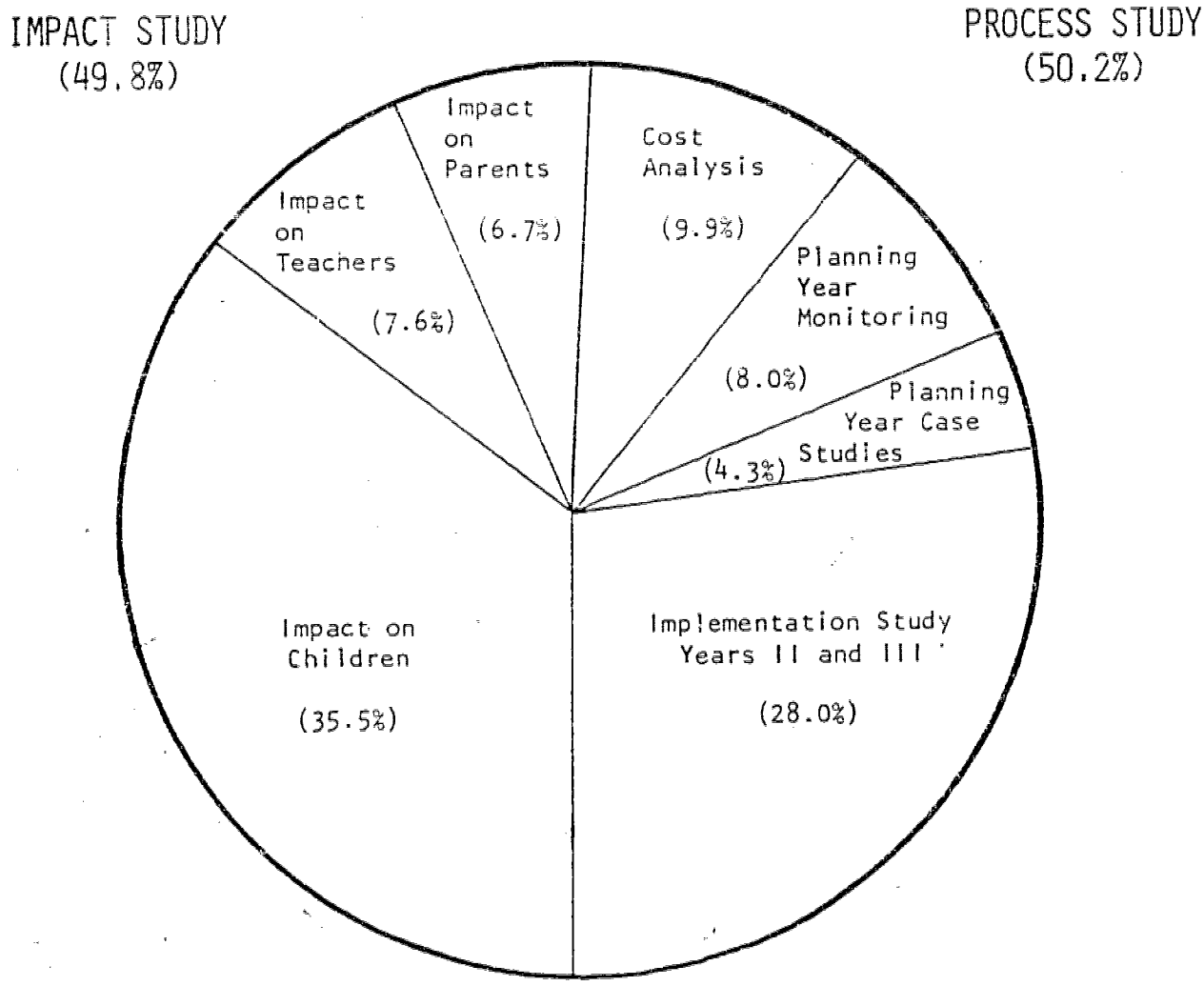
The major purpose of the PDC evaluation is to aid the Office of Child Development in its efforts to design effective programs for early childhood education. To accomplish this, the evaluation will ultimately have to provide answers to the following critical questions about PDC's impact:

- How does PDC affect children's social competence?
- How does PDC affect the school organization in terms of philosophy, methods, and social climate?
- How does PDC affect parents?
- How does PDC affect the attitudes and workstyles of teachers and other staff?

In addition to describing the consequences of PDC, the evaluation will describe and analyze the processes that led to those consequences. Figure 1 illustrates the proportions of the total evaluation effort that are devoted to each component of the study. Although the assessment of child social competence is very important and is emphasized in the present report, the relationship of this to the rest of the evaluation should not be neglected. Part B of Interim Report III delineates the process evaluation more fully; it is sufficient to emphasize here that the aims of the total evaluation are to produce conclusions about what happened (impact) and how and why it happened (process). This information will facilitate future decisions about whether the program should be replicated, and if so, how replication can best be accomplished in the light of past experience.

Figure 1

PDC EVALUATION EFFORT
(Total 3-Year Study)



Purposes of this Report

The present year, Program Year II, has been reserved as a time for sites to try out and refine the program strategies they developed during the planning year. There has so far been no measurement at all of program impact, and there will be none until 1976-1977, by which time the sites will have had a full year to pilot-test their strategies. During 1975-1976, the evaluation methodology also is being pilot-tested, and this report addresses questions about three issues fundamental to the integrity of the future evaluation:

1. Are the measuring instruments appropriate to the task?

2. Are the children in PDC schools and those in comparison schools really comparable?

3. Are enough children available in PDC and comparison schools at each site to permit a longitudinal study of program effects?

This executive summary presents the major findings from Interim Report III[1]. The methods followed in seeking answers to these questions are described in Chapter II. Chapter III presents the findings: sample characteristics, reliability and validity analyses (question 1), comparability analyses (question 2), and data related to sample size and attrition (question 3). Conclusions and recommendations for continuing the evaluation are presented in Chapter IV.

---

[1] A Process Evaluation of Project Developmental Continuity. Interim Report III, Part A. Status of the Impact Study. High/Scope Foundation, March 1976.

II

METHODS OF THE EVALUATION

Instrument Selection Procedure[1]

The two major objectives of the Impact Study, as set by the Office of Child Development (RFP-4-75-HEW-OS), are to assess the impact of PDC on the development of social competence in children and to assess the program's impact on preschools, schools, and other community organizations and groups. Accordingly, evidence of PDC's effects will be sought in four domains:

- the social competence of children,
- the attitudes and behavior of teachers, administrators, and PDC staff,
- the attitudes and behavior of parents,
- the structure and operation of the school and community organizations.

For reasons of economy and credibility, the decision was made to rely wherever possible on established instruments and procedures for measurement of program impact rather than to undertake the long, expensive, and uncertain process of instrument development. The search for existing measures was guided by standard criteria related to technical and practical characteristics of the instruments.

The measures selected for pilot-testing in fall 1975 included individual child tests, teacher-administered ratings, and classroom observations. These instruments were designed to assess child social competence in the areas of social-emotional development, psychomotor skills, and cognitive and language

[1]For a more complete review of the procedure, see Interim Report II, Part B: Recommendations for measuring program impact. High/Scope Foundation, June 1975.

development. A list of all the measures is included in Appendix A of this report. Two additional measures are being pilot-tested in spring 1976 and analysis of the spring data will lead to final recommendations for the child measurement battery.

Data Collection Procedures

Local testers were recruited from the PDC communities and were trained by High/Scope Foundation staff at a seven-day training conference in Michigan. In most cases, local PDC project staff participated in screening of initial applicants to assure the selection of testers compatible with and acceptable to the local programs. Tester training provided extensive practice in administering each of the child measures. Careful monitoring of each tester was carried out during training so that each tester achieved criterion-level performance on each of the measures before the end of the training session. Bilingual testers from the bilingual/bicultural demonstration sites were trained by bilingual staff. A system of on-site monitoring was established so that each tester at a site was responsible for monitoring each of the others on a weekly basis. In addition, testing did not begin at any site until all of the testers were monitored and found capable by one of the High/Scope trainers.

Data collection began the week of September 22 in four sites and the following week in the remaining sites. The length of data collection ranged from eight to fourteen weeks, with an average of ten weeks being required to observe and test all the children.

Data collection began with the classroom observations. This permitted a greater opportunity for children to become familiar with the testers before being taken out of class for individual testing. Testing was generally accomplished in two separate sessions--not on consecutive days, but less than ten days apart. In the bilingual/bicultural demonstration sites an additional session was required. Assignment of testers to children was made so that data collection in the PDC and comparison programs would progress in parallel. In addition, each tester was assigned both PDC and comparison children to avoid confounding group and tester effects.

Before tests were sent to the High/Scope Foundation for processing, the local site coordinator (one of the testers selected for this task) reviewed all protocols using a checklist of potential problems for each test. Completed test protocols were sent to the High/Scope Foundation at the end of each week.

Data Analysis Procedure

The baseline analyses of the child measures[1] used in Project Developmental Continuity were aimed at: 1) being able to determine the comparability of PDC and comparison groups within each site, and 2) determining the adequacy of the measures for use in the longitudinal evaluation. In order to accomplish these two objectives, the measures had to be shown to have acceptable levels of reliability and validity. The analyses included four major steps, as indicated in Figure B-1 (Appendix B): reliability analyses, validity analyses, comparability analyses, and aggregation of data across sites. The details of these four steps are described below.

Step 1: Are the Measures Reliable?

The procedures followed for the determination of reliability of measures are pictured in Figure B-2. Reliability was determined for each measure within each site. To be considered adequately reliable, a measure had to have a Cronbach alpha of at least .65. If a measure had an initial alpha of .80, no further reliability criteria were applied. If the measure did not initially obtain a Cronbach alpha of .80, item response distributions[2] and item-total

---

[1]The procedures described in this section refer to all child measures except the PDC Classroom Observation System and the PDC Child Rating Scale.

[2]Many items on the tests are constructed so that either almost all children will pass the item, or almost all children will fail the item. The easy items are included in the test to familiarize the child with the test, and the difficult items are included to allow for the determination of the upper bounds of children's abilities. Such items have little variance, and hence lower the magnitude of any reliability estimate. If the alpha was recomputed without these items in order to obtain a more accurate estimate of the measure's reliability, the items were still included in the score for the measure.

correlations[1] were inspected. Items were eliminated from the scale if they appeared to be lowering the internal consistency, and the alpha was recomputed on the modified set of items. If the original or recomputed alpha for the measure was over .65 for at least 10 sites and over .55 for the remaining sites, the measure was considered reliable for all sites. If the alpha was over .65 for fewer than 10 sites, it was considered reliable for only those sites. (In all cases, if any reliability estimates were less than .65, an effort was made to improve the administration procedures for future testing periods.)

Step 2: Are the Measures Valid?

The procedures followed for the determination of validity of measures are pictured in Figure B-3. As with reliability, the literature indicates to some extent the validity of the measures. But the validity of the measures also needed to be ascertained within the context of the PDC evaluation. Most of the measures were selected from larger existing batteries, and items on most of the measures have been modified, both to meet the needs of the sample being tested and for use by paraprofessional testers. Therefore, the validity of the measures within the PDC environment, and within the test battery in which they are administered, needed to be ascertained. The concern within this report is with concurrent validity, the correlation with other measures of the same construct as well as with measures of other constructs. A measure should correlate highly with other measures of the same construct, should correlate moderately with measures of similar constructs, and should not correlate at all with measures of independent constructs.

An hypothesized correlation matrix was constructed, based on the constructs the measures were selected to measure. The values in the matrix indicate the level of relationship that theoretically should obtain between the measures if they are valid measures of the constructs. The actual correlations (within sites) were then evaluated against the hypothesized correlations.

[1] Items with low item-total correlations appear not to be measuring the same construct as the rest of the measure. Items excluded for this reason ($r_{it} < .30$) were to be eliminated thereafter from the instrument if the rest of the instrument proved acceptable.

The hypothesized correlation matrix was constructed by determining first the correlations within the three areas of child tests; that is, within Cognitive-Language measures, within Psychomotor measures, and within Social-Emotional measures. Then the desired correlations were determined between the three groups of tests. Generally, higher correlations were expected within an area than between areas. But each area is composed of sub-constructs, so very few high correlations were expected.

The actual correlations between measures (the ones found reliable) were calculated within each site, and the following procedure used to determine whether a given measure was valid. First, the obtained intercorrelation matrix was compared with the hypothesized matrix of Table 4 and deviations of each correlation from the hypothesized one were calculated (e.g., if the hypothesized correlation was "medium" and the obtained was "low" a deviation of "-1" was scored; if the hypothesized correlation was "zero" and the obtained was "medium," a deviation of "+2" was scored). For each measure, the absolute values of the deviation were summed across all measures and divided by the number of measures. If this ratio had a value of 1.0 or less, the measure was considered valid. In effect this procedure says that a measure is considered to have adequate concurrent validity if, on the average, the obtained correlations with other measures are within the range adjacent to the expected value.

Step 3: Are the Groups Comparable Within Sites?

The two preceding steps in the analysis established which measures were useful for study of PDC's effects on children. The next task was to determine the actual comparability of PDC and comparison groups. The two groups in each site were compared on a number of demographic variables and on the performance measures found valid within that site. For every variable, all available data entered into a test of the equality of PDC vs. comparison group status. For categorical data (on ethnicity, for example) the equality of PDC-comparison group proportions was evaluated by means of the chi-square statistic; for metric data (all the test scores), equality of group means was determined by t tests. The criterion of significance for each statistical test was a probability value of less than .10.

Step 4: Are the Groups Comparable Across Sites?

After completion of Step 3, data were aggregated across sites for the subset of children who had no missing data on seven selected performance measures. This procedure was determined on a post-hoc basis in response to questions raised by the within-site analysis of group comparability. Data for the cross-site PDC and comparison group aggregations were analyzed in the same manner as that described in Step 3.

III

FINDINGS

Descriptive Characteristics of the Samples

Data were collected for 1179 children in the 14 Project Developmental Continuity sites which were in their first operational year in 1975-76. In each site an attempt was made to observe and test 30 to 45 PDC and 30 to 45 comparison children (except in Georgia, where 118 elementary children serve as the comparison group). Table 1 shows the actual number of children for whom any data were collected in each site ("Number in Full Sample"), as well as information regarding ethnicity and dominant language of those children. In California and Texas the PDC and comparison groups are further divided into English- and Spanish-dominant children, since the samples will be divided in that manner for the remainder of the evaluation. Note that in most instances the sample size of the PDC and comparison group is greater than 30, with the exception of Arizona and Florida (and of California and Texas when split by language).

Children were eliminated from the analytic sample (used for evaluation of the measures and for testing comparability of groups) if they were identified as being handicapped or as having a dominant language other than English in the non-bilingual/bicultural sites; other than English or Spanish in California, Colorado, and Texas; or other than English or Navajo in Arizona. Handicapped children are and will be included in some aspects of the evaluation, but are excluded for most aspects of this report.

The final column in Table 1 shows the number of children in each site and group who are included in the analytic sample for this report, a total of 959 children.

Table 1
Descriptive Characteristics of the Samples for Fall 1975 Data

		Number in Full Sample	Percent Handicapped	ETHNICITY						DOMINANT LANGUAGE				Number in Final analytic Sample
				% Black	% Hispanic	% Indian/ Alaskan	% White	% Asian	% Other	% English	% Spanish	% Navajo	% Other	
ARIZONA	PDC	43	0	0	0	100	0	0	0	28	0	72	0	42
	Comp	12	?	0	0	?	0	0	0	0	0	100	0	12
CALIFORNIA-Eng	PDC	33	0	9	82	3	3	3	0	100	0	0	0	33
	Comp	27	4	0	78	0	22	0	0	100	0	0	0	25
CALIFORNIA-Span	PDC	7	0	0	100	0	0	0	0	0	100	0	0	7
	Comp	13	0	0	100	0	0	0	0	0	100	0	0	13
COLORADO	PDC	38	18	0	82	0	18	0	0	100	0	0	0	33
	Comp	42	13	2	71	0	26	0	0	100	0	0	0	38
CONNECTICUT	PDC	47	4	70	23	0	6	0	0	79	21	0	0	33
	Comp	48	4	77	6	0	17	0	0	94	6	0	0	43
GEORGIA	PDC	40	20	60	15	0	25	0	0	100	0	0	0	33
	Comp	118	2	39	0	0	61	0	0	100	0	0	0	118
FLORIDA	PDC	24	0	100	0	0	0	0	0	100	0	0	0	24
	Comp	24	0	95	5	0	0	0	0	92	8	0	0	20
IOWA	PDC	45	5	64	0	2	34	0	0	100	0	0	0	43
	Comp	39	8	18	0	0	82	0	0	100	0	0	0	36
MARYLAND	PDC	47	2	51	0	0	43	2	4	98	0	0	2	45
	Comp	43	0	30	21	5	33	5	7	77	21	0	2	33
MICHIGAN	PDC	35	3	66	3	0	31	0	0	100	0	0	0	35
	Comp	58	2	95	0	0	5	0	0	100	0	0	0	57
NEW JERSEY	PDC	43	0	98	2	0	0	0	0	100	0	0	0	41
	Comp	32	3	100	0	0	0	0	0	100	0	0	0	31
TEXAS-Eng	PDC	13	0	0	62	8	31	0	0	100	0	0	0	13
	Comp	4	0	0	25	25	50	0	0	100	0	0	0	4
TEXAS-Span	PDC	29	0	0	93	0	7	0	0	0	100	0	0	29
	Comp	38	0	0	97	3	0	0	0	0	100	0	0	38
UTAH	PDC	39	15	3	36	0	61	0	0	100	0	0	0	28
	Comp	44	7	5	16	7	67	0	5	98	0	2	0	41
WASHINGTON	PDC	44	11	34	7	3	50	0	0	100	0	0	0	36
	Comp	44	23	34	2	9	52	2	0	100	0	0	0	34
WEST VIRGINIA	PDC	39	21	13	0	0	82	5	0	97	0	0	3	33
	Comp	33	17	9	0	0	91	0	0	100	0	0	0	26
ALL SITES COMBINED	PDC	560	7	40	24	9	26	1	1	86	8	4	1	508
	Comp	619	6	38	20	2	38	1	1	87	11	2	1	451

Reliability and Validity of the Instruments

Estimates of reliability and validity were already available for most of the instruments selected for use in this evaluation and these estimates were used as one of the bases for selection of measures. In addition, it is necessary to establish the usefulness of these measures for the particular populations of children within each PDC site. An estimate of reliability for each measure for each site was based on internal consistency, calculated as Cronbach's alpha. Validity for each measure for each site was assessed by comparing the obtained intercorrelations between measures with an hypothesized set of concurrent validity correlations.

Reliability of the Instruments

An instrument was accepted as reliable for a site if the final internal consistency coefficient (Cronbach's alpha) for the measure was greater than .65. In cases where the initial alpha was less than .80, efforts were made to refine the scoring procedure in such a way as to increase the magnitude of the coefficient. The refinement procedure involved determining whether the alpha might be suppressed due to (1) a large percentage of the responses to any item falling into just one scoring category (meaning the item was too easy or too difficult or irrelevant), or (2) presence of any item that was clearly unrelated to the rest of the items on the measure. If either condition was discovered, the item in question was deleted from the scale and the alpha was recalculated. Items deleted from alpha calculations for being too "hard" are still to be retained in the measure as administered; this is to allow for improved performance as the children mature. Items deleted from a scale due to low correlation with other items were to be deleted entirely from the measure if the remainder of the measure proved reliable, but this situation did not occur.

Based on the original and recomputed estimates of internal consistency, a decision was made regarding the reliability of each measure for each site. Those decisions are summarized in Table 2. The shaded portions of the table indicate that the measure was not administered at that site.[1]

---

[1]Note that the Stephens-Delys Reinforcement Contingency Interview does not appear in this or subsequent tables. The measure was administered at eight sites, but because probes and responses were not complete enough, scoring proved impossible.

Table 2
Summary of Reliability and Validity Decisions by Measure and by Site for Fall 1975 Data

R = Reliable for site
V = Valid for site
– = Not reliable/valid for site
X = Not calculated[b]
[ ] = Not administered

Child Measures	SITE: ARIZONA	CALIFORNIA-English	CALIFORNIA-Spanish	COLORADO	CONNECTICUT	FLORIDA	ILLINOIS	IOWA	MARYLAND	MICHIGAN	NEW JERSEY	TEXAS-English	TEXAS-Spanish	UTAH	WASHINGTON	WEST VIRGINIA
COGNITIVE-LANGUAGE																
BSM-English[a]		R V	R X	R V	R V		R V	R V	R V		R V	R V	R X	R V	R V	R V
BSM-Spanish[b]		X X	R –	X X	X X				X X			R X	R V			
Block Design (WPPSI)		R V	R V	R V	R V	R V	R V	R V	R V	R V	R V	R –	R V	R V	R V	R V
Conceptual Grouping		– –	– –	R V	– –	R V	– –	– –	R V		– –	– –	– –	– –	– –	– –
Say and Tell		– –	– –	– –	– –	– –	– –	– –	– –		– –	– –	– –	– –	– –	– –
Verbal Fluency		R V	R V	R V	R V	R V	R V	R V	R V		R V	R V	R V	R V	R V	R V
Verbal Memory-Part 1		R V	R V	R V	R V	R V	R V	R V	R V		R V	– –	R V	R V	R V	R V
Verbal Memory-Part 3		R V	R V	R V	R V	R V	R V	R V	R V		R V	R –	R V	R V	R V	R V
PSYCHOMOTOR																
Arm Coordination	R[a] –	R[a] –	R V	– –	– –	R[a] –	– –	– –	– –		R V	R[a] –	– –	– –	– –	– –
Block Building	– –	– –	– –	– –	– –	– –	– –	– –	– –		– –	– –	– –	– –	– –	– –
Draw-A-Child	R V	R V	R V	R V	R V	R V	R V	R V	R V		R V	R –	R V	R V	R V	R V
Leg Coordination	R[a] –	R[a] –	– –	– –	R –	R[a] –	– –	– –	– –		– –	R[a] –	– –	– –	– –	– –
SOCIAL-EMOTIONAL																
PIPS-Solutions[b]						X V					X V				X V	X V
PIPS-Locus of Control						X X					X X				X X	X X
POCL	R –	R V	R V	R V	R V	R V	R V	R V	R V		R V	R –	R V	R V	R V	R V
SITE-SPECIFIC																
Do You Know?							– –									R V
Opposite Analogies				R V										R V		

[a]For four sites, Arm Coordination and Leg Coordination were added to form one scale in order to improve internal consistency.

[b]Validity was not calculated on BSM-E for Spanish-speaking children or on BSM-S for English-speaking children. Due to small sample sizes, neither validity nor reliability was calculated for BSM-S in Colorado, Connecticut, or Maryland. Reliability was not calculated for either PIPS score since the scales contained only one "item." Validity was not calculated for the PIPS-Locus of Control since no hypothesized values were stated.

Table 2 shows that most instruments were reliable in every site at which they were administered. The exceptions are these:

- Say and Tell was reliable at none of the sites; no subset of items could be found which would yield an acceptable level of reliability.
- Block Building was reliable at none of the sites; nearly all of the children tested passed two of the four items and a substantial percentage passed a third item. It appears that this measure is not appropriate to the age level of children in this evaluation.
- Conceptual Grouping was reliable in three sites only.
- The Verbal Memory measure achieved acceptable alphas when the items were divided into Part 1 (repeating words) and Part 3 (repeating a story). Part 1 was reliable in all sites except Texas-English. Part 3 was reliable in all sites.
- Arm Coordination and Leg Coordination were reliable in only six and five sites respectively. It was possible to achieve reliability for some sites by combining the six Arm and the six Leg items into one scale.
- Do You Know?, a site-specific measure, administered in only two sites, was reliable in West Virginia, but not in Florida.

It should be noted that reliabilities and validities were calculated separately for the English and Spanish versions of the measures, the California-Spanish and Texas-Spanish sites representing the Spanish versions.

In order to summarize the level of internal consistency for the measures in the total sample, the values of the reliability estimates (Cronbach alpha) for each measure were computed for the total samples of English- and Spanish-speaking children, combined across sites (see Table 3). These alphas were calculated on the basis of all items in

Table 3

Estimates of Reliability of the Child Measures, Based on Cronbach's Alpha (Internal Consistency)[a] for Fall 1975 Data

Child Measures		English-Dominant Children		Spanish-Dominant Children	
		N	Alpha	*N*	Alpha
COGNITIVE-LANGUAGE					
Bilingual Syntax Measure-English		*691*	.82	*17*[b]	.94
Bilingual Syntax Measure-Spanish		*13*[b]	.88	*85*	.87
Block Design	(WPPSI)	*724*	.75	*87*	.80
Conceptual Grouping	(MSCA)	*721*	.63	*87*	.63
Say and Tell	(CIRCUS)	*720*	.47	*86*	.47
Verbal Fluency	(MSCA)	*726*	.75	*87*	.72
Verbal Memory, Part 1	(MSCA)	*724*	.64	*87*	.67
Verbal Memory, Part 3	(MSCA)	*725*	.85	*87*	.74
PSYCHOMOTOR					
Arm Coordination	(MSCA)	*738*	.54	*87*	.58
Block Building	(MSCA)	*738*	.28	*87*	.30
Draw-A-Child	(MSCA)	*737*	.82	*87*	.81
Leg Coordination	(MSCA)	*733*	.56	*87*	.53
SOCIAL-EMOTIONAL					
PIPS-Solutions		*233*	X	*0*	X
PIPS-Locus of Control		*203*	X	*0*	X
POCL (High/Scope)		*719*	.90	*87*	.87

[a]The samples consisted of all PDC and comparison children across all sites. Alphas were calculated separately for English- and Spanish-dominant children. All measures except the POCL (a rating scale) had an English and a Spanish version.

[b]The bilingual children within the English- and Spanish-dominant samples received both the English and Spanish versions of the BSM. Monolingual children received only the version appropriate to their group.

each measure, i.e., no attempt was made to "boost" them as was done at the site level of analysis. The alphas for the English versions of the measures were found to be very similar to those for the Spanish versions. As will be noted in later tables, the validities also tend to be very similar.

Validity of the Instruments

When an instrument was accepted as reliable within a site, a total score on the measure was calculated for each child to whom the measure was administered in that site. The validity of the measure was then evaluated for that site.

The method of evaluating validity for the purpose of this report is based on the concept of concurrent validity. The instruments were selected to measure specific aspects of a child's social competence. Those presented here focus on three areas of social competence: Cognitive-Language, Psychomotor, and Social-Emotional. A convergent-discriminant method of assessing validity was used; under this method the assumption is made that if an instrument is actually measuring the construct it was intended to measure, the results will correlate highly with other measures of the same construct, will correlate moderately with measures of similar constructs, and will not correlate at all with measures of independent constructs.

An hypothesized correlation matrix was developed (Table 4) which set an expected range of correlation values for each pair of measures based on the similarity of the constructs they are supposed to be measuring. In general, higher correlations were expected within the three areas of social competence than between the areas, but a degree of overlap between the areas was also expected. Actual inter-correlations were then calculated within each site for the measures that were judged to be reliable in that site (or for which reliability could not be calculated). The correlations were compared with the expected correlations in the manner described in Chapter II.

Table 2 summarizes the decisions made regarding the validity of the measures (in addition to the reliability). Most instruments which were judged to be reliable were also found valid. The exceptions were these:

- The combination of Arm and Leg Coordination was not valid in Arizona, Georgia, or within the Texas-English group.

- In Arizona, only Draw-A-Child was valid.

- In the California-Spanish group the BSM-Spanish does not appear to be valid, but the sample size is small, making validity more difficult to establish.

- In the Texas-English group (another small sample) only the BSM-English and Verbal Fluency appear to be valid.

- Although no expected correlation values were stated for the Internal Locus of Control scoring of the PIPS, the consistent low negative correlation of this measure with all other measures suggests that it is probably not measuring the focal construct.

Table 4 shows the hypothesized correlation values, and a summary of the site-level validity (correlation) matrices appears in Tables 5 and 6. Tables 5 and 6 show the obtained correlations for the total samples of English- and Spanish-speaking children across all sites (for children in sites where the measures were reliable). The N on which the correlation is based appears below the correlation value. The cells in which the obtained correlation value falls within the hypothesized range are demarcated by heavy lines. In general, many of the measures had very satisfactory levels of validity; even when the correlations were not in the expected range, they tended to be close.

Comparability of PDC and Comparison Groups

Information on background (demographic) characteristics of PDC and comparison children was collected, where available, to provide a basis for selection of the final analytic samples and to permit examination of the demographic similarity of the two groups. Children with handicaps that interfered with valid testing were excluded from analysis, as were

Table 4

Hypothesized Correlation Matrix
for Fall 1975 Data

Key:

0: -.1 to .1

Low: .1 to .3

Mod: .3 to .5

High: .5+

U: Undetermined

	CHILD MEASURES	Bilingual Syntax Measure	Block Design (WPPSI)	Conceptual Grouping	Say and Tell	Verbal Fluency	Verbal Memory-Part 1	Verbal Memory-Part 3	Arm Coordination	Block Building	Draw-A-Child	Leg Coordination	PIPS-Solutions	PIPS-Locus of Control	Pupil Observation Checklist
COGNITIVE-LANGUAGE	BSM	1.00													
	Block Design (WPPSI)	Med	1.00												
	Conceptual Grouping	Low	Med	1.00											
	Say and Tell	Med	Low	Low	1.00										
	Verbal Fluency	Med	Low	Low	High	1.00									
	Verbal Memory-Part 1	Med	Low	Low	Med	Med	1.00								
	Verbal Memory-Part 3	Med	Low	Low	Med	Med	Med	1.00							
PSYCHOMOTOR	Arm Coordination	0	0	0	0	0	0	0	1.00						
	Block Building	Low	Med	Med	Low	Low	Low	Low	Low	1.00					
	Draw-A-Child	Low	Med	Med	Low	Low	Low	Low	Low	Med	1.00				
	Leg Coordination	0	0	0	0	0	0	0	High	Low	Low	1.00			
SOCIAL-EMOTIONAL	PIPS-Solutions	Low	Low	Low	Low	Low	Low	Low	0	Low	Low	0	1.00		
	PIPS-Locus of Control	U	U	U	U	U	U	U	U	U	U	U	Low	1.00	
	POCL	Med	Med	Med	Med	Med	Med	Med	Low	Low	Med	Low	Med	Med	1.00

NOTE: All non-zero correlations are positive.

Table 5

Intercorrelations of Child Measures[a] for English-Dominant Children, Combined Across Groups and Sites for Fall 1975 Data

	CHILD MEASURES		Bilingual Syntax Measure-English	Block Design (WPPSI)	Conceptual Grouping	Say and Tell	Verbal Fluency	Verbal Memory-Part 2	Verbal Memory-Part 3	Arm Coordination	Block Building	Draw-A-Child	Leg Coordination	PIPS-Solutions	PIPS-Locus of Control	Pupil Observation Checklist
COGNITIVE-LANGUAGE	BSM-E		1													
	Block Design (WPPSI)	r	.29	1												
	Conceptual Grouping[b]	r	.52	.42	1											
	Say and Tell[b]															
	Verbal Fluency	r	.41	.28	.37		1									
	Verbal Memory-Part 1	r	.23	.15	.36		.32	1								
	Verbal Memory-Part 3	r	.33	.26	.42		.36	.33	1							
PSYCHOMOTOR	Arm Coordination[c]	r	.21	.16	X		.16	.17	.13	1						
	Block Building[c]															
	Draw-A-Child	r	.37	.36	.34		.30	.15	.22	.07		1				
	Leg Coordination[c]	r	.18	.10	X		.17	.32	.20	X		.18	1			
SOCIAL-EMOTIONAL	PIPS-Solutions	r	.36	.18	.44		.38	.29	.35	.45		.32	X	1		
	PIPS-Locus of Control	r	-.09	-.09	-.12		-.15	-.11	-.05	-.02		-.07	X	-.09	1	
	PMCL	r	.35	.34	.34		.41	.40	.36	.21		.29	.27	.40	-.18	1

[a]Calculated for total scores on measures judged to be reliable.

[b]Not reliable in any sites.

[c]No children represented in some cells.

Table 6

Intercorrelations of Child Measures[a] for Spanish-Dominant Children, Across Groups and Sites for Fall 1975 Data

	CHILD MEASURES		Peabody Picture Vocabulary Test-Spanish Adaptation	Block Design (WPPSI)	Conceptual Grouping	Say and Tell[b]	Verbal Fluency	Verbal Memory-Part 2	Verbal Memory-Part 3	Arm Coordination	Block Building[b]	Draw-A-Child	Leg Coordination[b]	PIPS-Solution[c]	PIPS-Locus of Control[c]	Pupil Observation Checklist
COGNITIVE-LANGUAGE	ASM-S		1													
	Block Design (WPPSI)		.40 (86)	1												
	Conceptual Grouping															
	Say and Tell[b]															
	Verbal Fluency	r	.52	.28			.1									
		n	(87)	(87)												
	Verbal Memory-Part 2	r	.36	.26			.23	1								
		n	(86)	(87)			(87)									
	Verbal Memory-Part 3	r	.75	.15			.46	.35	1							
		n	(87)	(87)			(87)	(87)								
PSYCHOMOTOR	Arm Coordination	r	.45	.26			-.01	.34	-.24	1						
		n	(19)	(20)			(20)	(20)	(20)							
	Block Building[b]															
	Draw-A-Child	r	.41	.32			.27	.21	.16	.19		1				
		n	(86)	(87)			(87)	(87)	(87)	(20)						
	Leg Coordination[b]															
SOCIAL-EMOTIONAL	PIPS-Solution[c]															
	PIPS-Locus of Control[c]															
	POCL	r	.45	.24			.42	.33	.29	-.02		.26				1
		n	(85)	(87)			(87)	(87)	(87)	(20)		(87)				

[a]Calculated for total scores on measures judged to be reliable.

[b]These tests were not reliable in any of the sites represented here (California-Spanish; Texas-Spanish

[c]Not administered to Spanish-speaking children.

children at non-bilingual sites whose primary language was not English; it was judged inappropriate to include the test scores of these two groups of children in the analysis because it seemed unlikely that their scores could be interpreted in the same way as those of other children. (Factors other than basic aptitude can complicate the performance of handicapped children, and the Spanish-translated tests administered to Spanish-dominant children cannot be presumed to be, a priori, equivalent to the English versions.)

Only at two sites, California and Texas, were there enough Spanish-dominant children to constitute statistically adequate Spanish-speaking samples. In these sites child data were grouped separately for English and Spanish speakers. Because of the sparse numbers of Spanish-dominant children at other sites, such children were excluded from site-level analysis rather than being incorporated into a special sample.

Once the final analytic samples had been established, analyses were performed to determine just how comparable the PDC and comparison groups really are at each site. Both background characteristics and test performance were examined and the results are shown in Table 7. Note that comparisons of the full samples, prior to exclusion of handicapped children and non-English speakers, are presented first because it is not otherwise possible to compare proportions of PDC and comparison children on the variables of Handicap and Language Dominance. Analyses of ethnic proportions were done both for the full and the analytic samples since exclusion of non-English speakers could artificially produce the appearance of ethnic similarity between treatment groups.

At the level of the analytic sample, the background variables examined represent characteristics that have been found in past research to be related to school performance. If the groups are not initially comparable on these dimensions, it is possible that the effects produced by PDC will be masked by extraneous differences unless these differences are somehow taken into account.

For each site and for each variable appearing in Table 7, the assumption of equal PDC and comparison group means was tested statistically (using the chi-square technique for categorical variables and t tests for metric variables). All available data entered into each analysis, meaning that

Table 7

Comparability of PDC and Comparison Groups at Each Site

* = statistically significant group difference ($p < .05$)

✓ = no significant test difference between groups

? = data insufficient for analysis

X = analysis not appropriate

– = test not given or not valid

	Site: ARIZONA	CALIFORNIA-English	CALIFORNIA-Spanish	COLORADO	CONNECTICUT	GEORGIA	FLORIDA	IOWA	MARYLAND	MICHIGAN	NEW JERSEY	TEXAS-English	TEXAS-Spanish	UTAH	WASHINGTON	WEST VIRGINIA
BACKGROUND CHARACTERISTICS OF THE FULL SAMPLES																
Handicap Presence/Absence	?	✓	✓	✓	✓	*	✓	✓	*	✓	✓			✓	✓	✓
Dominant Language	?	✓	✓	✓	*	✓	✓	✓	*	✓	✓	*		✓	✓	✓
Ethnicity	?	*	✓	✓	*	*	✓	*	*	*	✓	✓	*	✓	✓	✓
BACKGROUND CHARACTERISTICS OF THE ANALYTIC SAMPLES																
Ethnicity	?	*	✓	*	✓	*	✓	*	✓	*	*	✓	✓	✓	✓	✓
Sex	?	✓	✓	✓	✓	✓	✓	✓	✓	✓	*	✓	✓	✓	✓	*
Age	?	✓	✓	✓	✓	X	✓	✓	✓	?	✓	✓	*	*	✓	✓
Prior Preschool Experience	?	✓	✓	*	*	*	*	*	*	.	*	*	✓	✓	✓	✓
Number of Siblings	?	✓	✓	✓	✓	✓	✓	*	✓	*	*	?	?	✓	✓	✓
Mother's Employment Status	?	✓	*	✓	✓	✓	*	✓	?	*	*	✓	✓	*	✓	✓
Father's Presence/Absence	?	✓	*	✓	*	✓	*	✓	?	*	*	✓	✓	*	✓	*
TEST PERFORMANCE OF THE ANALYTIC SAMPLES																
COGNITIVE-LANGUAGE MEASURES																
BSM-English	No Comparison Possible (see text)	*	✓	✓	✓	No Comparison Possible (see text)	✓	✓	✓	No Comparison Possible (see text)	✓	✓	✓	✓	*	*
BSM-Spanish		✓	✓	✓	✓		✓	✓	X		✓	✓	✓	✓	✓	✓
Block Design (WPPSI)		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Conceptual Grouping		✓	*	✓	✓		✓	✓	✓		✓	✓	*	*	✓	✓
Verbal Fluency		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	*	✓	✓
Verbal Memory-Part 1		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
Verbal Memory-Part 3		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
PSYCHOMOTOR MEASURES																
Arm Coordination		✓	✓	✓	✓		✓	*	✓		✓	✓	✓	✓	*	*
Draw-A-Child		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	✓	✓
SOCIAL-EMOTIONAL MEASURES																
PIFS-Solutions		✓	✓	✓	✓		✓	✓	✓		*	✓	✓	✓	✓	✓
POCL		✓	✓	✓	✓		✓	✓	✓		✓	✓	✓	✓	*	✓
SITE-SPECIFIC MEASURES																
Do You Know?		–	–	–	–		–	–	–		–	–	–	✓	–	*
Opposite Analogies		–	–	✓	–		–	–	–		–	–	–	–	–	✓

even if data were missing for a particular child on one or more variables, data obtained for that child on other variables did enter into the respective analyses. A difference was declared to exist between PDC and comparison groups if analysis indicated the chance probability of the observed difference to be less than one in ten ($p < .10$). This criterion was judged to strike a balance between the need to be sensitive to small differences (which might be spurious) and the need to be fairly confident that any single difference found statistically significant is real and not just a random occurrence. In other words, it is important to be able to detect group differences in, say, ethnicity at a particular site, but it is also important not to declare a difference where none really exists.

Site-Level Findings

The asterisks in Table 7 mark statistically significant differences between PDC and comparison groups. If the analytic samples were perfectly matched, one could expect to find about ten significant group differences reported in the middle section of the table (in 10% of its 99 full cells); instead, the groups were found to differ in 28 instances. Some of these differences are more serious than others--ethnicity is invariably found to be an important background factor in studies of school effects; thus if PDC and comparison groups are not balanced ethnically, analysis of PDC effects in the future can be confounded; on the other hand, differences on the single variable, Number of Siblings, would not be expected to distort program effects very powerfully (but if a difference exists on this variable in conjunction with differences on Mother's Employment Status and Father's Presence/Absence, it is reasonable to presume that the groups are not equal socioeconomically, and this would be considered to be an inequality of some consequence).

The number of differences on performance measures shown in Table 7 is smaller than might be expected in view of the observed background differences--chance variation would result in about ten significant test differences, and only three more than that were found. This may mean that the background variables measured are relatively weak in their effects at this point or that the noted group imbalances on those variables are not great (even if statistically significant).

Because of the numerous demographic differences discovered between PDC and comparison groups at many of the sites, the prospect of conducting site-level analyses is uncertain at this time. Group imbalances can often be controlled statistically so that they do not confound analysis of treatment effects, but this is much easier with large samples than with small ones. Thus the feasibility of analyzing data aggregated across all sites was explored.

Aggregate-Level Findings

The fact that background data were not available for many children and that the tests were not all administered at all sites makes it impractical to add all the child data together to form a totally inclusive aggregate. For illustrative purposes, though, all children with complete data on seven major performance measures were pooled into aggregate PDC and comparison groups with a combined size of 634. Figure 2 shows the distributions of children in these groups on certain background variables and Figure 3 shows their relative standing on seven performance measures plus two more background variables. At the aggregate level, the similarities of the groups are more prominent than their differences. Although PDC children are likelier to be black and to have attended preschool before this year, there are no other demographic differences and only one difference in test performance (that one, as can be seen, is quite small in the perspective of the range of obtained scores).

The differences found at the site level appear not to consistently favor one group over the other, since they tend to disappear when the data are aggregated (except on the variables of Ethnicity and Prior Preschool Experience). Thus in performing future impact analyses, it appears that it will be feasible to aggregate data across sites to "smooth out" most of the imbalances that may exist and to control any remaining imbalances by statistical means.

Sample Size Requirements and Sample Availability

In 1974 and 1975, sites were asked to submit data documenting attrition rates from grade to grade (beginning in Head Start) in designated PDC and comparison schools. Based on the data provided, estimates were made of the

Figure 2

Background Characteristics of Aggregated Samples of PDC and Comparison Children

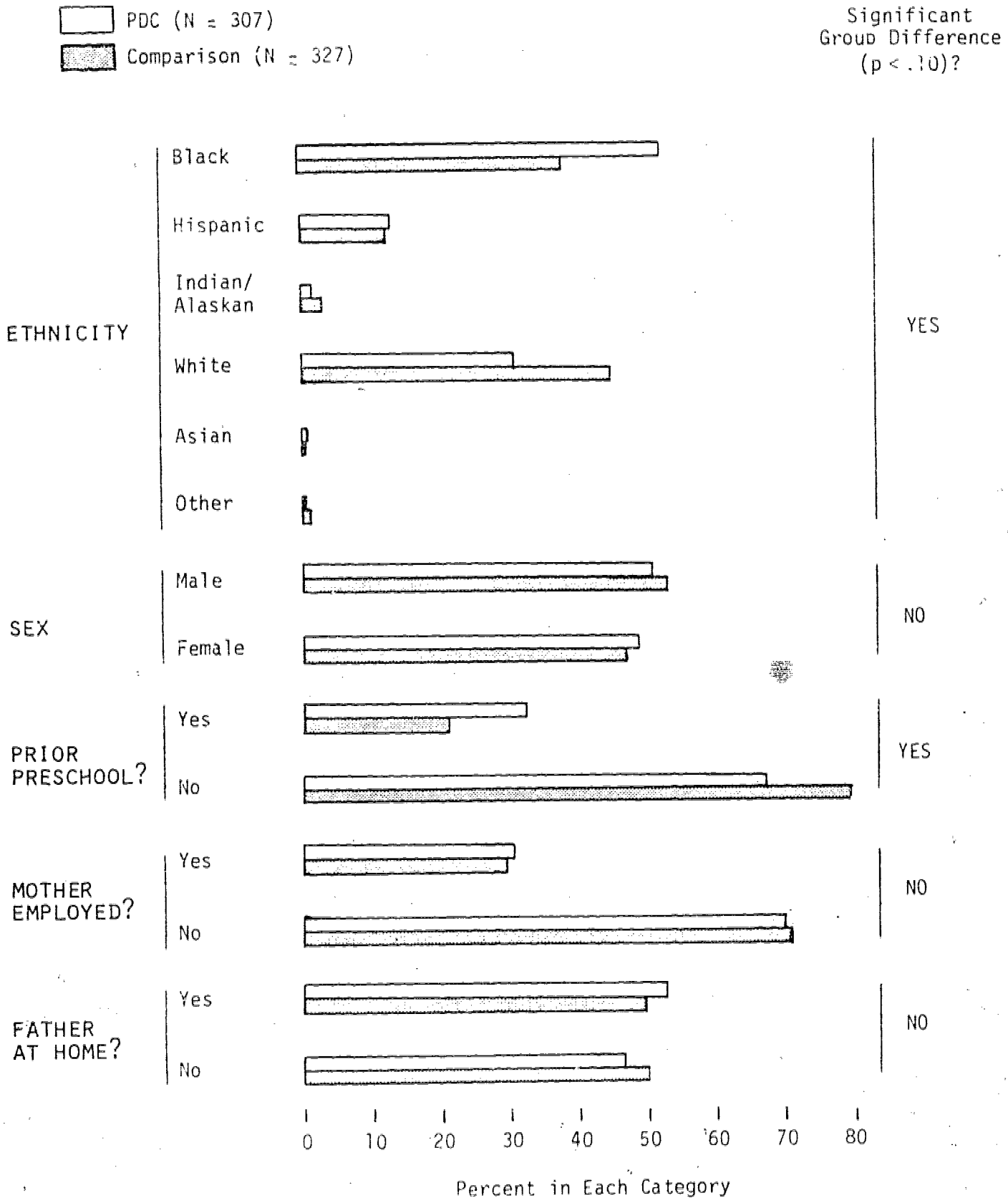
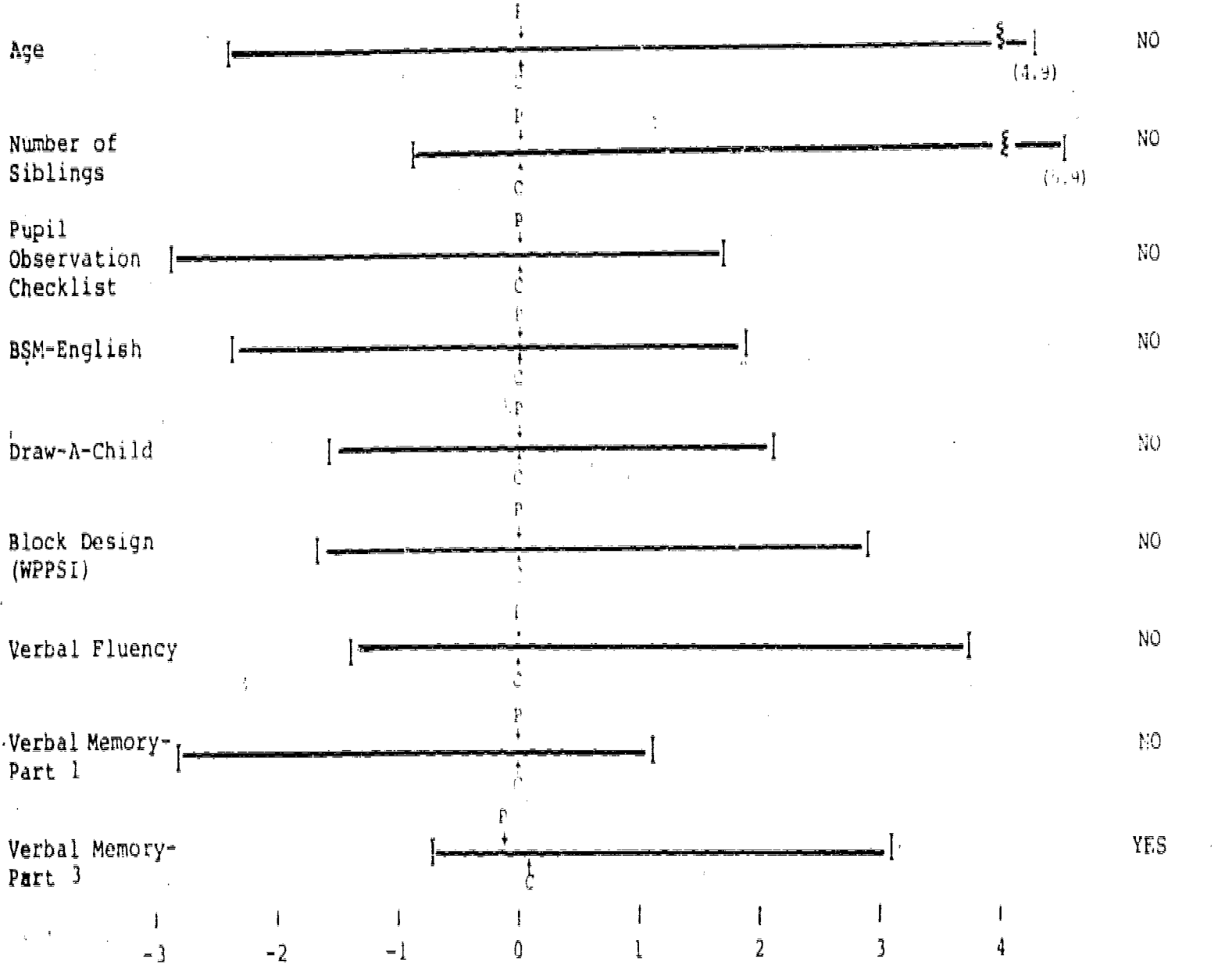


Figure 3

Mean Standard Scores of Aggregated Samples of PDC and Comparison Children
On Selected Demographic and Performance Measures



Head Start sample sizes necessary at each site to ensure that at least 30 children would remain in each of the two groups through the third grade (the terminus of the prospective longitudinal study). These estimates were first published in Interim Report II, Part A (1975) and are reprinted here, with some revisions, in Table 8. The table also shows each site's most recent projection of PDC and comparison group sizes for fall 1976 (the children entering at that time, Cohort 2, will be the focal samples of the Impact Study). It is evident in Table 8 that many of the sites do not expect to enroll the specified numbers of children. In cases where the number enrolled next fall turns out to be within 10-20% of the requirement, it might still be possible to perform analyses of program effects within site, depending on the degree of group comparability; in cases where the numbers are 10-20% low and the groups depart from comparability, the samples might still contribute usefully to a cross-site aggregation of PDC and comparison groups, upon which the analyses could be carried out. But in cases where enrollment for the coming school year falls drastically below the requirement, the utility of continuing the child testing phase of the Impact Study is questionable--there will be no possibility of a within-site analysis, and very small samples are as likely to complicate an analysis at the aggregate level as to facilitate it.

Table 8

Head Start Samples Required and Samples Projected at Each Site for Fall, 1976

	PDC Classes			Comparison Classes		
	Number Required	Number Expected	High-Low Range of Expectation	Number Required	Number Expected	High-Low Range of Expectation
ARIZONA	60	22	20-30	60	22	20-30
CALIFORNIA	45	47	45-50	45	57	55-61
COLORADO	60	68	68-68	60	68	68-68
CONNECTICUT	75	60	45-60	75	60	45-60
GEORGIA	60	60	50-60	--	--	-----
FLORIDA	75	45	30-50	75	37	30-40
IOWA	60	60	45-75	75	65	60-70
MARYLAND	75	70	?	75	60	?
MICHIGAN	60	75	75-75	75	60	60-65
NEW JERSEY	60	70	70-70	60	45	45-45
TEXAS	75	45	40-50	75	45	40-50
UTAH	65	65	35-70	75	65	40-80
WASHINGTON	75	60	60-60	75	100	92-100
WEST VIRGINIA	45	45	-----	45	45	-----

IV

RECOMMENDATIONS AND CONCLUSIONS

The evaluation activities reviewed in this report were designed to result in recommendations regarding the Impact Study component of the Developmental Continuity evaluation. The recommendations outlined in this chapter are based on two sets of data collected over the past year and on other considerations affecting the overall quality of the evaluation. The first set of data consists of information on the suitability of the instruments. Alterations in the battery based on analysis of the fall data are summarized in the first section of this chapter. The second set of data includes information on the number of children to be served by the programs and expected sample attrition, comparability of PDC and comparison Head Start groups on background characteristics, and initial comparability of groups on child performance measures. Recommendations stemming from these sources of information are presented in the second part of this chapter.

Other considerations have also been taken into account in preparing the recommendations. These include the tentative nature of some of the sample characteristics, the contribution of individual sites to an understanding of the implementation process, and the importance of a comprehensive Impact Study that examines parent, teacher, and institutional effects as well as child impact.

Suitability of the Instruments

Usefulness of the Present Battery

The instruments that provided the data for the present phase of the Impact Study were selected expressly because of their apparent relevance to the areas of social competence that PDC is designed to affect. Analyses of the internal consistency reliability and concurrent validity of the

instruments indicate that most of them, at most of the sites, seem suitable for assessing the behaviors of interest. Table 9 shows the decisions that have been made, based on these analyses, regarding future use of the instruments. Half the total number of instruments examined were judged adequate for future use with little or no further modification in content, administration, or scoring. While it is possible that continued efforts will be made to refine this first group of instruments, the task of refining those in the second group has immediate priority; the problems encountered with these instruments were judged to be soluble for one or more of several reasons:

- The instrument was judged to be reliable and valid in at least some sites, suggesting that the circumstances that caused it to fail at other sites may be correctable with further modifications;
- The instrument failed uniformly across sites due to circumstances that are correctable (e.g., confusing instructions);
- The instrument represents an attempt to tap important behaviors that develop later, so better results may be expected in the future as the children mature.

The three scales excluded from future use were found either to provide no useful discrimination among children (due to an extremely high percentage of children obtaining high scores) or to present difficulties in administration or scoring for which there are no acceptable solutions.

There is ample reason to believe that the set of instruments now proposed will be capable of detecting growth along some of the social competence dimensions of greatest interest. The battery as it is now constituted appears strongest in the cognitive-language area. It is augmented by a classroom observation system that focuses on social/interactive dimensions and other measures discussed in the following section.

Plans for Future Refinement of the Battery

Two instruments were administered in fall 1976 that have been described (see Appendix A) but not discussed elsewhere in this report. The PDC Classroom Observation System and the PDC Child Rating Scale are being developed especially for

Table 9

Conclusions Regarding Suitability of the Instruments

Measures to be Retained without Change. These measures were judged reliable and valid in all sites and will continue to be used in the evaluation with little or no modification:

Bilingual Syntax Measure - English
Bilingual Syntax Measure - Spanish
Block Design (WPPSI)
Verbal Fluency (MSCA)
Verbal Memory, Parts 1 and 3 (MSCA)
Draw-A-Child (MSCA)
POCL (High/Scope Foundation)

Measures to be Retained Provisionally. These measures are being modified to correct problems detected during data collection or during data analysis relevant to reliability and validity. They will continue to be used in the evaluation until reliability and validity are reevaluated:

Conceptual Grouping (MSCA)
Say and Tell (CIRCUS)
Arm Coordination (MSCA)
Leg Coordination (MSCA)
PIPS-Solutions

Measures to be Discontinued from the Evaluation. These measures were found to be inappropriate for the age levels spanned by the evaluation:

Block Building (MSCA)
PIPS-Locus of Control[1]
Stephens-Delys Reinforcement Contingency Interview

Measures to be Retained for Further Development. Two measures will continue to be retained:

PDC Classroom Observation System
PDC Child Rating Scale

[1]In the use of the PIPS, it is only the Locus of Control scoring procedure that will be discontinued; the PIPS will continue to be used, but with only the "solutions" score.

purposes of this evaluation (unlike the other instruments in the battery, which had been tested previously in their present forms), and it was deemed inappropriate to present premature reports on their psychometric properties or to judge group comparability on the basis of the scores they provided. These instruments will be administered again in spring 1976, however, and their utility will be reviewed in the next interim report. If they are judged to be reliable and valid, they will provide additional perspectives on the development of children's social competence, particularly in the social-emotional area.

Other promising instruments will be used experimentally in the future with an eye to complementing the present battery, or to replacing complex instruments with simpler ones. (For example, a measure of productive language will be pilot-tested in three sites this spring, and a measure of children's attitude toward school will also be obtained.) In addition, future analyses of test data will attend to the possibility of eliminating instruments that yield redundant information. Although it is valuable to provide for some redundancy of measurement as a way of certifying that the measurements obtained are valid, it is also important that the PDC evaluation not intrude on the program any more than is strictly necessary. Thus, in the child testing phase of the evaluation, where assessment activities directly involve the child and the teacher, special efforts are being made to maximize efficiency of measurement.

Recommendations Based on Suitability of the Samples

Since the current three-year evaluation effort is conceptualized as a feasibility study for conducting a five-year longitudinal impact evaluation, the six recommendations made here are concerned primarily with procedures that will best provide the essential information for judging the potential for a longitudinal study. There is not sufficient information at this time for making those long-range projections, but on the basis of fall 1976 data, recommendations for the longitudinal study will be made in March 1977.

Although the central focus of these recommendations is the Impact Study, and although this report comprises child impact data, the recommendations are made within the context of the entire Developmental Continuity project and the

comprehensive evaluation of that project. As a Head Start demonstration program, PDC is designed to demonstrate the linkage of Head Start and elementary programs to achieve greater continuity. An understanding of how this happens is critical to future decisions about programs of this nature. Thus, the process evaluation of PDC was seen by the original planners in the Office of Child Development as having an importance equal to that of the Impact Study. In fact, about fifty percent of the total evaluation effort is devoted to the process evaluation, including an analysis of the initial planning process, description of start-up and implementation activities, and assessment of the extent to which programs are successful in implementing the total concept of PDC (see Figure 1, p. 3).

1. *The PDC Impact Study should be continued at all fourteen sites.*

When all factors considered in this report are taken together with overall considerations surrounding the evaluation, there seems to be every reason for retaining all sites for participation in the Impact Study. Within the Impact Study, the assessment of impact on children is just one facet, albeit a very important one. The goals of PDC relate to changes in parents and teachers and to changes in the schools as institutions. These are important goals, and procedures have been developed for assessing program impact in these areas as well.

Even if the assessment of child impact were impossible (which does not appear to be the case), there would still be value in continuing the Impact Study. It is also felt that each site contributes important dimensions to the assessment of the implementation process and that the capacity of the evaluation to answer questions about relationships between implementation and impact will be enhanced by making use of all possible data. Consideration should also be given to examining child impact by means other than testing in sites where testing is judged to be infeasible.

2. Child testing should be continued at all but two of the sites.

The conclusions regarding child testing at each site are summarized in Table 10. After considering the factors of sample size and attrition, comparability of PDC and comparison groups, and suitability of the measures at each site, the fourteen sites clustered into four categories that describe the nature and extent of the problems with one or more of these factors.

Three sites seemed to have few problems: California-English, Colorado and Washington. If these sites continue with current plans for Head Start enrollment and if the attrition rate does not substantially increase, there should be little difficulty in conducting within-site analyses of the impact of PDC on child social competence.

In six sites (Georgia, Iowa, Maryland, Michigan, Utah, and West Virginia) there is little problem with the expected sample size, but differences between the PDC and comparison groups raise some concerns about the feasibility of within-site child impact analyses. The following summarizes the situation at each site:

- Georgia--the apparent imbalance on background characteristics is based on a comparison of Head Start and elementary children, since there is no contemporaneous comparison group; the elementary school enrolls a wide range of children, including those who would not meet Head Start eligibility requirements. The possibility of selecting from the total elementary sample a subset that best matches the PDC Head Start sample will be explored.

- Iowa--group differences were found in the analytic sample on three background variables; there was a higher proportion of whites in the comparison group than in PDC and a higher proportion of PDC children had previous preschool experience; the groups differed on only one performance measure.

Table 10

Summary of Recommendations for Continuing the Evaluation of Child Impact

	Site:															
	ARIZONA	CALIFORNIA-English	CALIFORNIA-Spanish	COLORADO	CONNECTICUT	GEORGIA	FLORIDA	IOWA	MARYLAND	MICHIGAN	NEW JERSEY	TEXAS-English	TEXAS-Spanish	UTAH	WASHINGTON	WEST VIRGINIA
Prospects for within-site analysis:																
1. Few problems exist for continuing the assessment of child impact.		X		X											X	
2. Child impact assessment possible if PDC-comparison comparability can be improved.					X	X		X	X	X	X			X		X
3. Child impact assessment possible if Head Start enrollment is increased or if attrition lessens.			X		X						X	X	X			
4. Problems with sample size, comparability, or instrumentation are severe and assessment of child impact might best rely on other methods.	X						X									
Prospects for combined-site analysis:																
1. Few problems exist for the assessment of child impact.		X	X	X	X[a]	X[b]		X	X	X	X	X	X	X	X	X

[a]Only PDC children could be aggregated because PDC and comparison children are not distinguished at the Head Start level.

[b]; analysis would not include the comparison children.

- Maryland--the full samples differed on important background characteristics (the PDC group had a higher proportion of black children, a lower proportion of Spanish-speaking children, and a higher proportion of whites); in the analytic sample, all of the PDC children had prior preschool experience while less than one-third of the comparison group had prior preschool. There were no group differences on performance variables.

- Michigan--the analytic samples differed on four background variables; the PDC group had a smaller proportion of black children, the children had more siblings, mothers were more likely to be employed, and families were more likely to have fathers present; the group differences in SES and ethnicity present more serious problems for this site than for any of the others.

- Utah--differences were found on three background variables and on one performance measure; the PDC children were slightly older than the comparisons, their mothers were less likely to be employed, and families were more likely to have fathers present.

- West Virginia--differences were found on two background variables and two performance measures; the PDC group had a higher proportion of boys than the comparison group and PDC families were less likely to have fathers present.

As explained in Chapter III, these imbalances appear to be largely corrected when sites are aggregated, but for within-site analyses to be feasible, a closer match on background and performance variables should be obtained for Cohort 2. Recommendation 4 addresses procedures for achieving improved PDC-comparison matches.

At two sites (California-Spanish and Texas-English) estimates of likely attrition and the projected numbers of children to be served by the programs raise some concerns for the possibility of a five-year longitudinal study when the sample at each site is divided into Spanish- and English-dominant children. California has a small proportion of

Spanish-dominant children, whereas in Texas, English-dominant children are in the minority. Because of the importance of the bilingual/bicultural demonstration projects, it is recommended that testing be continued in these groups, even though it may be necessary to aggregate data from the two sites to achieve sufficient sample sizes for a child impact study (see also Recommendation 6).

Problems with both group differences and projected sample size were found for three of the sites (Connecticut, New Jersey, and Texas-Spanish):

- Connecticut--PDC-comparison differences were found on three background characteristics in the analytic sample; the PDC group had a higher proportion of Spanish-speaking children, a larger proportion of children who had prior preschool experience, and PDC families were more likely to have fathers present; there were no group differences on any of the performance measures. Although attrition suggests the need for 75 children in each group next fall, funding levels will permit the program to serve only 60 children in each group. Attrition is compounded by the problem of tracking comparison children to 12 different elementary schools.

- New Jersey--group differences were found on four background characteristics and one performance measure; the PDC group had a lower proportion of boys and a larger proportion of children who had prior preschool; PDC families were less likely to have a father present and PDC children had more siblings. Problems with sample size are greatest for the comparison Head Start group.

- Texas-Spanish--group differences were found on one background characteristic and two performance measures and the sample size required is considerably larger than the number of children it is possible for the program to serve.

Finally, there are two sites where a combination of factors leads to the conclusion that continued child testing would not be the best use of evaluation resources. In Arizona, expected sample sizes are extremely small, but even

more important, adequate measures for assessing social competence of Navajo-speaking children simply are not available. In Florida, although there were relatively minor differences between groups on background characteristics and on one performance measure, the key consideration in this recommendation is the sample size. The program is designed to serve a small number of children and project attrition over a five-year period appears to be high. (It should be pointed out that, although school records indicate high attrition between kindergarten and third grade, recent information from the program indicates that school-wide attrition may not accurately reflect attrition within the Head Start migrant population.) At each of these sites, it would be best to focus the evaluation effort on those activities (e.g., interviewing parents, surveying teachers, assessing implementation) that can be most responsive to the unique characteristics of the site.

3. *Where sample size has been identified as a problem, sites should be encouraged to increase the number of children for fall 1976 in line with the numbers recommended.*

Sample size is likely to be a problem in at least five sites, but the evaluation could be improved at all sites if larger numbers of children were enrolled. At the same time, sites should be encouraged to provide more recent and more reliable data on attrition from Head Start to third grade to permit the most realistic assessment of necessary sample size.

4. *A high priority should be set on encouraging sites to recruit children in a way that will maximize the match between the PDC and comparison Head Start children.*

The data presented by site in Interim Report III should be used to guide recruiting decisions, and whenever there are more children available than can be served by the program, children should be enrolled in such a way as to balance PDC and comparison Head Start programs on background characteristics. It is also recommended that High/Scope staff work with OCD staff to develop procedures to assist the sites in following

this recommendation. Sites should also be encouraged to assist in maintaining complete records on the children enrolled so that information on background characteristics will be as complete and accurate as possible.

5. *The comparability of PDC and comparison groups at each site must be reevaluated when fall 1976 data are available.*

The importance of the match on certain background characteristics cannot be overemphasized. Factors such as SES, prior preschool experience and ethnicity are known to relate to performance measures, and it cannot be assumed, a priori, that statistical adjustments will restore balance completely. Problems of unmatched samples are especially critical where sample sizes are small, as is likely to be the case with the expected attrition in most of the sites.

If a site shows a small degree of imbalance on a small number of variables, a site-level analysis of children's test scores might still be possible so long as PDC and comparison samples are of adequate size. Larger imbalances on a larger number of variables for a given site may prohibit site-level analysis, but with adequate sample size, the site's data may be usefully pooled with data from other sites for an aggregate analysis. However, if a site has extreme problems with both comparability and sample size, it is unlikely that child test data from that site can contribute usefully to any analysis. After examining fall 1976 data, the feasibility of continuing the longitudinal study can be assessed.

6. *Special attention should be given to providing additional support to the two bilingual/bicultural demonstration sites that serve both Spanish- and English-speaking children.*

PDC is unique among national demonstration programs in its attempt to seriously address the needs of bilingual/bicultural children, and considerable effort has been devoted to developing both implementation and impact assessment procedures that are sensitive to the special characteristics of the programs and their children. At one BL/BC program (Colorado) there appears

to be no difficulty with projecting within-site analyses of child impact, but an evaluation of this site alone would not provide an adequate test of BL/BC programs, since all children are English dominant. The other two sites present some difficulties, partly because of insufficient numbers of children in each language group. At current and projected sample sizes, within-site analyses would only be possible for the English-speaking children in California; data from California and Texas would have to be pooled in order to assess impact on Spanish-speaking children. Although this may present conceptual problems because of differences among sites in the populations served and in the nature of the bilingual programs, the Office of Child Development should determine whether pooling data from these two sites would provide the necessary information on program effectiveness. If not, additional resources should be found to enable California and Texas to serve a larger number of children, if larger populations of eligible children exist who are not currently enrolled.

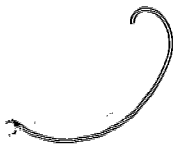
Conclusions

Two general conclusions emerge from the findings of this report. First, the child measures pilot-tested in fall 1976 are basically suitable for assessing the impact of PDC, at least at the Head Start level. By dropping a few measures, modifying others and pilot testing two additional measures this spring, it is believed that a sound measurement battery can be achieved.

The second conclusion is that the methodological problems assessed in this report, while real, are still manageable. If the recommendations can be followed (particularly with respect to reducing PDC-comparison group differences) there is every indication that a successful evaluation of the impact of PDC on children's social competence can be completed.

APPENDIX A

Descriptions of the Measures Selected



APPENDIX A

Descriptions of the Measures Selected

The test review process resulted in the selection of the following instruments:

Social-Emotional Measures

PDC Classroom Observation System
Preschool Interpersonal Problem-Solving Test (PIPS)
Pupil Observation Checklist (POCL)
Stephen-Delys Reinforcement Contingency Interview (S-D)

Psychomotor Measures

Arm Coordination [McCarthy Scales of Children's Ability (MSCA)]
Leg Coordination (MSCA)
Block Building (MSCA)
Draw-A-Child (MSCA)

Cognitive and Language Measures

Block Design (WPPSI)
Block Design (WISC)
Conceptual Grouping (MSCA)
Opposite Analogies (MSCA)
Verbal Memory (MSCA)
Verbal Fluency (MSCA)
Do You Know...? (CIRCUS)
Say and Tell (CIRCUS)
Bilingual Syntax Measure (BSM)

Other Measures

PDC Child Rating Scale
Height and Weight
Adult Language Check
Demographic Information Sheet
Wepman Auditory Discrimination Test

Each of these measures is described briefly below. For a more extensive review, see Interim Report II, Part B: Recommendations for Measuring Program Impact (1975).

PDC Classroom Observation System (High/Scope Foundation, unpublished). The PDC observation system was developed to provide information about children's classroom behavior along dimensions pertinent to the social-emotional goals of Project Developmental Continuity. The system focuses on aspects of an individual child's behavior, verbal or nonverbal, that reflect the child's attitude toward himself, and on the child's social competence as demonstrated in his interaction with peers and adults.

Using a time sampling method, trained observers observe each child for five minutes at four different times during the day and code their behavior into eight general categories. These categories include: "noninvolved," "involved," "interacts with peer," "uses peer as resource," "interacts with adult," "uses adult as resource," "expresses pride in personal achievements or attributes," and "dramatic play." A ninth category, "activity level," is included to provide information concerning the context in which these behaviors were observed. Each of these categories includes subcategories that are designed to identify the frequency and nature of specific behaviors within the general category.

Preschool Interpersonal Problem-Solving Test (Shure and Spivack, 1974). The PIPS attempts to assess the child's ability to name alternative solutions to a life-related problem--that of obtaining a toy from another child. Paper cut-outs of boys, girls and toys are used in presenting the problem. Among inner-city four-year-olds attending the Philadelphia Get Set day care program, those judged as better-adjusted by their teachers were able to conceptualize a greater number and a wider range of alternative solutions to real-life problems than were their more poorly adjusted classmates.

Pupil Observation Checklist (High/Scope Foundation, unpublished). This is a rating scale consisting of eleven 7-point bipolar adjectives derived from a similar scale used in the Home Start evaluation. The items are divided into two subscales, Child's Testing Behavior and Problem-Solving Behavior, and the ratings are all completed by the tester after he or she has administered all the other measures in the battery to a child.

Stephens-Delys Reinforcement Contingency Interview (Stephens and Delys, 1973). This instrument seeks to measure the extent to which a child believes that the behavior of others around him is contingent on his own behavior. The instrument was shortened to a version consisting of 12 items that ask questions such as: "What makes teachers smile?" Responses are coded internal if answered in a way that indicates attribution of control to oneself (e.g., "When I...") and external if the answer suggests that the behavior is under the control of someone else (e.g., "When Daddy...").

McCarthy Scales of Children's Abilities (McCarthy, 1972). These subtests consists of series of tasks tapping problem-solving, psychomotor, and conceptual abilities, and are similar to the Wechsler scales, but with emphasis on age-related maturational indicators. The particular McCarthy subtests used in the Impact Study are listed by category at the beginning of this section.

Wechsler Preschool and Primary Scale of Intelligence, Block Design (subtest) (Wechsler, 1967). The task requires reproducing (constructing) designs with flat colored blocks, either from the examiner's model or from a picture on a card. The measure taps problem-solving abilities, flexibility of response style, visual-motor organization, and execution.

Wechsler Intelligence Scale for Children, Block Design (subtest) (Wechsler, 1949). The task consists of reproducing (constructing) designs with colored blocks (cubes), either when modeled by the examiner or when presented on a card. The measure taps problem-solving abilities, flexibility in response style, visual-motor organization and execution.

CIRCUS subtest: Do You Know...? (Educational Testing Service, 1974). This is a general information test. The child chooses the picture which appropriately answers the examiner's question. This task taps the child's experience in a variety of areas (health, safety, social standards, consumer concepts).

Circus subtest: Say and Tell (Educational Testing Service, 1974). This test consists of two parts and taps children's language abilities. In the first part the child is given a pencil and asked attribute questions, e.g., "What color is it?"; in the second part the child is given two pennies and is asked to describe them. Scoring is based on categories of attributes which the child mentions.

Bilingual Syntax Measure (Burt, Dulay and Hernandez C., 1975). This test is designed to measure children's oral proficiency in standard English and/or standard Spanish grammatical structures. Simple questions are used with cartoon-type colored pictures to provide a conversational setting for eliciting natural speed. An analysis of the child's response yields a numerical indicator and a qualitative description of the child's structural language proficiency in standard English or standard Spanish. Responses are written down verbatim.

PDC Child Rating Scale (High/Scope Foundation, unpublished). This instrument is designed as a measure of social competence to be administered by the respective classroom teachers of the children rated. For each of the 39 items, specific behaviors such as "Uses words or wits to influence others" are rated on a 5-point scale according to frequency of occurrence ("Very frequently" to "Rarely). Summation of the ratings yields two aggregate measures: interpersonal competence and task competence (learning-to-learn ability).

Weight and Height. These two items of data are collected for all children in the sample. In most cases the tester personally weighs and measures the child, although in some instances Head Start records were used to avoid duplication of effort.

Adult Language Check. This measure is used in the bilingual/bicultural demonstration sites to obtain an indication of the languages the adults in the classroom use during their interactions with children. The interviewer sits in the classroom for a two hour period and records the language used by the teachers and aides approximately every five minutes.

Demograhic Information Sheet. Demographic information such as years of mother's education, presence of handicap, previous preschool experience, number of siblings, occupation of parents, etc. was collected primarily from the Head Start records.

Wepman Auditory Discrimination Test. This instrument tests children's ability to discriminate between sounds.

Not all the above measures were used in all PDC sites. After the basic battery was selected, each site was asked whether any of the five optional measures (Wepman, Do You Know...? (CIRCUS), How Much and How Many (CIRCUS), How Words Work (CIRCUS), and Opposite Analogies (MSCA) were related to their specific goals and objectives and, if so, whether they wanted them included in the battery for that particular site. Four sites did request at least one site-specific measure. Information on which measures were administered in each site is given in Tables A-1 and A-2.

Table A-1

Instruments Administered to PDC and Comparison Students at Each Site

Child Measures	Arizona	California	Colorado	Connecticut	Georgia	Florida	Iowa	Maryland	Michigan	New Jersey	Texas	Utah	Washington	West Virginia
gual Syntax Measure ish Version)		X	X	X		X	X	X	X	X	X	X	X	X
gual Syntax Measure ish Version)		X*	X*	X*				X*			X*			
Building rthy Scales)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
ordination rthy Scales)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
l Fluency		X	X	X	X	X	X	X	X	X	X	X	X	X
A-Child rthy Scales)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Design I)		X	X	X	X	X	X	X	X	X	X	X	X	X
l Memory rthy Scales)		X	X	X	X	X	X	X	X	X	X	X	X	X
ordination rthy Scales)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
hool Interpersonal Problem- ng Task (PIPS)					X				X	X			X	X
ens-Delys Reinforcement ngency Interview (S-D)		X	X	X		X	X	X			X	X		
ptual Grouping rthy Scales)		X	X	X	X	X	X	X	X	X	X	X	X	X
nd Tell US)		X	X	X	X	X	X	X	X	X	X	X	X	X
t and Weight	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Other Measures														
lassroom Observation System	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Observation Checklist (POCL)	X	X	X	X	X	X	X	X	X	X	X	X	X	X
raphic Information Sheet	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Language Check	X	X	X								X			
ating Scale			X		X							X		X

nistered to Spanish-speaking children in these sites

Table A-2

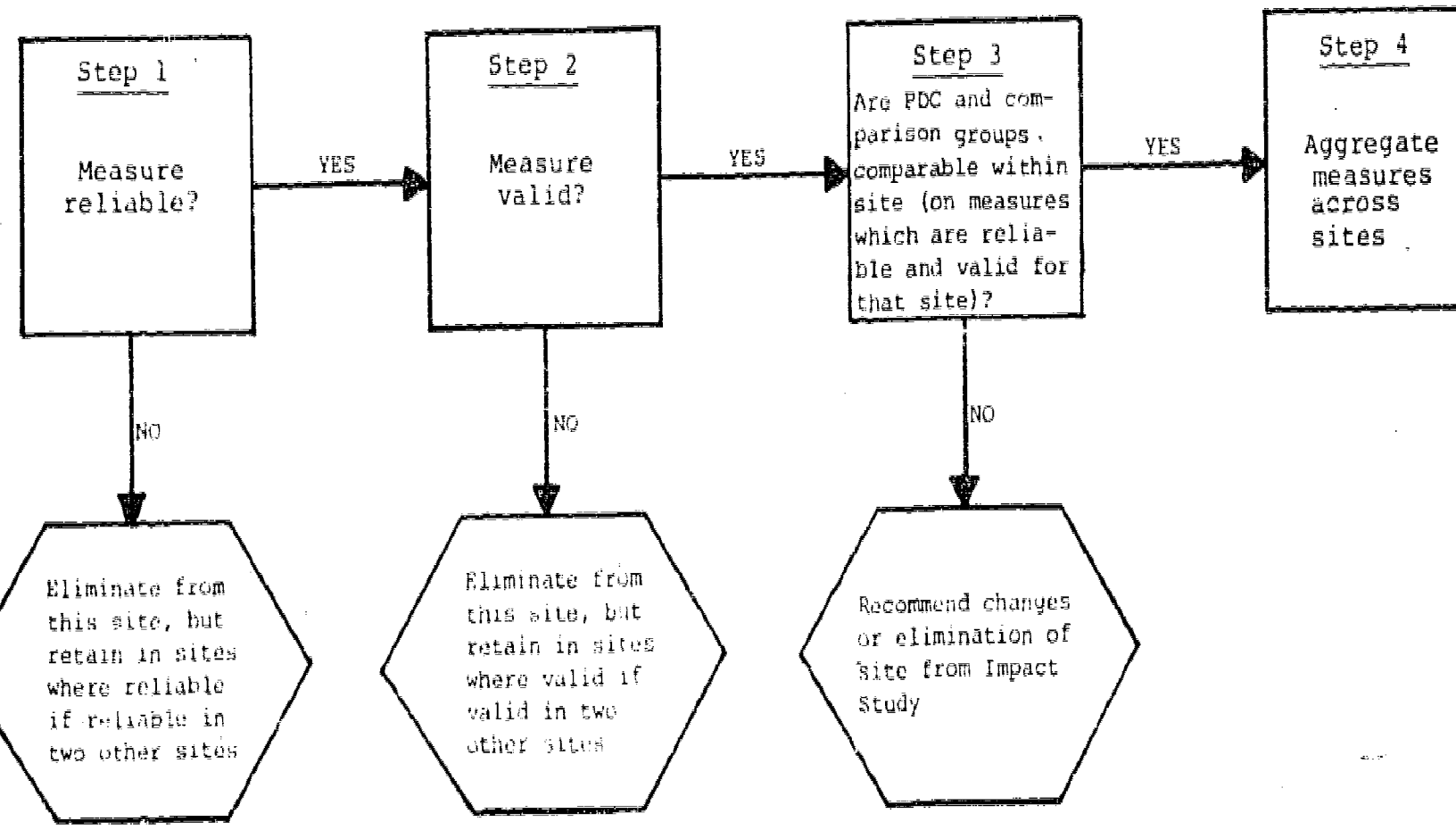
Site-Specific Instruments, Chosen Locally

Child Measures	Arizona	California	Colorado	Connecticut	Georgia	Florida	Iowa	Maryland	Michigan	New Jersey	Texas	Utah	Washington	West Virginia
Do You Know. . .? (CIRCUS)							X							X
Wepman Auditory Discrimination Test			X				X							
Opposite Analogies (McCarthy Scales)			X										X	

APPENDIX B

Flow Charts for the Analysis Procedure

Figure B-1
Overall Flow Chart for Basic Data Analysis



T = Table for this operation

= Operation

= Action decisions

= Go to

= YES/NO flows

= Mandatory flows

Figure B-2
Flow Chart for Step 1: Measure Reliable?

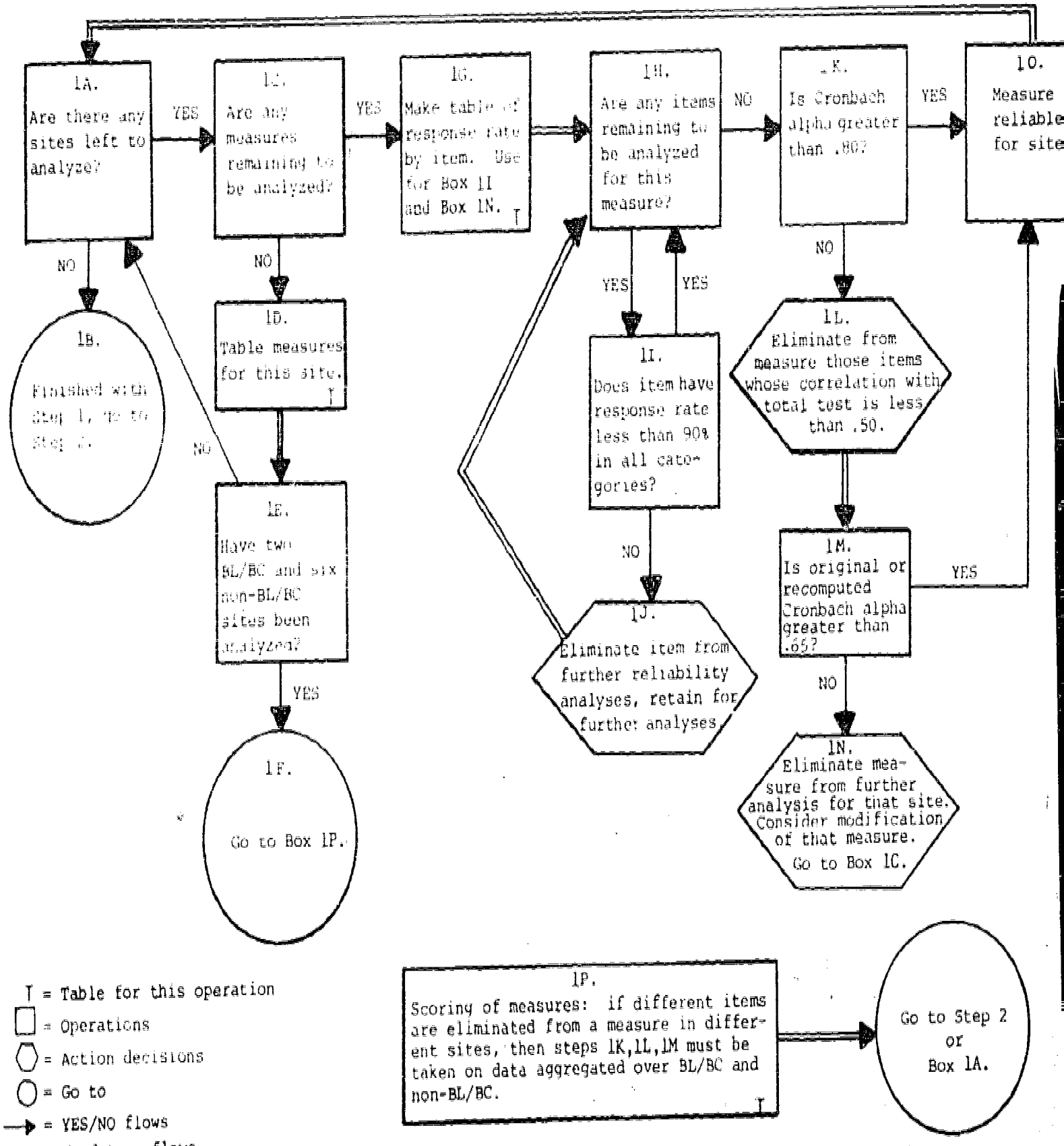


Figure B-3
Flow Chart for Step 2: Are the Measures Valid?

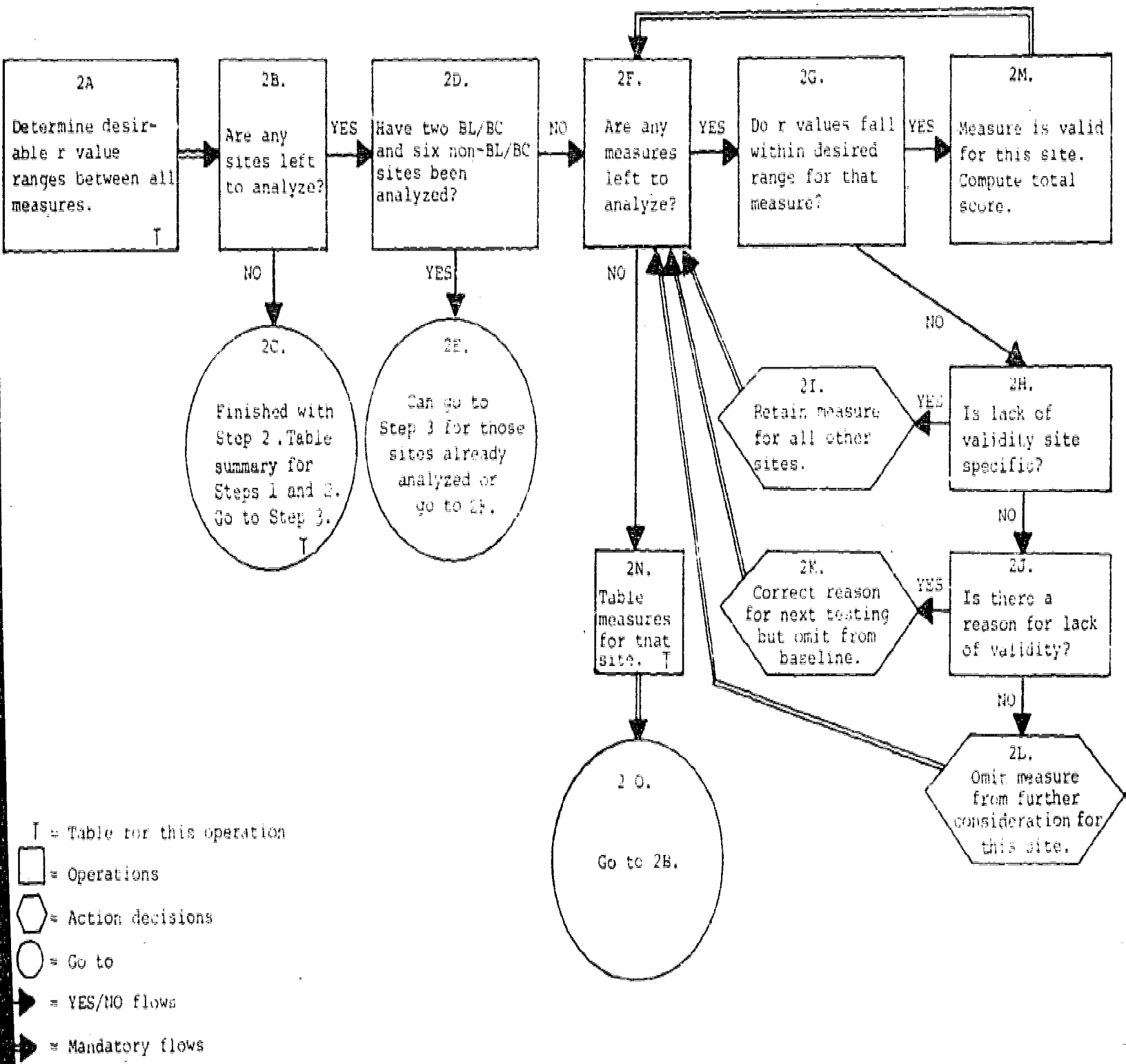
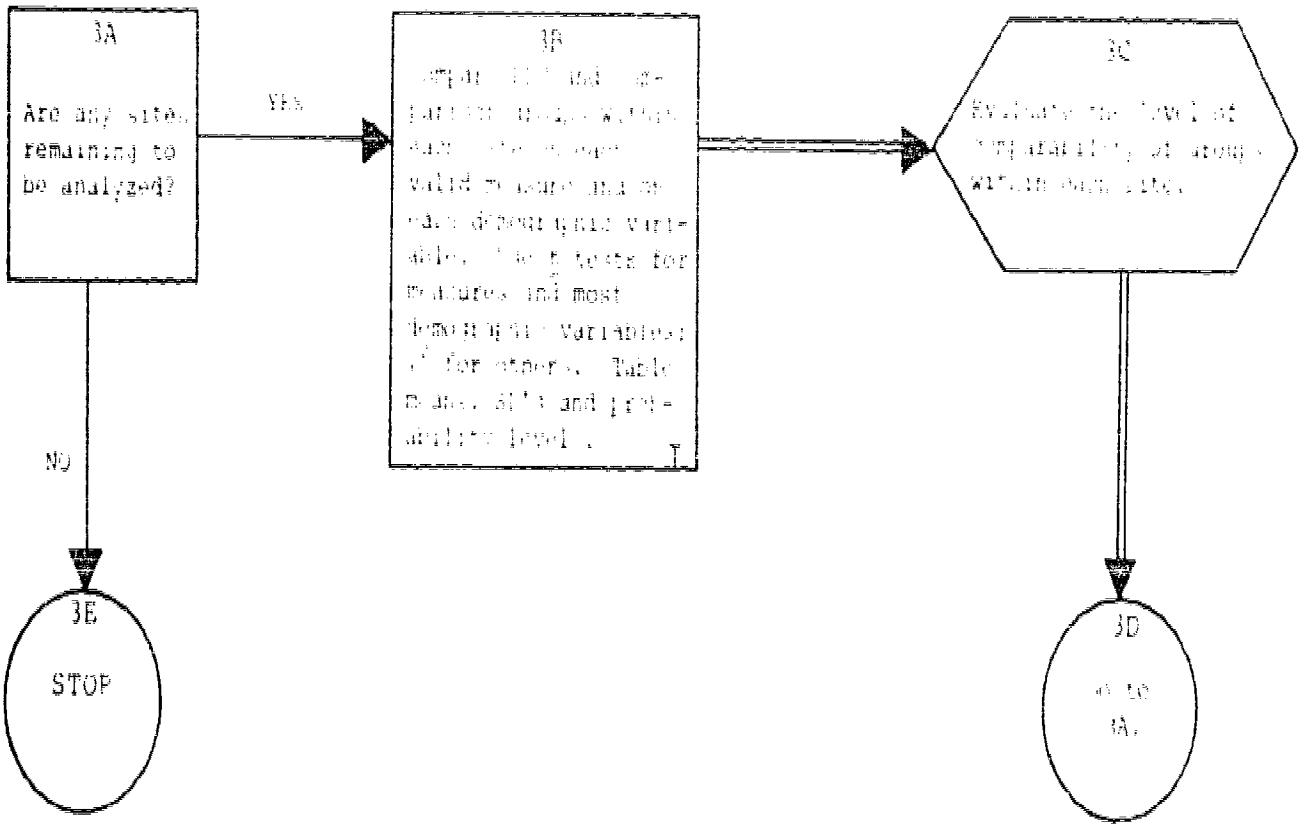


Figure B-4
Flow Chart for Step 3: Are the Groups Comparable?



= Table for this operation

= Operations

= Action decisions

= Go to

= YES/NO flows

= Mandatory flows