

Follow-up of Follow Through: The Later Effects of the Direct Instruction Model on Children in Fifth and Sixth Grades

The later effects of the Direct Instruction Follow Through program were assessed at five diverse sites. Low-income fifth and sixth graders who had completed the full three years of this first- through third-grade program were tested on the Metropolitan Achievement Test (Intermediate level) and the Wide Range Achievement Test (WRAT). Results were contrasted with those of children in local comparison groups using analysis of covariance procedures. Results indicated consistently strong, significant effects in WRAT reading scores (decoding), consistent effects in math problem solving and spelling, and moderate effects in most other academic domains. Students appeared to retain the knowledge and problem-solving skills they had mastered in the primary grades. However, without a continuing program, most students demonstrated losses when compared to the standardization sample of the achievement tests. Implications for improved instruction in the intermediate grades were discussed.

Project Follow Through evaluated a variety of educational approaches to teaching low-income children in various communities from

kindergarten (or first grade) through third grade. The originators of Follow Through believed that gains made by students in Head Start could be enhanced and solidified in a comprehensive, systematic three- or four-year program. One of the approaches found to be most effective in the longitudinal evaluation conducted by Abt Associates and Stanford Research Institute under the auspices of the U.S. Office of Education (USOE) was the Direct Instruction Model (Bereiter & Kurland, 1981-82; Kennedy, 1978; Stebbins *et al.*, 1977).

The present study investigated the later effects of Direct Instruction Follow Through; that is, what happened to fifth- and sixth-grade graduates of the Direct Instruction Follow Through program. These children were tested on all subtests of the Intermediate Form of the Metropolitan Achievement Test (Durost, Bixler, Wrightstone, Prescott, & Balow, 1970) and the Reading subtest of the Wide Range Achievement Test (Jastak & Jastak, 1965). Results were compared to children in local comparison groups with similar demographic characteristics using analysis of covariance (ANCOVA) with multiple covariates. In addi-

This research was supported in part by U.S. Office of Education grant no. G007507234.

Robert St. Pierre and Robert Goodrich of Abt Associates offered some valuable feedback on an earlier version of this paper. Paul Williams of Dallas Independent Schools, Bill White of the University of Oregon, and Peter Sharpe of Rustin College (Melbourne) assisted in the data analyses. Robert and Jane Donahue organized computer processing of the data. Harriet Kandelman, Doug Carmine, Bill White, and Robert Taylor of the University of Oregon gave careful, thoughtful readings to earlier drafts of this manuscript.

Reprinted from *American Educational Research Journal*, Spring 1982, Vol. 19, No. 1, Pp. 75-92

tion, the longitudinal progress of the samples for the three years of the program and the three years after the program was compared to the norm samples of the Metropolitan and Wide Range Achievement Tests.

Background

The Direct Instruction Model represents a highly structured approach to early-childhood education with an emphasis on high levels of academic engaged time through small-group instruction in reading, oral language, and arithmetic. The Distar curriculum materials used in this approach are designed to explicitly teach general principles and problem-solving strategies. Teachers and paraprofessional aides are trained to teach these programs in a fast-paced, dynamic fashion with high frequencies of unison group responses and systematic corrections of student errors.

In some communities, Follow Through began in kindergarten and lasted four years; in others, Follow Through began in first grade and lasted three years. For a variety of logistical reasons (primarily having to do with difficulties in obtaining cooperation in larger northern cities) this follow-up study deals only with the three-year programs. (See Becker & Englemann [1978] for further details.)

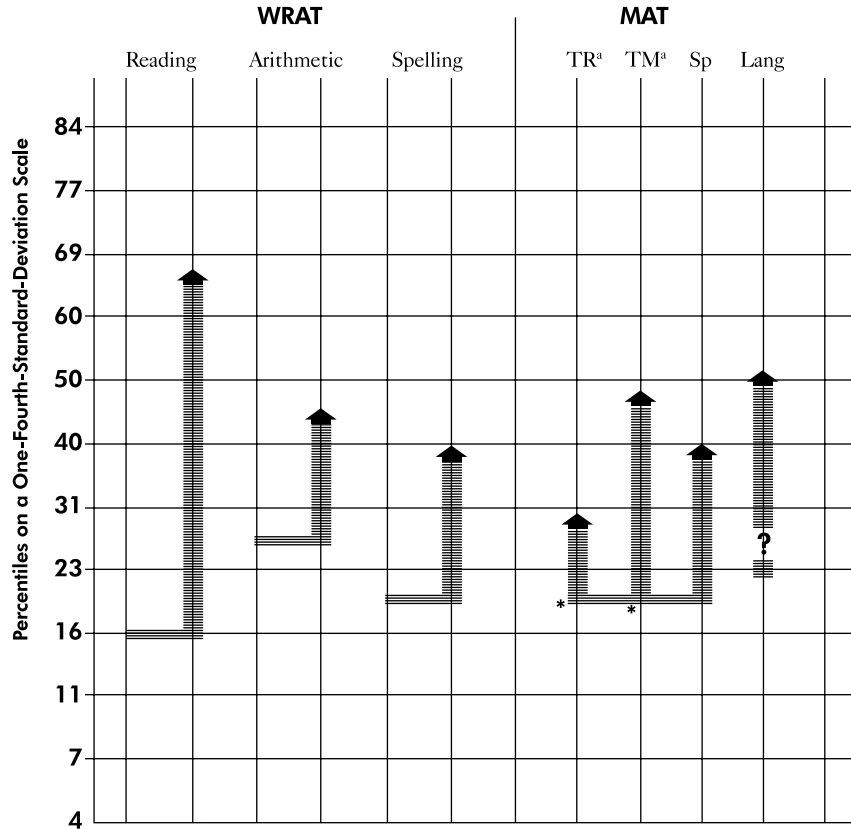
Results of the national evaluation (Stebbins *et al.*, 1977) indicated a high proportion of significant positive effects for both three- and four-year Direct Instruction sites. When third-grade students completing the Direct Instruction Follow Through program in the three-year sites were compared to a pooled national comparison group, they performed significantly higher in 60 percent of the instances for total reading, 80 percent for total math, 100 percent in language, and 50 percent in spelling.

The absolute level of performance on standardized achievement tests was typically higher for students in the four-year programs than

the three-year programs, particularly in reading. Because of this fact, the results reported here represent a low estimate of what might have been achieved. Figure I presents mean scores on all subtests of the Wide Range Achievement Test (WRAT) Level I, and 1970 Metropolitan Achievement Test (MAT), Elementary Level, for all low-income students in the three-year program. These data were collected for four groups of children in eight communities. The WRAT scores, on the left, are presented in the form of a norm-referenced comparison (Horst, Tallmadge, & Wood, 1975; Tallmadge, 1977). The mean standard scores at the beginning of the program (entry into first grade) and end of the program (end of third grade) are converted to percentiles in order to assess growth against the norm sample of the WRAT. The right half of Figure I presents MAT scores. Again, the mean standard scores at the end of third grade are converted to a percentile. No pretest level of the MAT was available. The performance of Direct Instruction Follow Through students is contrasted with typical third-grade performance of low-income, minority students in math and reading, according to USOE reports (Ozenne *et al.*, 1974; 1976). Note that MAT math and language scores are within a few percentile points of the national median, and all scores are significantly above the typical level of low-income, minority students. The large discrepancy between decoding skills (word reading as assessed by the WRAT) and reading comprehension scores (as assessed by MAT reading) is probably due primarily to low-income children's problems with the large, virtually uncontrolled reading vocabulary required on third-grade achievement tests (which reflects the content of fourth-grade textbooks) (Becker, 1977). Schools expect vocabulary development to occur at home, yet Becker makes a clear case that this is not happening in many low-income homes. Note that no corresponding discrepancy appears in Math, where the MAT tests computation, high-order problem solving, conceptual skills, and the WRAT tests only computation.

Figure I

Norm-references gains on the WRAT and end-of-third-grade scores on MAT low-income, direct instruction Follow Through students (3-year program).



	WRAT		WRAT		WRAT		MAT			
	Reading	Arithmetic	Reading	Arithmetic	Spelling	Spelling	TR ^a	TM ^a	Sp	Lang
PRE	1	1	1	1	1	1	1	1	1	1
POST	3	3	3	3	3	3	3	3	3	3
N	2418	4100	2451	4112	2428	4068	2392	2385	2410	2402
%-TILE	16	67	26	45	20	40	30	48	40	51
S. S.	84.9	106.3	90.5	98.0	87.4	96.4	55	70.0	61.0	70.7
S. D.	14.4	19.0	13.7	9.7	15.1	15.4	9.0	11.0	11.0	12.3
G. E.	.51	4.6	.77	3.8	.64	3.6	2.9	3.7	3.3	4.1

^a TR = Total Reading; TM = Total Math.

* Typical end-of-third-grade performance of low-income minority students (Ozenne *et al.*, 1976).

It seemed important to examine the later effects of the Direct Instruction Follow Through program to see if the students maintained and built on the gains they made during the first three years of elementary school, and to determine in which academic domains these gains were maintained. To do so we compared Follow Through graduates with children in local comparison groups. It also seemed worthwhile to trace the longitudinal progress of the Follow Through children through their entire six years in elementary school, and contrast their scores with the standardization sample of the achievement tests.

Method

A quasi-experimental design (Cook & Campbell, 1979) was used. In each of five Direct Instruction Follow Through sites, roughly equivalent comparison groups were located by the school district. Demographic information was collected on income level, sex, primary home language, and mother's education. These variables were used as potential covariates in the analyses. Despite the problems in using analysis of covariance (Campbell & Boruch, 1975; Elashoff, 1969; House, Glass, McLean, & Walker, 1978), it was the only feasible option for a follow-up study. As Cook and Campbell (1979) state, when one is using a quasi-experimental design and imperfect data analysis techniques, it is crucial to replicate. In this case, replications were conducted (a) across five communities, (b) across two grade levels, and (c) across two cohorts of children.

Subjects, Sites, and Testing

In the spring of 1975, fifth- and sixth-grade students who had previously completed the Follow Through program were tested on all subtests of the MAT (Durost *et al.*, 1970), Intermediate Level, and Levels I and II of the reading subtest of the WRAT (Jastak & Jastak, 1965). The same battery was given to children

in the local comparison group who had undergone traditional education in the community.

The only selective factors operating in choosing sites were the availability of a local comparison group and district cooperation. All 15 sites affiliated with the model were invited to participate provided (a) the district would allow additional testing of fifth- and sixth-graders, and (b) a comparison group of children with similar demographic characteristics could be found. Eight of the 15 sites agreed to participate. Generally, these were rural sites in the South and moderate-sized cities in the Midwest. One small, rural site was eliminated from the final analyses because the sample size was too small ($N = 6$) to produce reliable results. In another site, a Native American reservation in North Carolina, it was impossible to find a reasonable comparison group. Because five of the six remaining sites were three-year programs, analyses were limited to the three-year sites. (Results for the one four-year site are available in Becker and Englemann [1978] and Becker and Gersten [1979]. They show basically similar patterns to the sample analyzed in the large study.)

The five sites which agreed to participate were Dayton, Ohio and East St. Louis, Illinois (with urban black populations); Tupelo, Mississippi (with a rural, black population); Smithville, Tennessee (with a rural white population); and Uvalde, Texas (with a primarily Hispanic population). In 1976, the study was replicated in four of the five sites. (Dayton did not participate in the 1976 replication study.)

These five sites appear to offer a representative sample of the 15 sites affiliated with the Direct Instruction Model. The wide range of populations typical of Follow Through is represented. Two of the sites deemed most effective in the national evaluation of Follow Through (Stebbins, 1976; Stebbins *et al.*, 1977) Williamsburg Count, SC and New York, NY were not able to participate, whereas

Tupelo, MS, a site with inconsistent, often non-significant results in the national evaluation, was able to participate.

The 1975 study involved 624 Follow Through graduates and 567 non-Follow-Through students; the 1976 study included 473 Follow Through graduates and 403 non-Follow-Through children. Table I presents fifth grade sample sizes and demographic information for each site (see Becker & Englemann, 1978; Becker & Gersten, 1979 for sixth-grade demographics, which are quite similar). Low-income students were sought in each case. As it turned out, there were a few students in most samples who were not from low-income families, so income level was used as a potential covariate.

Testing was conducted by local staff after training by University of Oregon supervisors on the procedures specified in the publishers' test manuals. Actual testing was monitored by supervisors to ensure that standard procedures were followed. In Tupelo, MS, the California Achievement Test results (collected by the local district) were used in place of the MAT and the reading scores were converted to MAT equivalents using the Anchor Test Study (Loret, Seder, Bianchini & Vale, 1974).

Strategies for Data Analysis

Analysis of covariance using ethnicity, income, home language, mother's education, sex, and number of siblings was performed on each sub-test at each site. One strategy that was consid-

Table I

Comparison of Follow Through (FT) and Non-Follow Through (NFT) Groups on Selected Demographic Variables for the fifth Grade

Site	Sample Size		Mother's Ed. Scale ^a		Proportion Low Income		Proportion Non-White	
	FT	NFT	FT	NFT	FT	NFT	FT	NFT
E. St. Louis, IL								
1975	43	45	4.35	3.68	.93	.98	1.00	1.00
1976 ...	45	44	4.76	...	1.00	...	1.00	...
Smithville, TN								
1975	47	51	3.30	3.70	.87	.54	.17	.00
1976	71	38	3.43	3.40	.86	.86	.00	.05
Uvalde, TX								
1975	117	86	2.86	2.34	.87	.95	1.00	.95
1976	103	74	2.77	2.35	.98	.98	.98	.80
Dayton, OH ^b								
1975	104	87	4.75	5.00	.85	.71	1.00	.91
Tupelo, MS								
1975	46	35	3.22	3.50	.95	.97	.93	.53
1976	56	42	4.11	4.15	.87	.95	.95	.64

Note. Leaders indicate missing data.

^a 5 = High school graduates.

^b Did not participate in 1976 study.

ered, and ultimately rejected, was pooling together all Follow Through and all comparison (non-Follow Through) children in all five sites (see Goodrich & St. Pierre, 1979). Despite the immense gain in sample size and statistical power, this option was deemed inappropriate because subjects in both Follow Through and comparison groups came from at least four highly distinct populations (urban black, rural black, rural Anglo, and Hispanic), and it seemed highly unlikely that the assumptions of ANCOVA would be met. Thus, separate ANCOVAs were performed on each site for each subtest for each year.

The remaining problem was one of meaningfully synthesizing and collating the results of the multitudinous analyses. Using the site as a unit of analysis was rejected because there were too few sites to give adequate power to any test. Three analytic strategies were adopted for synthesizing the results.

In the first analysis, the results of each ANCOVA were classified as (a) significant ($p < .05$, two-tail), (b) suggestive of a trend ($.05 < p < .15$), or (c) non-significant. For each subtest, the number of sites falling into each category was tabulated. Because sample sizes tended to be small (ranging from 25 to 117, but averaging around 50), it seemed appropriate to record those sites in which $.05 < p < .15$. The researchers reasoned that, if, for a particular subtest (e.g., math problem solving), only one out of six comparisons was positive at $.05 < p < .15$, this would rightfully be considered a chance finding. If, on the other hand, eight of the ten comparisons were found to be "suggestive" at the .15 level, this would be evidence of a replicable phenomenon.

The second procedure used was the meta-analysis technique advocated by Gage (1977) (after Jones and Fiske, 1953). For each site-level ANCOVA the exact p values (both significant and nonsignificant) are converted to chi-square ratios (with two degrees of freedom).

Total chi-square values were then tested for significance with $2(n-1)$ degrees of freedom (where n = number of studies in the meta-analysis). This technique is one of the only meta-analytic techniques to offer statistical significance levels for the comparisons.

Finally, the average magnitude of effect in pooled standard deviation units for each subtest at each grade level was calculated (Glass, 1976; Pillemer & Light, 1980; Smith & Glass, 1977). (In this case, the standard deviation was computed by pooling the comparison sample. This seemed the most reasonable procedure since the larger sample size gives more stability to the estimate and the sample of F-T graduates is not a treatment group in Glass' sense.) The method gives an estimate of the treatment effect that is not biased by the differential sample sizes at the various sites.

Results

Table II, which summarizes the data from the first method of analysis, indicates the number of site level ANCOVA comparisons in both the 1975 and 1976 studies falling into each of the three categories outlined in the previous section. Table III presents the results of the Gage (1977) and Jones and Fiske (1953) meta-analysis procedures; chi-square values are shown for Grades 5 and 6, and then for the total number of comparisons. Table IV, which presents results of the third method of analysis, shows the mean magnitude of effect in pooled SD units (Glass, 1976) for each subtest.

The strongest, most consistent finding is for reading decoding, as assessed by both Level I and Level II of WRAT reading: $X^2(30) = 134.1$, $p < .005$ for Level I; $X^2(32) = 134.4$, $p < .005$ for Level II. The mean magnitudes of effect (in Table IV) range from .38 to .56 pooled standard deviation units, well over the conventional criteria set for educational significance of .25 or .33 pooled standard deviation units (Horst, Tallmadge, & Wood, 1975; Stebbins *et al.*, 1977). This test measures chil-

dren's ability to accurately read isolated words. The consistency is demonstrated across sites, grade levels, and levels of the test. The teaching of decoding (or word attack skills) is one of the strongest early outcomes of the Direct Instruction Model; mean end-of-third grade performance corresponds to the 67th percentile on the WRAT for entering-first-grade students. It appears that these skills are maintained two–three years after the program ends.

A strong, consistent effect is also found on MAT spelling (Table III), with significant effects for both grades five ($p < .05$) and six ($p < .05$) and the combined sample ($p < .005$). It is possible that the enduring effect in spelling is related to the phonic and word-attack skills the students mastered in the early grades.

The other strong, consistent effect is in math problem solving. Table II shows that in each year, half of the site level ANCOVAs significantly favor Follow Through at either the .05 or .15 level, and no comparisons significantly favor the comparison children. The chi-square analyses (Table III) are also significant at the .05 level for both Grade 5, Grade 6, and at the .005 level for the combined sample. Mean magnitude of effects (Table IV) are .27 for Grade 5 and .18 for Grade 6. Note that the math problem solving effects are consistently stronger than math computation. At first, this would seem unusual for a program with a heavy emphasis on acquisition of basic skills. Yet the finding is consistent with the emphasis in the Distar arithmetic programs on teaching general-case problem-solving strategies, including basic algebraic principles.

Table II

Significance Levels (two-tailed test) for Fifth- and Sixth-Grade Follow Through/Non-Follow Through Comparisons with Covariance Adjustment

	WRAT Reading		Metropolitan Achievement Test										Summary of Effects
	Level I	Level II	Reading			Math				Sci	Lang	Spelling	
			Word Know	Rdg	TOTAL Rdg	Comp	Concepts	Prob. Solv.	TOTAL MATH				
a) 1975 Study													
Favoring FT ($p < .05$)	7	7	3	2	2	2	3	3	3	2	2	1	37
FT (.15 > $p > .05$)	0	0	0	1	1	2	0	1	0	0	0	4	9
ns ($p > .15$)	1	1	7	7	7	4	5	4	5	6	6	3	56
Favoring NFT	0	0	0	0	0	0	0	0	0	0	0	0	0
Total	8	8	10	10	10	8	8	8	8	8	8	8	102
b) 1976 Study													
Favoring FT ($p < .05$)	5	4	0	1	1	0	2	0	0	1	3	2	19
FT (.15 > $p > .05$)	0	2	1	0	1	1	0	3	0	1	0	2	11
ns ($p > .15$)	2	2	7	7	6	1	4	3	5	4	3	2	46
Favoring NFT (.15 > $p > .05$)	0	0	0	0	0	1	0	0	1	0	0	0	2
Total	7	8	8	8	8	3	6	6	6	6	6	6	78

More variable effects are found for the word knowledge, math concepts, science, and language subtests, as well as the composite scores for total reading and total math. For example, for word knowledge and science, the chi-square analysis (Table III) indicate significant effects for the combined ($p < .01$) and the fifth grade ($p < .05$) samples, but not for the sixth grade. Yet the magnitude of effects (Table IV) is actually somewhat higher for Grade 6. Site level analyses (Becker & Englemann, 1978; Becker & Gersten, 1979) do not shed any great light on these patterns, other than indicating that the largest site, Uvalde, seemed to have consistent effects in science and word knowledge.

Overall, there is reasonable evidence of significant later effects. Of the total of 180 comparisons in Table II, 56 favor the Follow Through sample at the .05 level; none favor the comparison group. If one uses the more liberal .15 significance level to explore potential trends (Carver, 1978), 76 effects (42%) favor Follow Through, and only two (less than 1%) favor the comparison groups. The tests with the strongest effects are WRAT reading, MAT spelling, and MAT math problem solving. Meta-analysis techniques revealed a similar pattern. The mean magnitude of effect (Table IV) is well above .33 pooled SD units for all levels of WRAT Reading and in the .17 to .26 range for most MAT subtests. The two tests with consis-

Table III

Pooled Results from 1975 and 1976 Follow Up Studies in Seven Follow Through Sites Chi-Square Ratios Calculated by Jones & Fiske (1953) Meta-Analysis Procedures

Subtest	Grade 5			Grade 6			Combined		
	Chi Square	df	P	Chi-square	df	P	Chi-square	df	P
a) Metropolitan Achievement Test (Intermediate) (1970)									
Word knowledge	33.8	18	<.05	26.7	18	ns	60.5	36	<.01
Reading	30.3	18	<.05	20.3	18	ns	50.6	36	ns
Total Reading	36.2	18	<.05	22.7	18	ns	58.9	36	<.01
Language	22.7	14	ns	24.8	15	<.05	47.5	28	<.05
Spelling	29.1	14	<.05	27.2	14	<.05	56.3	28	<.005
Math computation	16.3	14	ns	25.5	14	<.05	41.8	28	<.05
Math concepts	21.1	14	ns	33.3	14	<.05	54.3	28	<.01
Math Problem	33.2	14	<.05	28.0	14	<.05	61.2	28	<.005
Total math	16.8	14	ns	24.8	14	<.05	41.6	28	<.05
Science	26.6	14	<.05	22.4	14	ns	49.0	28	<.01
b) WRAT Reading (Decoding)									
Level I	80.0	14	<.005	54.0	16	<.005	134.1	30	<.005
Level 11	73.3	16	<.005	61.1	16	<.005	134.4	32	<.005

tently low magnitudes of effect are math computation and reading (comprehension). Using the Jones and Fiske (1953) statistical tests for the combined sample, there is evidence of significant, enduring effects in all domains except MAT Reading (a test of reading comprehension). There are no consistent patterns indicating that one particular site or one particular grade level displayed more lasting effects. Thus, the effects appear to be due to the model.

To round out the picture, Table V presents the unadjusted percentiles for WRAT reading and total reading, total math, spelling, and science for each site in the 1975 study. These percentiles are converted from the unadjusted mean standard scores. Table VI presents the mean magnitude of effects for adjusted scores on a site level basis for the children who were in fifth grade in 1975 and sixth grade in 1976. Table VI shows reasonably high consistency across grade levels when the same children are followed.

Table IV
Mean Magnitude of Effects in Pooled Standard Deviation Units Between Follow Through and Non-Follow Through Samples Pooled from 1975 and 1976 Studies

Test	5th Grade	6th Grade
WRAT Level II	.50	.51
WRAT Level I	.56	.38
MAT Word Knowledge	.19	.23
MAT Reading	.16	.14
MAT Total Reading	.20	.19
MAT Math Computation	.09	.13
MAT Math Concepts	.18	.24
MAT Math Problem Solving	.27	.18
MAT Total Math	.18	.26
MAT Science	.20	.26
MAT Language	.21	.20
MAT Spelling	.24	.17

Table V
Unadjusted Percentiles for Follow Through and Non-Follow Through in 1975 Study

	MAT									
	WRAT Reading		Total Reading		Total Math		Spelling		Science	
	FT	NFT	FT	NFT	FT	NFT	FT	NFT	FT	NFT
Grade 5										
E. St. Louis	17	17	16	18	41	39	11	14
Smithville	63	39	34	34	49	56	39	39	36	36
Uvalde	45	23	16	17	24	21	37	37	22	19
Dayton	61	27	20	16	19	12	34	27	20	13
Tupelo	42	27	18	18
Grade 6										
E. St. Louis	32	18	33	21	53	43	20	12
Smithville	73	25	36	21	52	28	59	39	42	26
Uvalde	32	18	15	15	22	17	39	28	21	17
Dayton	50	27	22	22	19	17	13	25	23	21
Tupelo	44	13	36	28

* Level I for Grade 5. Level II for Grade 6.

Longitudinal Tracebacks: Gains And Losses Of Follow-Through Children Against The National Norm Sample

Table VII traces the growth of the Follow Through children against the norm sample of the WRAT for the children during the three years of the Follow Through program, followed by their decline during the intermediate grades. East St. Louis is omitted because data were unavailable. Entry (pretest) scores were unavailable from Tupelo and Dayton; however, on the basis of similar Follow Through children tested at entry in these communities in later years, one can estimate the entry scores at approximately the 14th percentile for Dayton and the 9th percentile for Tupelo (Becker & Engelmann, 1978).

Note in Table VII that the major growth in reading decoding occurs during the first two years of school, when a major program empha-

sis is on word-attack skills. The level is basically maintained in Grade 3. By Grades 5 and 6, there are appreciable drops in Smithville and Tupelo. Although the Follow Through students significantly outperform non-Follow Through students at all sites on the WRAT, they are losing a bit when compared to the norm sample.

This decline is even more dramatic on the Metropolitan Achievement Test (Table VIII). The MAT is a well-normed, comprehensive test of reading, math, and language, including both basic skills and higher order, cognitive operations (Bereiter & Kurland, 1981-82; Wolf, 1978). To conserve space, Table VIII presents only percentile equivalents for the composite scores total reading, total math, and spelling. (The mean scale scores, standard deviations, and sample sites, and detailed site level analyses are available in Becker and Engelmann [1978] and Becker and Gersten [1979].)

Table VI

Mean Magnitude of Effects in Pooled Standard Deviation Units A Within-cohort Follow-up of 1975 5th Graders as 6th Graders in 1976

Site Date	Uvalde		Smithville		E. St. Louis		Tupelo	
	1975	1976	1975	1976	1975	1976	1975	1976
N =	117	108	46	38	43	37	56	55
Test:								
WRAT reading level II	.47	.41	.61	.53	...	-.15	.78	.46
WRAT reading level I	.47	.52	.52	.48	...	-.15	.63	...
MAT word knowledge	.03	.19	.06	.02	.00	-.03	.38	.18
MAT reading	.21	.20	.25	.02	-.12	-.20	.20	.01
MAT total reading	.11	.20	.15	.02	.01	-.11	.28	.09
MAT language	.13	.39	.09	.03	.03	-.06
MAT spelling	.01	.25	.36	.37	.13	-.06
MAT math computation	.11	.04	.09	.15	.05	-.30
MAT math concepts	.10	.31	.09	-.08	.21	-.30
MAT math prob. solv.	.33	.23	.20	-.28	.02	-.22
MAT total math	.10	.16	.15	-.08	.01	-.33
MAT science	.15	.36	.23	-.04	-.15

Note: Leaders indicate missing data.

By the end of the third grade, with three years of Direct Instruction Follow Through, all sites were within a few percentile points of the national median in total math and within one-fourth standard deviation in spelling. In total reading, Smithville students are at the national median, the Dayton sample within one-fourth of a standard deviation unit, and Tupelo and Uvalde one-half standard deviation below. Yet two years after the program had ended, all samples made appreciable, significant drops against the national norm group in both math and reading. In the case of Smithville, there are even further drops during sixth grade. Though in many domains Follow Through graduates outperform the control students in Grades 5 and 6 (Tables

II-V), low income Follow Through students are losing against the national normal sample. The same phenomenon occurs for other low-income students in the intermediate grades (National Center for Educational Statistics, 1978). The losses are much smaller in MAT Spelling and WRAT Reading (decoding) than MAT Reading and Language.

Conclusions

There are two basic findings in this study. The first is that there is evidence that, in most domains assessed by standardized achievement tests, low income graduates of a three-year Direct Instruction Follow Through program

Table VII
Longitudinal Analysis of WRAT Reading at Entry and at the End of Grades 1, 2, 3, 5, 6

Site	WRAT Reading (Level 1)					
	Pre-Grade I (Fall, 1970)	Grade I (1971)	Grade 2 (1972)	Grade 3 (1973)	Grade 5 ^a (1975)	Grade 6 ^b (1976)
Uvalde, TX						
Percentile	9th	43rd	64th	55th	45th	42nd
Mean Std Score ^c	79.7	97.4	105.5	102.0	98	97
N	81	110	110	110	117	108
Smithville, TN						
Percentile	20th	66th	74th	75th	63rd	39th
Mean SS	87.6	106.0	109.4	110.1	105	96
N	40	46	43	46	46	38
Tupelo, MS						
Percentile	...	39th	57th	55th	42nd	34th
Mean SS	...	96.0	102.5	102.1	97	94
N	24	25	25	56	55
Dayton, OH						
Percentile	...	52	73	67	61	...
Mean SS	...	100.7	108.8	106.7	104	...
N	97	104	102	93	...

Note. Leaders indicate missing data.

^a Estimate based on raw score for Level I.

^b Estimate based on raw score for Level II.

^c Mean = 100, SD = 15.

perform better than comparable children in their communities who did not attend the program. These enduring effects are strongest and most consistent in WRAT reading (decoding), math problem solving and spelling. There are lesser effects in MAT science, math concepts, math computation, and word knowledge. The fact that this study was conducted at five quite diverse sites across two grade levels and replicated in 1976 adds confidence to the results. Because none of the outcomes significantly favored the comparison groups at the .05 level, and 31 percent favored Follow

Through, it is extremely unlikely that these results are due to chance. These results are from the second and third cohorts of Follow Through children, when the program was not fully developed. Also, only sites with a three-year (rather than four-year) program participated. For both these reasons, the results cited here, in all likelihood, represent a low estimate of program effectiveness. For example, a quasi-experimental follow-up study by Gersten, Gutkin, and Meyer (Note 1) demonstrated that low-income fifth graders from the four-year program in New York City were significantly outperforming comparison group children, and were well above national median levels in reading on the Comprehensive Test of Basic Skills (CTBS). Also, Weber and Fuhrmann (Note 2) reported significant later effects in reading and math on the California Achievement Test for ninth graders who had completed the program; Follow Through graduates were .24 standard deviation units (or .8 grade equivalent units) ahead of comparable students in the district.

The second finding is less optimistic. When compared to the national norm sample, these children invariably lose ground in the three years after they leave Follow Through.

Two reasonable conclusions can be formed from these findings. The first is that if students learn skills and problem solving strategies well, they do not lose this knowledge. Follow Through graduates often perform significantly higher than other low-income fifth and sixth graders in their communities, especially in the areas of reading decoding, math concepts, math problem solving, and science.

The second conclusion is that without effective instruction which continues to build on these skills in the intermediate grades, the children are likely to lose ground against their middle-income peers. They are failing to master new computational skills (such as long division and complex multiplication), and are failing to develop their vocabularies

Table VIII

Longitudinal Analysis of Metropolitan Achievement Test: Same Children Followed from Grades 3 to 6

	Grade 3 (1973)	Grade 5 (1975)	Grade 6 (1976)
a) Total Reading Percentiles ^{ab}			
Site			
Uvalde, TX ...	31st	16th	16th
Smithville, TN ...	52nd	34th	26th
Tupelo, MS ...	28th	18th ^c	17th ^c
Dayton, OH ...	40th	20th	...
b) Total Math Percentiles			
Site			
Uvalde, TX ...	53rd	24th	19th
Smithville, TN ...	78th	49th	36th
Dayton, OH ...	55th	19th	...
c) Spelling Percentiles ^a			
Site			
Uvalde, TX ...	40th	37th	35th
Smithville, TN ...	62nd	39th	39th
Dayton, OH ...	40th	34th	...

Note. Leaders indicate unavailable data.

^a Percentiles converted from mean standard score for the sample.

^b Third grade tested on Elementary Form. Grades 5 and 6 on Intermediate Form.

^c CAT scores converted by Anchor Study tables.

and reading comprehension abilities at the rate of middle- and higher-income students. Limited English speaking students appear to lose the most. In order for these children to become fully literate adults, it appears that they need high-quality instructional programs in the intermediate grades (and probably beyond). Key areas for program development are instruction in reading comprehension (Jenkins, Stein, & Osborn, 1981; Perfetti & Lesgold, 1977; Resnick, 1981); vocabulary development (Becker, 1977); independent study skills (Adams, 1980; Chall, 1979; Durkin, 1978–1979); mathematical problem solving (Silbert, Carnine & Stein, 1981); expressive writing (Frederiksen, Whiteman & Dominic, 1981); and independent reading for information and pleasure (Brown & Smiley, 1977).

An ever-increasing body of knowledge has accrued from correlational and experimental studies of effective classroom practices in the elementary grades (Brophy & Evertson, 1976; Fischer *et al.*, 1980; Gersten, Carnine & Williams, in press; Good & Grouws, 1979; Stevens & Rosenshine, 1981). These studies isolate teaching practices that are consistently effective, and that have always been central to the Direct Instruction Follow Through Model, such as high student success rate, clarity of tasks, amount of guided practice, and systematic use of correction procedures. There is now a need to implement and evaluate instructional programs in the intermediate grades that systematically utilize principles of Direct Instruction, which include mastery learning, high levels of feedback, and incremental steps to develop independent reading, writing, and critical thinking.

Reference Notes

1. Gersten, R., Gutkin, J., & Meyer, L. *Evaluation of P.S. 137* (New York City) *Follow Through*. Final report submitted to Joint Dissemination Review

Panel, Department of Education, Washington, D.C., 1981.

2. WEBER, B., & FUHRMANN, M. *A study of District A's former Follow Through students' retention of basic skills after six years out of the program*. Paper prepared for District A, April 1978.

References

- ADAMS, A., CARNINE, D., & GERSTEN, R. Instructional strategies for studying Content area texts in the intermediate grades. *Reading Research Quarterly* in press.
- BECKER, W. C. Teaching reading and language to the disadvantaged — What we have learned from field research. *Harvard Educational Review* 1977, 47, 518–543.
- BECKER, W. C., & ENGELMANN, S. *Analysis of achievement data on six cohorts of low income children from 20 school districts in the University of Oregon Direct Instruction Follow Through Model* (Follow Through Project, Technical Report #78–1). Eugene, Ore.: University of Oregon, 1978.
- BECKER, W. C., & GERSTEN, R. M. *Follow-up study of fifth and sixth graders: The 1976 replication study*. (Follow Through Project, Technical Report #78–1). Eugene, Ore.: University of Oregon, 1979.
- BEREITER, C., & KURLAND, M. A constructive look at Follow Through results *Interchange* 1981–82, 12, 1–22.
- BROPHY, J., & EVERTSON, C. *Learning from teaching: A developmental perspective* Boston: Allyn & Bacon, 1976.
- BROWN, A. L., & SMILEY, S. S. Rating the importance of structural units of prose passages: A problem of meta-cognitive development. *Child Development* 1977, 48, 1–8.
- CAMPBELL, D. T., & BORUCH, R. F. Making the case of randomized assignments to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In C. A. Bennett & A. A. Lumsdaine (Eds.), *Evaluation and experiment: Some critical issues in assessing social programs*. New York: Academic Press, 1975.
- CARVER, R. P. The case against statistical significance testing. *Harvard Educational Review* 1978, 48, 378–399.
- CHALL, J. S. The great debate: Ten years later, with a modest proposal for reading stages. In L. B. Resnick & P. A. Weaver (Eds.), *Theory and practice of early reading* (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum, 1979.

- COOK, T., & CAMPBELL, D. T. Quasi-experimentation: *Design and analysis issues for field settings*. Chicago: Rand McNally, 1979.
- DURKIN, D. What classroom observations reveal about reading comprehension instruction. *Reading Research Quarterly* 1978-1979, 15, 481-533.
- DUROST, W. N., BIXLER, H. H., WRIGHTSTONE, J. W., PRESCOTT, G. A. & BALOW, I. H. *Metropolitan Achievement Tests*. New York: Harcourt Brace Jovanovich. 1970.
- ELASHOFF, J. D. Analysis of covariance: A delicate instrument. *American Educational Research Journal* 1969, 6, 383-401.
- FISCHER, C. W., BERLINER, D. C., FILBY, N. N., MARLIAVE, R., CAHEN, L. S., & DISHAW, M. M. Teaching behaviors, academic learning time and student achievement: An overview. In C. Denham & A. Lieberman (Eds.), *Time to learn*. Washington, D.C.: USOE/NIE Printing, 1980.
- FREDERIKSEN, C. H., WHITEMAN, M., & DOMINIC, J. (Eds.). *Writing: The nature Development, and teaching of written communication*. Hillsdale, N.J.: Lawrence Erlbaum, 1981.
- GAGE, N. *The scientific basis of the art of teaching*. New York: Columbia Teachers College Press, 1977.
- GERSTEN, R., CARNINE, D., & WILLIAMS, P. Measuring implementation of a structured educational model in an urban setting: An observational approach. *Educational Evaluation and Policy Analysis*, in press.
- GLASS, G. V. Primary, secondary, and meta-analysis of research. *Educational Researcher*. 1976, 5, 3-8.
- GOOD, T. L., & GROUWS, D. A. The Missouri Mathematics Effectiveness Project: An experimental study in fourth-grade classrooms. *Journal of Educational Psychology* 1979, 71, 355-362.
- GOODRICH, R., & ST. PIERRE, R. *Opportunities for studying later effects of Follow Through*. Cambridge, Mass.: Abt Associates, 1979.
- HORST, D. P., TALLMADGE, G. K., & WOOD, C. T. *A practical guide to measuring project impact on student achievement*. Monograph No. I on Evaluation in Educational. Washington, D.C.: U.S. Government Printing Office, 1975.
- HOUSE, E. R., GLASS, G. V., McLEAN, L. D., & WALKER, D. E. No simple answer: Critique of the "Follow Through" evaluation. *Harvard Educational Review* 1978, 48, 128-160.
- JASTAK, J., & JASTAK, S. *Wide Range Achievement Test*. Wilmington, Del.: Jastak Associates, 1965.
- JENKINS, J. R., STEIN, M., & OSBORN, J. What next after decoding: A look into instruction and research in reading comprehension. *Exceptional Education Quarterly* 1981, 2, 27-40.
- JONES, L., & FISKE, D. W. Models for testing the significance of combined results. *Psychological Bulletin* 1953, 50, 375-381.
- KENNEDY, M. Findings from the Follow Through planned variation study. *Educational Researcher* 1978, 7, 3-11.
- LORET, P. G., SEDER, A., BIANCHINI, J. C., & VALE, C. A. *Anchor test study: Equivalence and norms tables for selected reading achievement tests (grades 4, 5, 6)*. (Office of Education Report No. 74-305). Washington, D.C.: U.S. Government Printing Office, 1974.
- NATIONAL CENTER FOR EDUCATIONAL STATISTICS. *The condition of education*. Washington, D.C.: U.S. Government Printing Office, 1978.
- OZENNE, D., et al. United States Office of Education. *Annual evaluation report on programs administered by the U.S. Office of Education FY 1973*. Washington, D.C.: Capital Publications, Educational Resources Division, 1974.
- OZENNE, D., et al. United States Office of Education. *Annual evaluation report on programs administered by the U.S. Office of Education FY 1975*. Washington, D.C.: Capital Publications, Educational Resources Division, 1976.
- PERFETTI, C. A., & LESGOLD, A. M. Coding and comprehension in skilled reading and implications for reading instruction. In L. B. Resnick & P. A. Weaver (Eds.), *Theory and practice of early reading* (Vol. 1). Hillsdale, N.J.: Lawrence Erlbaum Associates, 1977.
- PILLEMER, D. G., & LIGHT, R. J. Synthesizing outcomes: How to use research evidence from many studies. *Harvard Educational Review* 1980, 50, 176-197
- RESNICK, L. B. Instructional psychology. In M. R. Rosenzweig & L. W. Porter (Eds.) *Annual Review of Psychology* Palo Alto, Calif.: Annual Reviews, 1981.
- SILBERT, J., CARNINE, D., & STEIN, M. *Direct instruction arithmetic*. Columbus, OH Charles E. Merrill, 1981.
- SMITH, M. L., & GLASS, G. V. Meta-analysis of psychotherapy outcome studies *American Psychologist* 1977, 32, 752-760.
- STEBBINS, L. (Ed.). *Education as experimentation: A planned variation model* (Vo IIIA). Cambridge, Mass.: Abt Associates, 1976.
- STEBBINS, L., ST. PIERRE, R. G., PROPER, E. L., ANDERSON, R. B., & CERVA, T. R. *Education as experimentation: A planned variation model* (Vols. IVA-D). Cambridge, Mass.: Abt Associates, 1977.
- STEVENS, R., & ROSENSHINE, B. Advances in research on teaching. *Exceptional Education Quarterly* 1981, 2, 1-10.

- TALLMADGE, G. K. *JDRP ideabook*. Washington, D.C.: USOE/NIE Printing, 1977.
- WOLF, R. Review of Metropolitan Achievement Test. In O. K. Buros (Ed.), *The sixth mental measurements yearbook*. Highland Park, N.J.: Gryphon Press, 1978.

Authors

- WESLEY C. BECKER, Associate Dean for Division of Counseling and Educational Psychology, University of Oregon, Eugene, Oregon 97403. *Specializations*: Education of the Disadvantaged, behavior theory and education, evaluation of instruction.
- RUSSELL M. GERSTEN, Research Associate, Follow Through Project. University of Oregon, Division of Special Education. *Specializations*: Special education and compensatory education, program evaluation, assessment.