DOCUMENT RESUME

ED 094 884 PS 007 440

AUTHOR	Hantman, Stephen A.; Acland, Henry D.
TITLE	Final Report on the Follow Through City Data.
INSTITUTION	Huron Inst., Cambridge, Mass.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
PUB DATE	30 Jun 73
CONTRACT	OEC-0-72-0718
NOTE	56p.
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	*Academic Achievement; *Comparative Analysis; Comparative Statistics; Data Analysis; *Disadvantaged Youth; *Elementary School Students; Evaluation Criteria; Federal Programs; Group Norms; Lower Class; National Norms; *Program Evaluation; Reliability
IDENTIFIERS	*Project Follow Through

ABSTRACT

The purpose of this City Data study was to explore sources of data outside the National Evaluation study of Follow Through to determine the usefulness of such information in the overall assessment. The two major objectives of the study were to judge the representativeness of the Project Follow Through sample and to check on the credibility of information collected in the National Evaluation. The first part of the present study used data and information collected by sponsors and local sources and addressed issues of the credibility of conclusions about the achievement test success of Follow Through children. The second part was designed to use data and information collected from the National Title I survey, from local sources, and from the Stanford Research Institute data bank on Follow Through, and focused on the issues of the representativeness of the Follow Through group in comparison with the population of low income children enrolled in Title I. (CS)

FINAL REPORT ON THE FOLLOW THROUGH CITY DATA

STEPHEN A. HANTMAN

HENRY D. ACLAND

June 30, 1973

Prepared for:
The Office of Education
Follow Through Evaluation
Contract No. OEC-0-72-0718

TABLE OF CONTENTS

The purpose of the City Data study was to explore sources of data outside the National Evaluation study to determine the usefulness of such information in the overall assessment of Follow Through. Two major uses were anticipated: First, to judge the representativeness of the FT sample. If we found that the sample of tested FT children were broadly representative of the pouplation of lower-income children reached by compensatory education programs; and if effects of FT overall, or specific projects or models within FT, were found to help the children, then we could generalize with more confidence about the probable effects of FT models on the larger population of low-income children. We would be on firmer ground in our judgment that a successfully implemented FT model which worked in Brooklyn, N.Y. could be exported to Chicago and have the same salutary effect. If, on the other hand, we found that FT children who tested better in reading were different in important ways (e.g., race or income or family size) from children elsewhere, we would have less confidence in proposing that a successfully implemented FT model could produce the same desirable results in other locations.

The second anticipated major use of data sources outside the National Evaluation was to check on the "credibility" of information collected in the National Evaluation. The state of

instrumentation required to properly judge the effects of a vast and ambitious undertaking like FT clearly lags behind the broad and complex goals of the program. From the first generation of evaluation studies in compensatory education (roughly from 1965 to 1969), we came to be more fully aware of just how uncertain our knowledge of both instrumentation and of choice and execution of research design is; of just how many obstacles and difficulties even the best-planned and thought out evaluation scheme would encounter. One important lesson we learned was that wherever and however possible, we should seek to use a variety of data sources collected independently. Clearly no single source should be entirely trusted. But a convergence of information coming from a variety of sources collected by different groups which all gave a similar picture should be more trustworthy. Likewise, a serious divergence of findings would be important information enabling us to enquire more skeptically and in greater detail as to the reasons for the divergence. We would be less quick to judge a complex program a failure as the result of one evaluation using one type of measure of goal fulfillment.

The course of this Study of the usefulness of City Data has consequently broadened. Where once we believed that major reliance could be placed on data collected from a sample of large cities, we now are inclined to look beyond large cities to additional data sources; to all local education agencies (LEA's)

which have sites involved in the FT program; and to all sponsors who also collect information about the effects of the Program on children they enroll. A corollary of this broader scope of exploration has been a more restricted sense of the usefulness of any one source of data. Where once we felt that a small number of large cities could provide us with relatively reliable and valid information that would confirm or disconfirm findings from the National Evaluation, we are now less sure that such supplementary information, especially information coming from standardized achievement tests, can by itself be of much use in assessing program effects in the immediate term. This is true whether one is interested in the long or short-term effects, whether one wants data to be obtained and interpreted quickly or only after careful analysis.

Organization of the Report

This report is divided into two parts or substudies. Each substudy uses data from outside the National Evaluation to explore an important issue in the overall assessment of FT. The first part uses data and information collected by sponsors and local sources and addresses issues of the credibility of conclusions about the achievement test success of FT children. The second part was designed to use data and information collected from the National Title I survey, from local sources and from the SRI data bank on FT, and focuses on the issues of the

the representativeness of the FT group in comparison with the population of low income children enrolled in Title I. A major appendix (Appendix I) provides the results of the "Availability of Local Data" survey conducted by Huron and collected from all local education agencies with FT sites in 1972-73. This appendix should be of use to FT-Washington or to outside researchers who seek to explore achievement testing issues that arise out of the FT program. Appendix II contains details of our recommendations on guidelines for sponsor annual reports submitted earlier to FT-OE.

PART I

SPONSOR AND LOCAL DATA AND THE NATIONAL EVALUATION

Apart from data collection activities specifically undertaken as part of the National Evaluation of FT, the sponsors represent a potentially valuable and prolific source of information about the effects of FT on children, their families and their schools. The major emphasis of a sponsor's time and budget is properly directed toward program activities and not toward evaluation. Nonetheless, all sponsors have at least thought about process and product evaluation and some sponsors have done a considerable amount of independent data collection and analysis of the effects on children, families and schools. Sponsors vary widely in the amount and type of evaluation information they collect.

Sources of information from sponsors. Since 1971, The Huron Institute has received copies of the annual reports sponsors are required to submit to FT-Washington. We have tried to obtain and examine all sponsor annual reports for the year 1971-1972 for this study. This has been our major source of sponsor information about the measured effects on children, families and schools. A list of the sponsors whose annual reports we examined is shown in Table I. With the exception of the University of Arizona, these include all the "major" FT sponsors: They cover a large majority of all FT sites.

We looked at each annual report for evidence of measured effects, primarily on children and secondarily on families and schools. We placed special emphasis on measures of school achievement not because it is the only effect of interest and importance. Rather, our interest focuses on this measure because it permits us to make comparisons with the National Evaluation data on achievement, and becuase improved school achievement is certainly one of the most significant goals of the FT program. We chose the year 1971-72 because it is the latest year annual reports would be available to our study and because we had greater faith in the national evaluation test and in the meaningfulness of norms from the Metropolitan Achievement Test in the Spring of 1972. Finally, we expected that sponsor evaluation efforts would show the greatest refinement and sophistication in the latest possible year we could examine, especially for those sponsors who had been associated with the FT program since 1968 or 1969.

In addition, we sent letters to all sponsors whose projects were included in the 1973 Interim Report Analysis of Selected FT Data, prepared by Abt Associates. We asked each sponsor to comment on and provide supporting documentation "which either would strengthen or weaken the conclusions they reached about the achievement of FT pupils in your model." We suggested examples of such evidence: achievement tests collected independently either by the sponsor or the LEA for FT pupils; other test results and measures of pupil progress and development;

examples of problems of test administration in the National Evaluation; irregularities in the SRI data bank.

Having examined all the available sponsor annual reports, we chose to concentrate attention on the data from three sponsors, the University of Oregon, the University of Kansas, and the University of Pittsburgh. These sponsors were selected because they had the most complete and continuous achievement testing programs of all major sponsors and because they represented an approach to early childhood and primary education which has been called "structured." There is developing evidence from experimental preschool and primary programs and from Head Start and Follow Through Planned Variations which suggests (with some ambiguity) that more "structured" programs in preschool and early primary years seems to produce enhanced academic achievement. (For a complete review, see White, et al., 1972, Part III.) We felt that a more thorough study of sponsors who employ a more structured approach might provide valuable evidence confirming or disconfirming this tentative pattern of findings about "structure." Early in our investigation, the three structured sponsors were asked to send to Huron as much additional data as they had available on the results of their 1971-72 testing program.

We report the results of this substudy in two sections. The first section discusses 1971-72 annual reports from sponsors whose reports were made available to us. The second section

discusses in greater detail the achievement test findings from the three most highly structured sponsors.

1. ANNUAL REPORT SUMMARIES

Table I
1971-72 Sponsor Annual Reports Examined

Bank Street College of Education Approach
Behavior Analysis Approach (University of Kansas)
California Process Model (California State Department of Education)
Cognitively Oriented Curriculum Model (Hi/Scope Educational Research Foundation)
Cultural Linguistic Follow Through Approach (University of California, Riverside)
EDC Open Education Follow Through Program
Florida Parent Education Model (University of Florida)
Individualized Early Learning Program (U. of Pittsburgh)
Language Development (Bilingual) Education Approach (Southwest Educational Development Laboratory)
Mathemagenic Activities Program (University of Georgia)
Responsive Educational Program (Far West Laboratory for Educational Research and Development)
University of Oregon Engelmann-Becker Model

Our overall impression of the annual reports that we examined is that they are of uneven quality. Often they contain large amounts of scattered, undigested, uninterpreted and unanalyzed information about children, families and schools. Some variety is wholesome and expected for, after all, they are operating in uncharted territory. There is a dearth of solid knowledge or even of concepts and frameworks which point out information that is relevant. On the other hand, it strikes one even more strongly that the format and content of many reports are such that it is hard to imagine making any headway so long

as information continues to be reported in such a scattered and disorganized fashion.*

Annual Reports for 1971-72 by Sponsor

We comment below on specific annual reports of six major FT sponsors, leaving a discussion of the three most structured sponsors for the second section.

Bank Street. There is no evidence of any achievement testing of children in the Report. The only evidence of any kind of measurement activity is the "Analysis of Communication in Education" (ACE) instrument, developed by the sponsor for quantifying aspects of the open classroom. Information provided about reliability is insufficient for assessment. No mention is made about validity. No references are provided about other studies which used the same instrument. The study which reported the use of the ACE instrument covered all 14 Bank Street FT sites, representing 78 classrooms and 468 hours of observations. While details about research methodology are inadequate, the little information provided casts doubt as to the meaningfulness of the comparisons presented. The non-FT group seems to have been opportunistically selected from two sites without any evidence of representativeness or comparability of children or teachers at these non-FT sites. FT classrooms were

---

* Appendix II contains a letter we sent to Ms. Frieda Denemark detailing our suggestions for improvements in the format and content of sponsor annual reports which would make them more useful for research and evaluation purposes.

selected because of effective implementation of the Bank Street approach. Finally, the sampling of time for measurement of classroom interaction was not representative. Only one day was observed, and teachers were notified in advance when they would be observed. There is good reason to believe that observers were aware of which classrooms were part of the sponsor sample. Thus the results on this instrument, while of possible use to the sponsor and the local site people, are useless for FT research and evaluation purposes. Bank Street should be encouraged to use this instrument under improved experimental conditions if any valid generalizations are to be made about the effects of the model at all Bank Street sites. Other sponsors with an open education approach should be encouraged to use the ACE instrument so that comparisons can be made.

Additional material supplied by Bank Street covered results of measurements on children or parents made before 1971 or after 1972. No systematic sampling was undertaken and not all sites were included, so it is not possible to assess the representativeness of sites selected or parents or children measured.

Educational Development Center. The EDC annual report for 1971-72 contains no pupil or parent measurement data. It reports "community data" by site. Appendices detail services rendered at each site and provide anecdotal evidence of pupil,

teacher and parent interaction with EDC personnel.

In one appendix (XIV) and in a letter sent to Huron, the research director of EDC FT explains why achievement tests are felt to be hostile to open education, particularly in the kindergarten. Alternatives are suggested: looking at goals of the program apart from basic academic skills; using teachers' records to record invidual child development. However, the annual report contains no documentation of teacher records used in this way, nor of measurement of other aspects of child development.

Far West Laboratory for Educational Research. In the early years of sponsorship, a considerable amount of standardized test baseline data and other measures of child cognitive and affective status were collected. However, after 1970, no pupil achievement data was collected. In 1971-72, pupils were not tested by the sponsor at all. Thus there is no way to compare sponsor and SRI-collected data.

Although the report promises local data collection, using measures devised by the sponsor, in fact the only data presented are IQ scores on the WPPI over two years for one community ending in 1971. In that study, no comparison groups were used.

Florida Parent Education Model. There are a great deal of data in this report, but none directly measures changes in pupil academic achievement. Most of the research work on the Florida Model done by the sponsor concerns measures of

change in teachers (Purdue Teacher Opinionnaire), in parent educators (How I See Myself Inventory and Social Reaction Inventory) and in mothers of target children (Home Environment Review). Children are measured on two "affective" instruments said to correlate with achievement, the "I Feel Me Feel" (IFMF) and the Cincinnati Autonomy Test Battery (CATB).

The IFMF yields scores on five factors (general adequacy, peer, teacher-school, academic and physical). It was administered pre-post in 11 centers. Children in the FT program made statistically significant gains on all five factors while non-FT children gained significantly on three of five factors. Results on the CATB were disappointing, inconclusive, and difficult to interpret, partly because of the very small sample size.

The format of this long report, without even a table of contents, makes understanding what is being reported extremely difficult.

Cognitively Oriented Curriculum Model. The Comprehensive Test of Basic Skills, an achievement test, was given only to third graders in FT and compared with a third grade comparison group tested the year the FT program began so as to minimize treatment contamination. The results for Cohort I of FT show either no difference (2 comparisons) or a difference favoring the control group (1 comparison). Other comparisons made for children who entered in the 1968-69 year showed similar results.

The trends, incidentally, almost consistently favored the controls.

On the Stanford-Binet IQ testing, conducted every year, statistically significant gains in mean score from fall to spring for the FT group were found in 5 of 10 centers. Statistically significant higher mean scores of third grade FT children compared with a group of non-FT third graders were found in 3 of 5 centers.

One appendix of the annual report contains sections which discuss results of local evaluation of achievement test scores in three sites. Unfortunately, none of these sites is part of the National Evaluation sample for the years that test scores are reported locally. Thus, comparisons between local and National Evaluation scores could not be made.

Mathemagenic Activities Program. There are no data whatsoever regarding any child measure, parent measure or classroom measures. The only data consist of ratings of sites on project assessment and implementation criteria.

Recommendations on Sponsor Annual Reports

Many sponsors included no measures of pupil development at all. Some sponsors measured aspects of pupil cognitive growth at a few sites, but only two measured academic achievement. Only Bank Street attempted to systematically measure classroom interaction for a sample of sites which was, unfortu-

nately, non-representative. For evaluation purposes the results are useless since Bank Street sites were compared with only two non-FT sites opportunistically selected. Only one sponsor, High/Scope, mentioned the results of local testing. Clearly, there is little meaning and interpretable information about pupil change in achievement or even cognitive development in the 1971-72 sponsor annual reports we examined. There is no reason to suspect that the annual reports not reviewed (i.e., those not included in Table I) are any different.

One can recognize that sponsors have varying goals and place different emphasis on aspects of child, family and school development. Many of these aspects are at present imperfectly measurable. However, measures do exist for many aspects of child and classroom development. One way to perfect imperfect measures, to learn more about the validity of such instruments, as well as to learn about the effects of program activities, is to use a limited number of the best existing measures systematically over a large number of subjects or sites.

There remains the criticism that measurement of children, as such, is hostile to the kind of experience that some models are seeking to create in their classrooms. We quote here from a letter sent to us by the EDC Evaluation Research Committee:

> The teacher following test directions talks a great deal to children but tells them nothing that would interest them. Our children are not accustomed to detailed directions about how and when to do their work. They are expected to proceed independently.

> Nor are they accustomed to sitting at separate desks for long periods of time with no communication with their fellow students.
>
> Many of the children in our program will be handicapped on this kind of pencil-paper test because we stress sharing, playing and working with a wide range of materials, art and music activities, sand, water and block activities, rather than workbook or mimeographed "test-like" materials.

The only proper response to such a critique is that a program model of this type should never have been included in a Planned Variation experiment. If the sponsor cannot imagine and then devise any type of measurement with validity and reliability which would not infringe on the child's accustomed mode of operation, then clearly the program model cannot be evaluated in an experimental situation.

In the letter to Ms. Denemark we detail our recommendations for improvements in annual reports (see Appendix II). We suggest there that sponsors be asked to specify in advance, in their proposals, what measures they intend to use. FT should provide the necessary funds and technical assistance to sponsors both in the formulation and in the execution of research design proposals and plans. Sponsors should be required to collect and report the results of testing undertaken at each site by the LEA for all children in the FT grades, broken down by FT and non-FT, with further refinement of the non-FT population if possible.

2. ACHIEVEMENT TEST FINDINGS FROM THE "STRUCTURED" SPONSORS

We report here comparisons of scores on achievement tests taken by the same children at about the same time in the more highly structured programs sponsored by the University of Oregon (Engelmann-Becker Model), the University of Kansas (Behavior Analysis Approach), and the University of Pittsburgh (Individualized Early Learning Program). All comparisons are made in the grade equivalent (GE) metric. While serious problems exist in the interpretation of grade equivalents, there is simply no other way to make comparisons given the information provided by the sponsors.

Using the GE metric to draw inferences about pupil growth or status involves several difficulties. Depending on the correlation of grade with achievement in the particular skill area measured, a GE six months behind the norm may represent a serious lack of achievement or just one or two questions missed. If the average scores are near the national norm, GEs fluctuate much more in relation to raw scores than at far out ends of the distribution. GEs for tests taken between points of standardization are linearly interpolated values, yet linearity of growth throughout the school year is just an assumption -- one which has little empirical support. Standard scores are thus far more desirable, but we had no data from which standard scores could be computed.

The three sponsors tested either all or a random sample of

the children on the Wide Range Achievement Test (WRAT) in the Spring of 1972. Additional comparisons from previous years of sponsor testing and from local testing are included, although emphasis has been placed on the spring 1972 testing point. The comparison is between a sponsor-reported (or locally reported) score on an achievement test and the SRI-administered Metropolitan Achievement Test (MAT). Only children in grades one through three are included since there are no validated GE's at the kindergarten level for the MAT.

The charts which follow present data on test scores for pupils in sites sponsored by the Universities of Oregon, Kansas and Pittsburgh, the highly structured sponsors. Each row represents site scores on achievement tests at one time of test administration for a specific cohort at the designated grade level. Sites are separated by single dark horizontal lines. Underlined are the GE are on the various tests. Thus, the first row of the charts in Table II gives the Spring 1971 test results from the second grade FT pupils at the Dayton, Ohio (University of Oregon) site. The tests compared are the WRAT reading subtest and the Stanford Reading Achievement Test. The SRAT was administered by the local district (note the column the information appears in) to 158 FT second graders as well as to a "comparison" group of 126 second grade pupils in schools adjacent to the FT project school in Dayton. The GE comparison shows FT pupils at grade 3.5 on the WRAT compared with grade 2.1 on the SRAT, Primary II. The FT students scores .1 GE

TABLE II

ACHIEVEMENT TEST COMPARISONS
SELECTED SITES

Sponsor/Site	Time	Grade (Years in FT)	Sponsor Test	n	Score (FT Children) Raw	Score (FT Children) GE	SRI Test	n (FT)	Score FT Raw	Score FT GE	Score NFT Raw	Score NFT GE	District Test	Sample & Number	Score FT GE	Score NFT GE	Notes
U. of Oregon																	
7.01 Dayton	S'71	2	WRAT - Reading	?		3.5							SRAT Primary II	158 (FT) 126 (adjacent schools)	2.1	2.2	
	S'71	1	WRAT - Reading	?		1.9							SRAT	198 (FT) 164 (adjacent schools)	1.6	1.6	
7.03 E. St. Louis, Ill.	S'72	2	WRAT - Reading	149		3.6	MAT - Rdg.	169	23	2.26	22	2.20					
							Total Rdg.	169	44	2.3	43	2.29					
			Arithmetic	149		2.7	Total Math	162	47	2.09	45	2.05					
	S'72	1	WRAT - Reading	186		2.7	MAT - Rdg.	151	23	1.86	17	1.57					
							Total Rdg.	151	46	1.86	36	1.62					
			Arithmetic	186		2.0	Total Math	151	31	1.57	20	1.19					
7.04 Grand Rapids	S'72	3 (5 yrs FT)	WRAT - Reading	91		5.8							MAT - Reading	77 (3 yrs in FT program)	2.9	2.4	(inner city children)
														29 (FT students who left FT)	3.1		

TABLE II (cont'd)

Sponsor/Site	Time	Grade (Years in FT)	Sponsor Test	n	Score (FT Children) Raw	Score (FT Children) GE	SRI Test	n (FT)	Score FT Raw	Score FT GE	Score NFT Raw	Score NFT GE	District Test	Sample & Number	Score FT GE	Score NFT GE	Notes
7.04 Grand Rapids (cont'd)			WRAT - Arithmetic	91		3.7							MAT - Math	77 29	3.2 3.2	2.8 (inner city)	
	S'71	2 (4 yrs in FT Same cohort as above)	WRAT - Reading		4.1	3.0 Control Project Read							MAT - Rdg.	135 (FT)	2.4	2.1 Control Project Read	
	S'70	1 (3 yrs in FT Same cohort as above)	WRAT - Reading Arithmetic			2.3 1.9							SAT Word Mng. ¶ Meaning Vocab. Arith.		1.5 1.5 1.5 1.5	1.5 1.6 1.8 1.7	
7.07 PS137K	S'72	1	WRAT - Reading Arith.	70 70		2.7 2.2	MAT - Rdg. T.Rdg. T.Math	70 70 72	22 45 28	1.8 1.9 1.5							
	S'72	2	WRAT - Arith.	67		2.7	MAT - T.Math	66	38	1.8							
7.08 Racine	S'72	2	WRAT - Reading Arith.	119 119		3.6 2.9	MAT - Rdg. T.Rdg. T.Math	115 115 115	27 50 60	2.51 2.52 2.60	23 45 65	2.29 2.46 2.77					
	S'72	1	WRAT - Reading Arith.	121 121		2.9 2.4	MAT - Rdg. T.Rdg. T.Math	117 117 117	28 58 39	2.25 2.37 1.90	18 39 30	1.65 1.73 1.53					

TABLE II (cont'd)

Sponsor/Site	Time	Grade (Years in FT)	Sponsor Test	n	Score (FT Children) Raw	GE	SRI Test	n (FT)	Score FT Raw	FT GE	NFT Raw	NFT GE	District Test	Sample & Number	Score FT GE	NFT GE	Notes
7.11 Tupelo, Miss.	S'72	3	WRAT -				MAT -										
			Reading	120		3.9	T.Rdg.	119	33	2.54	61	4.34					
			Arith.	120		3.5	T.Math	118	42	2.84	77	4.60					
	S'72	2	WRAT -				MAT -										SRI tested 5 classes while sponsor reported 4 classes
			Reading	91		3.5	Reading	114	20	2.08	34	3.11					
							T.Rdg.	114	38	2.15	64	3.14					
			Arith.	91		2.5	T.Math	114	43	1.93	74	3.15					
7.12 Uvalde, Texas	S'72	3	WRAT -				MAT -										SRI tested 6 classes while sponsor reported 7 classes
			Reading	117		5.0	T.Rdg.	122	36	2.67	54	3.89					
			Arith.	117		4.0	T.Math	122	52	3.34	64	3.96					
U. of Kansas																	
8.01 PS 77X N.Y.	S'72	1	WRAT -				MAT -										
			Reading	52	37	1.9	Reading	51	16	1.5							
							T.Rdg.	50	37	1.7							
			Arith.	52	24	2.4	T.Math	50	34	1.7							
	S'72	2	WRAT -				MAT -										
			Arith.	61	29	3.6	T.Math	60	55	2.4							
8.03 Phila.	S'72	1	WRAT -				MAT										Note: Difference of 77 in number of children reported
			Reading	315	37	1.9	Reading	239	21	1.8							
							T.Rdg.	239	44	1.9							
			Arith.	315	22	2.0	T.Math	238	37	1.9							
	S'72	2	WRAT -				MAT -										Note: Difference of 130-150 in number of children tested
			Reading	330	46	2.5	Reading	200	21	2.1							
							T.Rdg.	200	41	2.2							
			Arith.	352	26	2.8	T.Math	196	52	2.3							

TABLE II (cont'd)

Sponsor/Site	Time	Grade (Years in FT)	Sponsor Test	n	Score (FT Children) Raw	Score (FT Children) GE	SRI Test	n (FT)	Score FT Raw	Score FT GE	Score NFT Raw	Score NFT GE	District Test	Sample & Number	Score Raw	Score GE	Notes
8.04 Portageville, Mo.	S'72	2	WRAT - Reading	97	52	3.1	MAT - Reading	96	27	2.7	33	3.1					
							T.Rdg.	96	52	2.7	63	3.1					
			Arith.	97	28	3.2	T.Math	95	65	2.8	66	2.8					
	S'72	1	WRAT - Reading	95	37	1.9	MAT - Reading	95	21	1.8	22	1.8					
							T. Rdg.	95	43	1.9	45	1.9					
			Arith.	95	23	2.2	T. Math	95	36	1.8	33	1.7					
8.08 Mounds, Ill.	S'72	1	WRAT - Reading	117	40.18	2.16	MAT - Reading	112	21	1.76							
							T. Rdg.	110	44	1.84							
			Arith.	117	23.00	2.20	T. Math	109	36	1.80							
U. Pittsburgh																	
12.01 Elkins W. Va.	S'72	3	WRAT - Reading	112		4.4	MAT - T. Rdg.	81	57	4.00	38	2.84					Sponsor tested 112 children while SRI tested 81. Sponsor reports 124 pupils in program.
			Arith.	112		3.9	T. Math	81	68	4.16	53	3.33					
	S'72	2	WRAT - Reading	135	47.8	2.7							MAT - Reading	129	23.7	2.4	
12.03 Lock Haven, Pa.	S'72	2	WRAT - Reading	133	47.3	2.6							MAT - Reading	126	23.9	2.4	
12.04 Montevideo, Minn.	S'72	2	WRAT - Reading	143	61	4.5	MAT - Reading	144	37	3.66	36	3.58	MAT - Reading	143	37	3.0	
							T. Rdg.	144	71	3.80	69	3.60					
			Arith.	143	29	3.6	T. Math	144	79	3.44	77	3.31					

TABLE II (cont'd)

Sponsor/Site	Time	Grade (Years in FT)	Sponsor Test	n	Score (FT Children) Raw	GE	SRI Test	n (FT)	Score FT Raw	FT GE	NFT Raw	NFT GE	District Test	Sample & Number	Score Raw	GE	Notes
12.04 Monte-video (cont)	S'72	1	WRAT - Reading	44	50	2.9	MAT - Reading	130	33	2.56	28	2.20					Sponsor tested random sample taken from 6 rooms. However, only 2 class-rooms are in program.
			Arith.	44	24	2.4	T. Rdg.	130	64	2.58	52	2.21					
							T. Math	129	46	2.37	43	2.20					
12.05 Akron, Ohio	S'72	2	WRAT - Reading	146	42.9	2.3							MAT - Reading	144	19.4	2.2	

level below the comparison group on the SRAT, but both groups scores well below grade level (2.8 or 2.9).

Discussion of the Comparisons. The comparisons will be discussed mainly in terms of differences in grade equivalents for the same FT children taking achievement tests about the same time. However, it is possible, if one has faith in the comparability of the SRI-designated NFT group, or of the district-designated "comparison" group, or of the national norm tables, to make formal and informal comparisons between FT children and children not in FT.

FT children who take the WRAT at about the same time as another standardized achievement test score higher on the WRAT than on the other test. Overall, the advantage in GE for taking the WRAT seems to be about 1 year GE in reading and about 1/2 year GE in arithmetic (Table III). In one dramatic instance, the Grand Rapids (7.04) third grade, Spring 1972 reading test scores, the WRAT GE is 5.8 compared with the MAT GE of 3.0, obtained by local testing -- a difference of 2.8 grade level! In only one case do FT children ever do worse on the WRAT than on another standardized achievement test (in site 12.01, 3rd grade total math).

Several other items of interest should be noted. The number of children tested by the sponsor, by SRI or by the local district for the same FT cohort is not the same. Some difference is to be expected since the tests were not given at

TABLE III

DIFFERENCES IN GRADE EQUIVALENT
ON SPONSOR vs. LOCAL OR NATIONAL EVALUATION DATA

Reading Differences

First Grade		Second Grade		Third Grade	
Site & Year	Diff.*	Site & Year	Diff.*	Site & Year	Diff.*
7.01 (1971)	+0.3	7.01 (1971)	+1.4	7.04 (1972)	+2.8
7.03 (1972)	+0.8	7.03 (1972)	+1.3	7.11 (1972)	+1.4
7.04 (1970)	+0.8	7.04 (1971)	+1.7	7.12 (1972)	+2.3
7.07 (1972)	+0.8	7.08 (1972)	+1.1	12.01 (1972)	+0.4
7.08 (1972)	+0.6	7.11 (1972)	+1.4		
8.01 (1972)	+0.3	8.03 (1972)	+0.3		
8.03 (1972)	0.0	8.04 (1972)	+0.4		
8.04 (1972)	0.0	12.01 (1972)	+0.3		
8.08 (1972)	+0.4	12.03 (1972)	+0.2		
12.04 (1972)	+0.3	12.04 (1972)	+1.5		
		12.04 (1972)	+0.8		
		12.05 (1972)	+0.1		

Arithmetic Differences

First Grade		Second Grade		Third Grade	
Site & Year	Diff.*	Site & Year	Diff.*	Site & Year	Diff.*
7.03 (1972)	+0.4	7.03 (1972)	+0.6	7.04 (1972)	+0.5
7.04 (1970)	+0.4	7.07 (1972)	+0.9	7.11 (1972)	+0.6
7.07 (1972)	+0.7	7.08 (1972)	+0.3	7.12 (1972)	+0.7
7.08 (1972)	+0.5	7.11 (1972)	+0.5	12.01 (1972)	-0.3
8.01 (1972)	+0.7	8.01 (1972)	+1.2		
8.03 (1972)	+0.1	8.03 (1972)	+0.5		
8.04 (1972)	+0.4	8.04 (1972)	+0.4		
8.08 (1972)	+0.4	12.04 (1972)	+0.1		
12.04 (1972)	0.0				

---

* + means sponsor G.E. higher than comparison;
- means sponsor G.E. lower than comparison.

TABLE III (cont'd)

Summary - Reading Differences (number of sponsor sites in parentheses)

Sponsor	First Grade	Second Grade	Third Grade
U. of Oregon	0.66 (5 sites)	1.38 (5 sites)	2.17 (3 sites)
U. of Kansas	0.18 (4)	0.35 (2)	---
U. of Pittsburgh	0.30 (1)	0.58 (5)	0.40 (1)
Average over all sites	0.40	0.90	1.70
			Overall average 1.0

Summary - Mathematics Differences (number of sponsor sites in ( ))

Sponsor	First Grade	Second Grade	Third Grade
U. of Oregon	0.50 (4 sites)	0.58 (4 sites)	0.60 (3 sites)
U. of Kansas	0.40 (4)	0.70 (3)	---
U. of Pittsburgh	0.00 (1)	0.10 (1)	-0.30 (1)
Average over all sites	0.40	0.57	0.38
			Overall average .50

exactly the same time. Large and unexplained differences are noted in the far right column for site 7.11 (2nd grade, spring 1972), site 7.12 (third grade, spring 1972), site 8.03 (first and second grade, spring 1972), site 12.01 (third grade, spring 1972) and site 12.04 (first grade, spring 1972). When differences are so large, the possibility of nonrandom "attrition" is certainly present and should be investigated. We have called these discrepancies to the attention of SRI and the sponsors and they are attempting to resolve them.*

Looking in detail at the sponsor contributions to the overall GE differences (Table III, Summary), it appears as though the differences between sponsor WRAT and the National Evaluation MAT scores is far greater in Oregon sites than in Kansas and Pittsburgh sites. But even among the latter two sponsors, differences are sometimes considerable.

The difference in GE between FT and a comparison group, either the National Evaluation's NFT group or a locally created comparison group, usually favors the FT with the startling exception of Tupelo, Mississippi (7.11) which must be suspected of initial non-comparability. But the differences between FT and a comparison group are rarely as large as differences within a single FT group as a result of merely taking a different achievement test. The implications of this pattern may be profound.

---

* The difference in site 8.03 (Philadelphia) may be due to a decision by SRI not to test all classes at this site since the N is so large.

What it might suggest is that measured increase in achievement, which we are trying to assess in the National Evaluation, may be more a function of the particular test administered than of the Program itself. If greater differences in GE are found between achievement tests administered to the same children at the same time than are found between different children, FT compared with non-FT, it becomes clear than any conclusion arrived at concerning the effects of FT using just one achievement test may be a conclusion not about the effects of the FT program but about the sensitivity of that particular test to the effects of FT or of a particular model in FT.

For example, one can properly infer from the WRAT that FT children (in the highly structured models we have comparison information on) generally score well above grade level in the first grade (reading, 2.5; arithmetic, 2.2) and in the second grade (reading, 3.3; arithmetic, 3.0). In the third grade they are at grade level in arithmetic (3.8) but still one year above grade level in reading (4.8). For the same children on the MAT total reading and total math subtests, the children are above grade level in total reading (2.1) and at grade level in total math (1.8) in the first grade. They are behind grade level in the second grade (2.6 for total reading and 2.4 for total math) and well behind in the third grade (3.1 for total reading and 3.4 for total math).

Considerations in Interpreting the Findings. If it is

granted that the comparisons presented above are intriguing enough to pursue -- given that the differences in a single group of FT children are typically much greater than differences between FT and non-FT -- we ought to speculate and, if possible, investigate why this is so and whether this affects the credence we give to the MAT, the WRAT or to any standardized test of achievement.

To anticipate the conclusion we have reached, it has created greater skepticism than we had initially about the use of all standardized achievement tests in making meaningful assertions about the affect of compensatory programs like FT on school achievement. Among other things, it further suggests to us that no single standardized test ought to be relied on in forming a complete judgment about the effects of the FT program on children's achievement in school.

(1) Differences in test content: The simplest explanation for the WRAT-MAT reading discrepancy (which covers the majority of comparisons in Table II) is that the tests measure different skills, especially in reading, where the GE discrepancy in tests is greatest.

A brief inspection of the MAT Primary I and II and Elementary batteries reveals important differences in the reading skills tested as between the MAT and WRAT. The WRAT requires that a student be able to name and recognize letters and then read aloud individual words of varying difficulty. In the Primary I battery, for example, the total reading score consists

of the sum of score from two subtests -- word knowledge and reading. Word knowledge requires an ability to read words silently, understand their meaning(s), and then interpret pictures, matching the appropriate word meaning with the pictures. It also requires many complex abilities in test-taking and following directions. Reading, in Primary I, requires comprehension of sentences, together with matching sentences with pictures. The child is also required to read stories, answer questions based on the literal meaning of the stories and draw inferences from the stories to answer other questions not based on the literal meaning of the stories.

The MAT total reading score obviously demands a large number of complex skills. Many of these skills require abilities to interpret pictures and make other kinds of judgment and above all, ability and motivation to follow directions. These skills are judged by the test publishers and by some authorities in reading instruction to be a crucial part of what it means to read in the early primary grades. Other authorities disagree, believing that the essence of early reading ability is the decoding of words and comprehension of word meaning. The debate between these two interpretations of reading, as exemplified by the MAT and WRAT tests, cannot be resolved by pointing out the predictive validity of the MAT, which is considerable. The MAT may test important school-related skills which predict well to future school achievement, but whether the

MAT reading subtests are efficient and accurate tests of reading ability is not an issue in predictive validity. To put it another way, it may be that children who can read well will not score high on the MAT but will score high on the WRAT. If these children do not succeed in later school work, it may be because they lack other skills which the MAT reading test taps, but not because they can't read well. Such children will not need a compensatory reading program. Granted the importance of succeeding in terms of existing public school criteria, they will need entirely different kinds of instruction. Thus, to diagnose reading deficiency from a low score on the MAT would be wasteful of resources in compensatory education.

(2) Differences in test administration: The WRAT as administered by the highly structured sponsors is intended to be given individually; the MAT is group administered. It is likely that children who might not be motivated to perform to capacity in a group situation might do so when tested in a one-to-one relationship, especially if they know and trust the tester. It is also likely that children would be more easily discouraged from attempting items if directions are complex and incorrectly or incompletely understood. The MAT, as mentioned before, makes far greater demands on the child in following complicated directions both in the interpretation of questions and in the marking of responses.

We solicited comments from sponsors about the procedures used by SRI in test administration in the Spring of 1972. Did

the children understand the directions? Did irregularities occur? Were test conditions, the testing environment, adequate and free from serious distraction? Several sponsors replied with general comments disapproving of testing children per se. This was deemed not relevant to the issue we were exploring. EDC raised questions about the qualifications and training of the testers who administered the SRI tests.

> At one of our sites parents complained that the testers were employed who were associated with another model. Many testers tell our children they are "going to play a game with them" (even though the test direction do not specifically instruct them to do so). These tests are not games. Our children expect to be treated honestly. One child, when the tester told him he "was going to play a game with him," listened for a little and then said, "If this is a game it's a dumb game, mister," and got up and left.

No sponsor provided any evidence of irregularities, although occasional claims were made that children were tested in large, noisy auditoriums, etc. From the generally disappointing returns from sponsors, we have no way of knowing whether children did understand the directions, irregularities did not occur and testing conditions were reasonable, or alternatively, whether the sponsors don't bother to learn and collect such information.

(3) Differences in norming procedure: The MAT is normed on a national sample of schools stratified geographically, by size of school location, by public or private sponsorship and by SES. Adjustments were made to assure that the sample was representative of the national population on mental ability test scores.

For the WRAT, such information on the norming procedure is scanty. The manual states that "no attempt was made to obtain a representative national sampling. Nor is such a sampling considered essential for proper standardization." While this is true as far as it goes, it is incumbent upon the test publisher to provide information about the characteristics of the norming sample so that a user can determine the comparability of his sample to the norming sample. The WRAT publisher does provide information on the age and sex of the norming sample. It claims that IQ information was used to develop norms corresponding to the "achievement of mentally average groups with representative dispersions of scores..." because of the incomplete information furnished in the WRAT manual about sampling, the test has been heavily criticized (Buros, 1972).

(4) Other explanations for differences: There are two forms of the WRAT, of which only one is appropriate to the grade level of FT children. By the third grade, some FT children will have taken the same sponsor-administered WRAT test as many as four times. They will also have taken the SRI-administered WRAT, a modification of the publisher's test, two or three times. The possible disortion of scores owing to test familiarity is compounded by the danger of increased susceptibility to teaching the test items. On the other hand, three forms of the MAT exist at four different battery levels appropriate to the grade range of FT children. This clearly

minimizes both of the distortions which may have occurred in the WRAT.

Another possible explanation for difference in GE between the WRAT and the MAT is that not all the same children are taking both tests. We have noted previously that sites exist where the differences between the number of children tested by the sponsor and those tested by SRI is considerable. Even where the difference in N's is less dramatic, the mean grade equivalent might have been raised or lowered if there were systematic differences between tested and non-tested children.

We know, for example, that most testing by the sponsors took place in the ninth month of school during 1971-72. In some sites this meant that children were tested just one or two weeks before the end of the school year. This might have resulted in biasing the scores, if lower-achieving students' attendance drops off at the end of the school year as we suspect.

Another systematic decision that has been made at some sites is not to test children when it was felt that they would be unable to attempt a test above their ability level. We know that this happened in New York City on the MAT. There, the school system's policy is that children whose achievement is considerably below grade level should not take the MAT battery approrpiate to their actual grade placement. For example, a second grade child reading at early first grade level would not take the MAT Primary II reading subtests along with the rest of

his classmates. When class means were computed for the second grade in New York City, the reported mean score was for those children who took the Primary II reading subtest and not for all children in the classroom. The reading score, therefore, looks impressively high, but the N is very small. Having discovered this, we excluded the 2nd grade New York City reading scores from our comparison. Unfortunately, we do not know the extent to which this practice was followed, either formally or informally, in other sites either on the MAT or on the WRAT. But the validity of the class mean scores either for the MAT or the WRAT remains open to question until this information is obtained for every site.

Recommendations on Achievement Test Findings. An interesting, provocative and potentially important pattern of findings emerges if one compares the scores of the same children in highly structured models taking different reading and math achievement tests at the same time. Looking at one achievement test, the WRAT, children are well above grade level and national norms in reading past the third grade, and well ahead in math until the end of third grade when they are at the national norm. Looking at another achievement test, the MAT, children are doing far less well in the secondgrade. By the end of the third grade they are considerably below national norms. If we were to judge the success of FT on the basis of achievement tests in reading and math alone, we would be inclined

to give FT an overwhelming vote of confidence using one test (the WRAT) and to have some serious reservations using another test (the MAT). Which test are we to believe?

At this time, the answer has to be "neither." We have discussed and analyzed a series of possible reasons for the one-year overall GE discrepancy in reading and the 1/2 year overall GE discrepancy in math. Any one of them could be used to explain away the difference. In some cases, there are clear indications that the MAT and its procedure for administration is considerably underestimating real effects of the sponsors. There are equally clear indications in other cases that the WRAT might be overestimating effects.

Some possible explanations could be tested if enough information were available from the sponsors, from local sites and from SRI. For other explanations, no satisfactory resolution seems attainable. It is difficult to imagine resolving the debate between those who argue that the "true" meaning of reading skill in the early primary grades is properly tested by the complex items of the MAT subtests and those who testify that reading skill in the early primary grades is more basically a matter of decoding as tested by the WRAT. For yet other explanations, knowledge from applied research that has not yet been done on the nature, interpretation and behavior of achievement tests is absolutely necessary before we can decide which test is more valid and which means of analysis are more appropriate in measuring program effects.

Perhaps the most important lesson of this study is that National Evaluation data are just one of a number of sources of fallible information about the effect of FT on the academic achievement in basic skills of primary school children enrolled in the Program. While achievement test data in the National Evaluation ought to be collected with meticulous care and analyzed using the most sophisticated techniques available, they will never be able to provide at the present the kind of unambiguous indication of the effect of FT on skill achievement that is desirable.

One implication is that FT Research ought to be involved either directly or indirectly in the kind of applied research in testing and methodology that will lead to less ambiguous interpretations of data.

A second implication is that FT Research ought to spur efforts throughout the FT Program at collecting better and more complete achievement test data both from sponsors and from LEA's. This involves not only providing funds for sponsor and LEA data collection but more important, providing technical assistance and uniform guidelines and standards which would enable FT and outside research personnel to use information from sponsors, LEA's and the National Evaluation in order to arrive at better and more accurate estimates of the effects of FT on children's achievement.

A third and related implication is that until data collection procedures and testing guidelines are created, the collec-

tion of massive amounts of test information from cities or other local education agencies at FT sites will be of little use. It will result in a deluge of new numbers and create additional problems of analysis and interpretation which will never be resolved. The first priority ought to be the gathering of better information from sponsors. Once this information is collected and analyzed and bugs ironed out, it will then become feasible to embark on the more ambitious undertaking of confirming these findings using additional sources of data from localities.

PART II

THE REPRESENTATIVENESS OF THE FT SAMPLE

An exploration of the early history of Follow Through clearly indicates that by 1968, it had been decided that FT was to be an experimental program. It was designed to produce useful information for the time when the Program could be expanded nationwide as a service program for disadvantaged children, their families and the schools that served them. A primary purpose of FT since then has been to compile evidence to help guide decisions regarding the design and implementation of compensatory education. For this reason, issues concerning the generalizability of findings from the FT population to the larger target population of poor children are of crucial concern for policy making.

Our investigation of the question is divided in three sections. Each section attempts to address the question: How representative is a sample of children (from whom we have data about FT effects) of a larger population? The variables on which representativeness is assessed, as well as the samples, vary from section to section.

The first section looks at a sample from the entire population of FT children whose parents were interviewed. NORC interviewed parents of entering children (kindergarten or first grade) each year from the Spring of 1970. The section compares

selected background information about the child and his family with similar background information about children and their families in Title I schools.*

The second section makes a different FT-Title I comparison. Here we look at the FT and Title I populations in two large cities, New York and Baltimore. The comparison variable is achievement scores expressed in G.E. in the upper grades. Achievement scores in the upper grades are used as a proxy variable for the host of child, family and school factors which influence school achievement. The logic of using this variable is: If we look at achievement scores in grades where the FT program has not yet reached, we have a relatively clean measure (barring massive year-to-year SES mobility) of how similar children were before the advent of the FT intervention. If FT schools and non-FT Title I schools show similar school achievement profiles in upper grades, it is highly likely that the similarity will extend to the lower grades where the FT program has begun. Note, however, that even if similarity is found, we can only generalize to these two cities and possibly to other large cities like New York and Baltimore. The other major drawback of this comparison is that if not all children in the lower grades are in the FT program, and if there is a selective process for picking children to receive the FT program, comparability

---

* This section is not included in the present report since we have not yet obtained printouts from Abt on the parent interviews. As soon as Abt provides us with this information, we will be able to make the comparison with Title I and will forward that section of the report. We expect that Abt will furnish us with these data in the next two weeks.

is brought into question. On the evidence we have available, it is unusual for the FT program not to cover the entire grade of an elementary school, and hence this problem would not arise.

The third section, a comparison of pupils within the National Evaluation sample, addresses a related but distinct issue of comparability. We tend to assume that data produced on tested FT children are representative of the population of all children in the FT program. If, however, tested FT children are considerably different in background characteristics from rostered but untested children, we would have to limit any conclusions from the National Evaluation to the group of tested or potentially testable FT children in the program. The study we did here should be considered exploratory. Only one background characteristic, race, was looked at. This is the only possible and meaningful comparison that could be made given the existing data tape. This comparison raises the disturbing possibility that rostered and tested FT children do differ on race in several sites. Because the sample is small, the conclusion arrived at is necessarily tentative. But it points decisively to the need for further study about the representativeness of test FT pupils for the whole FT population.

2. COMPARISON OF FOLLOW THROUGH SCHOOLS AND OTHER "DISADVANTAGED" SCHOOLS

The question is: Are FT schools representative of the wider population of disadvantaged schools? Follow Through is a compensatory program aimed at "disadvantaged children in the primary grades of schools throughout the nation." It is reasonable to ask whether FT schools are reaching this special population, or rather, some subset of that population. We know enough about the history of Follow Through to expect that the practices followed in selecting FT schools varied widely from one place to another.

Elmore (1972) has commented:

> The process used to nominate and select Follow Through sites was neither an arbitrary and irrational construct of some bureaucratic imagination nor a willful and perverse attempt to undermine good experimental design. It was founded on a very rational desire to minimize administrative difficulties.

Despite this, we know that the OEO Poverty Index was used to select FT schools, first to identify disadvantaged pupils and then, through aggregation of these data, to identify disadvantaged schools.* One component of the Poverty Index is a measure of parental income, and an income level is also used to define schools eligible for Title I funds.** So there are good a priori

---

* It should be pointed out that about one-third the FT pupils do not in fact fall within the limits defined by the OEO Index (SRI Longitudinal Evaluation of Selected Features of the National Follow Through Program, March 1971).

** AFDC eligibility is also used in conjunction with income level.

reasons for using Title I schools as the comparison group representing the wider population of disadvantaged children. The investigation reported here asks how far FT schools are representative of the Title I population.

Since the analysis depends on existing data sources rather than purpose-gathered data, it was natural that a very limited range of comparison criteria could be found. In the end only one has been used: the school mean reading test score. This requires some justification, not just because it is an imperfect proxy for many other important background which it would be interesting to take into account.* The problem is simply that test score differences between FT and Title I schools might be explained in terms of the effects of the FT program. This difficulty can be avoided in large measure if it is accepted that the test scores of pupils who could not have experienced Follow Through are an adequate means of characterizing the populations of the schools. Thus, the data presented here will refer to pupils in the FT and Title I schools who were too old to have been involved with Follow Through. The assumptions is that the test scores of these pupils in higher grades reflect important characterisitics of the populations of these schools. Further, it is assumed that the populations of these schools are

---

* It might be pointed out that school level variations in tested achievement are closely associated with school level variations in social background variables such as the traditional measures of socio-economic status.

sufficiently stable, at least over the short term of two or three years, to make use of test scores in this way. In fact we know that for these data, the grade mean scores at one grade level correlate very highly with the grade means at another level. Thus, for the New York schools, the correlation between school mean reading scores in second grade and fourth grade is 0.862 for a cohort of pupils (Acland, 1972).

A further limitation concerns the unit of analysis: the school. We know that variations in tested achievement among pupils within the same school are nearly as large as variations among all pupils. That is to say, within school variations are typically 60%-80% of the variation among the whole population of pupils. Now, we know that Title I funds are meant to be allocated to particular pupils within the schools. Rather than use Title I funds for general improvements to the schools, they are meant to be used for the most needy pupils. If this practice were followed in fact, the correct comparison would be between disadvantaged pupils in the FT program and those pupils <u>within</u> Title I schools who should be receiving Title I benefits. We are dealing here with school average scores which do not tell us about special sub-groups within the school.

On the positive side, it may be pointed out that Title I funds may, in reality, be distributed in a great variety of ways, some of which diverge from the guidelines concerning allocation.

Murphy (1973), for example, has pointed out:

> Currently it is not even clear to what extent Title I is expended on eligible disadvantaged children in poverty neighborhoods. Even when it reaches them, it is uncertain that the money buys services in addition to the level provided other school children in each district.

It was necessary to limit the study to large cities which had more than one of two FT schools. Constraints imposed by using existing data sources further limited the investigation. In the event two large cities were studied, Baltimore and New York. Achievement test data were collected for all schools in these cities. The comparison of school mean scores for FT schools and Title I schools is presented in Tables IV through VI. In all these Tables, school means are presented by grade level for most of the elementary grades.

The Baltimore data (Table IV) suggest that the FT schools have considerably lower reading scores than the whole population of elementary schools, and the same scores as Title I schools. In the third grade, for example, FT schools are seen to score, on the average, around three months below the city-wide average.

It may be added that the Baltimore city average is appreciably lower than the national norm for large cities (Baltimore Schools, 1971). These data, then, indicate that Follow Through really does reach the target population, at least in terms of the assumptions which have been defined here.

A less consistent finding emerges from the New York City data (Tables V and VI). Two boroughs have been chosen, which had the largest number of Follow Through schools, Manhattan and

TABLE IV

BALTIMORE SCHOOLS -- GRADE EQUIVALENT SCORES FOR SCHOOLS ON IOWA TEST OF BASIC SKILLS (Average of Vocabulary, Reading, Language Skills, Work-Study Skills and Arithmetic Skills Subtests). BY GRADE AND BY YEAR AND BY "WHETHER FT SCHOOL OR NOT"

1969	Grade:	3	4	5	6	N
FT Schools	Mean	2.473	3.215	4.085	5.167	13
	S.D.	0.195	0.248	0.294	0.284	
Title I Schools	Mean	2.504	3.277	4.135	5.187	74
	S.D.	0.173	0.252	0.271	0.338	
All Schools	Mean	2.791	3.561	4.510	5.601	155
	S.D.	0.510	0.568	0.633	0.686	

1970	Grade:	2	3	4	5	N
FT Schools	Mean	2.669	3.362	4.208	5.409	13
	S.D.	0.330	0.340	0.309	0.378	
Title I Schools	Mean	2.654	3.311	4.268	5.321	74
	S.D.	0.311	0.274	0.345	0.476	
All Schools	Mean	2.926	3.680	4.664	5.780	155
	S.D.	0.522	0.598	0.639	0.733	

Brooklyn. In the first, nearly all the schools were defined as eligible for Title I funds, so in this case a comparison has been made between FT schools and all the schools in the borough (Table V).

It is evident that FT schools in Manhattan tend to score below the average for the borough. There are variations from grade to grade; for example, in grade 3 (1970) FT schools score roughly three months ahead of the other schools. But in general, the Follow Through schools are about a month or so behind. From this it seems safe to conclude that FT schools are not atypical of Manhattan schools; at least they are not clearly superior to the average.

The results for Brooklyn schools (Table VI) are suspect because they are based on a very small number of FT schools. The differences in school averages for FT and Title I schools lie in no consistent direction here, but, as in the case of Manhattan schools, it appears that FT schools are roughly comparable to Title I schools. Certainly, FT schools are neither clearly superior or inferior to other "disadvantaged" schools.

Perhaps the greatest weakness of this analysis is that we lack a consensus about the appropriate comparison group. It would, after all, be surprising if there was agreement about the target population of compensatory programs. Allowing that, the case can be made that the Title I population approximates to this comparison group, and if this assumption is granted,

TABLE V

ELEMENTARY SCHOOLS IN MANHATTAN. GRADE EQUIVALENT SCORES ON THE METROPOLITAN ACHIEVEMENT TEST (Average Score from Word Knowledge and Reading Subtests) BY GRADE, AND YEAR, AND BY "WHETHER FOLLOW THROUGH SCHOOL OR NOT"

(Note: All but 10% of the schools in Manhattan are eligible for Title I funds, so the comparison presented here is between FT and all the schools in Manhattan.)

1970	Grade:	2	3	4	5	6	N
FT Schools	Mean	2.715	3.791	4.125	4.916	5.725	8
	S.D.	0.345	0.391	0.335	0.468	0.179	
All Schools	Mean	2.806	3.481	4.372	5.155	5.934	87
	S.D.	0.558	0.700	0.807	1.058	1.101	

1971	Grade:	2*	3	4	5	6	N
Ft Schools	Mean	2.757	3.238	4.066	4.812	5.375	8
	S.D.	0.382	0.573	0.543	0.868	0.432	
All Schools	Mean	2.675	3.355	4.012	4.837	5.865	87
	S.D.	0.376	0.747	0.956	1.105	1.041	

* Pupils in second grade in 1971 could have had FT experience.

TABLE VI

BROOKLYN SCHOOLS. GRADE EQUIVALENT SCORES ON THE METROPOLITAN ACHIEVEMENT TEST (Average Score Based on Reading and Word Knowledge Subtests) BY GRADE AND YEAR AND WHETHER FOLLOW THROUGH SCHOOL OR TITLE I SCHOOL.

1970	Grade:	2	3	4	5	6	N based on
FT Schools	Mean	2.657	3.590	4.127	5.283	5.800	4
	S.D.	0.268	0.302	0.181	-.359	0.002	
Title I Schools	Mean	2.555	3.229	3.946	4.716	5.431	143
	S.D.	0.314	0.382	0.482	0.642	0.750	

1971	Grade:	2	3	4	5	6	N based on
FT Schools	Mean	2.462*	3.107	3.872	4.700	5.500	4
	S.D.	0.127	0.168	0.411	0.660	0.0**	
Title I Schools	Mean	2.555	3.073	3.750	4.488	5.545	143
	S.D.	0.281	0.423	0.582	0.713	0.767	

* Pupils in second grade could have had FT experience.
** One school in this cell.

these analyses demonstrate that the FT sample is not seriously unrepresentative of the disadvantaged group. Certainly, the investigation has been restricted to two local areas, and it may be that evidence of unrepresentativeness could be found with a wider-reaching analysis. However, the indications of this analysis are that the FT sample reaches it target population.

3. COMPARISON OF PUPILS WITHIN THE NATIONAL EVALUATION SAMPLE

A further way of looking at the question of the representativeness of the Follow Through pupils makes use of the existing data gathered for the national evaluation. Again, our knowledge of procedures used to identify FT and NFT schools is sufficient to warn against expecting too much. It has already been mentioned, for example, that an unexpectedly small proportion of pupils in FT schools fall below the OEO Poverty line. Similarly, the selection of NFT schools leaves room for doubting their utility as comparison groups. Analysis of the national evaluation data bears this out. Stanford Research Institute found that little more than 40% of the FT/NFT matches were "good" in terms of their definition.*

---

* Seven baseline variables were used to estimate the quality of the match between FT and NFT samples. "For each project, the number of these variables showing a FT/NFT difference of 10 percentage points or more was tabulated. Three or less discrepancies of 10 percent or more results in the classification of an FT/NFT comparison as a 'good' match." [p. 278, SRI Interim evaluation of the national Follow Through Program 1969-1971, February 1973]

The strategy adopted here was to compare the characteristics of two groups of pupils, both falling within the FT sample, one being tested during the year, the other being excluded from the testing program. These comparisons were replicated for each entering cohort, that is, the entering K and entering first grades in the three years 1969-70, 1970-71 and 1971-72. There are two issues which these comparisons address. First it will be asked if the tested pupils, who form the basis for the key analyses of the national evaluation, are representative of the larger sample covered by the Follow Through program. Second, it will be asked if there are substantial variations in the background characteristics of the tested pupils from one year to the next.

We had to rely on pupil background information contained in the SRI Index Tape, information which was limited in scope. In the event we decided to use an index of the racial composition of the tested and untested groups; the proportion of blacks. For each site, the proportion of black pupils was computed for FT-tested and FT-untested groups. The results are presented in Table VII. Admittedly, this variable does not capture many aspects of background differences. But, on the positive side, racial background has been regarded, traditionally, as one of the key baseline variables in most evaluation analyses. Those who have sought to evaluate the effectiveness of programs such as Follow Through have been sensitive, above all, to the possibilities of variations in racial background as a causative

or confounding factor. The comparisons between FT-tested and FT-nontested are presented in Table VII. The FT-tested group are those who received some kind of achievement test during the year in question, either during the fall or during the spring.

Two observations can be made from these findings. The first is that there is a high degree of stability in the racial composition of successive FT-tested cohorts. Take, for example, site 03.09. The proportion of blacks in the FT-tested sample changes from 56.7% in Cohort I to 56.3% in Cohort II to 49.6% in Cohort III. Similarly high consistency can be found for other sites, with one exception (01.04).

The second observations is limited to a rather small numb er of sites in which we have both FT-tested and FT-non-tested pupils. For this small number of cases we find that the two groups are generally similar in terms of racial composition, but that there are dramatic exceptions to this rule. For example, site 01.14 has 66.1% blacks in the FT-tested sample in 1969-71 compared to 29.1% black in the FT-nontested sample. Similar differences can be found in other sites (e.g., 03.07, 1969-70). On the other hand, there are also sites in which the tested and untested pupils have very similar racial composition. For example, in site 05.10 the FT tested pupils were 85% black compared to the FT-untested group which was 83% black. The same holds true for the next year.

Not surprisingly, perhaps, we cannot reach firm conclusions from these analyses, but one cautious implications might be

TABLE VII

COMPARISON OF FT-TESTED AND FT-NON-TESTED PUPILS, BY YEAR, BY SITE. PERCENTAGE BLACK FOR ENTERING GRADE (EITHER K OR 1ST)

	1969-70		1970-71		1971-72	
SITE #	Tested	Non-Tested	Tested	Non-Tested	Tested	Non-Tested
0104	25.0	-	90.4	-	-	79.5
0114	66.1	29.1	68.1	56.9	-	42.2
0201	53.4	-	48.1	-	-	51.9
0204	2.7	8.5	3.0	-	3.6	-
0302	98.5	87.7	98.9	-	-	98.8
0307	69.9	29.1	58.2	49.2	60.4	-
0308	25.0	-	20.3	-	20.5	-
0309	56.7	-	56.3	-	49.6	-
0510	85.1	83.2	91.3	99.1	95.2	-
0506	98.6	-	97.6	92.9	98.0	-
0604	5.8	-	4.4	-	10.7	-
0701	89.0	-	91.9	-	92.0	-
0711	69.2	-	82.0	63.3	74.8	-
0712	-	-	0.5	-	0.6	-
0801	46.3	-	57.7	-	38.3	-
0804	24.1	-	36.4	-	33.3	-
0901	75.7	100.0	86.1	100.0	-	96.2
0902	66.7	-	73.0	-	76.0	-
1002	12.3	-	23.3	-	12.7	25.0
1102	26.8	-	28.3	-	-	27.5
1301	81.4	-	97.2	80.1	90.2	-

suggested. The results raise the possibility that the sample of pupils selected for testing may be unrepresentative of the whole FT sample. It is likely that the degree of representativeness varies from site to site. Naturally, it is not possible to say if this happens. Just what determines this cannot be discovered with these data, but the process of sample selection may well be biased inadvertently. Whatever the cause, the consequence is that one should exercise extreme caution in making any assumptions about the referent population of the Follow Through sample.

References

Acland, Henry. The effects of schooling in the elementary grades. Cambridge, Massachusetts: Center for Educational Policy Research, Harvard University, Reprint Series, 1972.

Baltimore Public Schools. Untitled report (mimeo). 1971.

Buros, O. K. (Ed.) The Seventh Mental Measurements Yearbook. Highland Park, N.J.: The Gryphon Press, 1972.

Elmore, R. F. The politics and administration of an educational experiment: The case of Follow Through. Cambridge, Mass.: Harvard Graduate School of Education, 1972.

Murphy, J. T. The education bureaucracies implement novel policy: The politics of Title I of ESEA, 1965-72. 1973.

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