DOCUMENT RESUME

ED 097 120 PS 007 493

AUTHOR Siegel, Martin A.
TITLE An Experimental Investigation of Teacher Behavior and Student Achievement in the Distar Instructional System.
PUB DATE Apr 74
NOTE 40p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Academic Achievement; Classroom Observation Techniques; Curriculum Evaluation; Discriminant Analysis; *Early Childhood Education; *Instructional Systems; *Programed Instruction; Program Effectiveness; Sequential Learning; Social Reinforcement; Standardized Tests; *Teacher Behavior; Teacher Education; Teacher Workshops
IDENTIFIERS Criterion Teaching; *Distar

ABSTRACT

Fifty randomly selected moderate and low implementing Distar Language teachers were retrained in techniques of correcting students' mistakes and recycling through an instructional task until all of the children in the group responded without error (criterion teaching). As a result of retraining, the experimental teachers performed at a significantly higher level of implementation than the control teachers. In addition, there were significant differences (favoring the experimental group) in achievement (for both a program-specific and program-general measure) on the students' statistically adjusted posttest scores. Implications for research and training in teacher education are discussed. (Author)

AN EXPERIMENTAL INVESTIGATION OF TEACHER BEHAVIOR AND STUDENT ACHIEVEMENT IN THE DISTAR INSTRUCTIONAL SYSTEM

Martin A. Siegel

College of Education
Computer-Based Education Research Laboratory
University of Illinois at Urbana-Champaign

Paper presented at the annual meeting of the
American Educational Research Association
Chicago, April, 1974
Session 26.10

Introduction

Among the many curriculum programs that have been developed, there has been little research that has investigated the relationship between teacher behaviors as prescribed by the curriculum developers and student outcomes such as achievement or attitudes. The research on teacher behaviors within curriculum programs generally falls into two major categories. First, there are studies which describe curriculum specific teacher behaviors--those instructional activities or behaviors which are hypothesized to be important for the success of a given program--but do not relate these activities to student gains (for example, Olivero, undated; Gallagher, 1966, 1968; Katz, 1968; Lindvall and Cox, 1970; Niedermeyer and Dalrymple, 1970). Second, there are studies which relate curriculum general teacher behaviors--those instructional activities or behaviors which are hypothesized to be important for the success of all or many programs--to student outcomes (for example, LaShier, 1967; Walberg, 1969; Flanders, 1970; Soar, 1971; Soar, Soar, and Ragosta, 1971).[1/]

Unfortunately the results of especially the first group of studies can have limited impact on the development or assessment of the teacher training programs within particular curriculum packages, or on the modification of the curriculum materials themselves. The descriptive studies, although suggesting wide variation in events within classrooms using a particular curriculum package, do not relate the variation to student outcome measures. For example, Gallagher (1966) counted various types of activities which occurred in the classrooms of six teachers who were teaching the same unit from the Biological Sciences Curriculum Study (BSCS)

1/ Some behaviors which are curriculum specific may also be curriculum general. That is, instructional activities or behaviors which are important for a particular program may be important for other programs as well. The distinction between the two types of behaviors reflects two different research approaches. One research approach is to study those behaviors which are important across all or most teaching situations. Another research approach is to study those behaviors which are important within specific teaching situations.

program. On almost all measures of teacher behavior there were significant differences among the six teachers. Regretably, the investigator did not relate this variation to measures of student outcomes. Does an increase in inquiry-strategy behaviors which are intended by the BSCS curriculum planners enhance or suppress student achievement or is the effect negligible? Given a behavior that affects cognitive gains, what are the concommitant effects in attitude towards the curriculum, towards the school, or towards the child?

While the second group of studies do attempt to relate instructional activities to measures of student outcomes, the observational instruments used were designed to apply to all types of programs and educational settings. For example, Soar (1971; Soar, Soar, and Ragosta, 1971) has been monitoring eight classrooms in each of seven Follow-Through programs along with two comparison classrooms for each program. Instead of developing program-specific observation instruments Soar used four general observational systems: the Reciprocal Category System (Ober, #61,[2/] an expansion of the Flanders system, #5), the Florida Taxonomy of Cognitive Behaviors (K-1 Form, (Brown, et al, #37), the Teacher Practices Observation Record (Brown, #36), and the Florida Climate and Control System (Soar, 1966; Soar, Soar, and Ragosta, 1971). Although the investigators correlated the factor scores derived from the four instruments with measures of class mean residual gain, it is plausible that the most critical variables which affect student gains are those which were not included in the general observational instruments. The ability to follow a prespecified format without even minor deviations may be an important variable in the Engelmann-Becker program, whereas in the Bank Street Program, the ability to elaborate on a child's experiences may be essential to the realization of the program's goals and objectives. However, a general observation instrument is likely to be insensitive to either of these program-specific variables. Therefore, in addition to general instruments, development of observational measures specific to the instructional activities most emphasized by the curriculum designers seems useful.

2/ Numbers such as this refer to those assigned each observational system in Mirrors for Behavior, (Simon and Boyer, 1967, 1970a, 1970b).

Rosenshine and Furst (1973) suggested that research on a particular curriculum materials package should consist of five phases:

1. Train a group of teachers to use a certain package of materials which have already received extensive trial and modification within special settings (for example, any of the Follow-Through programs like the Bank Street Program, Bushell's Behavior Analysis Program, or Engelmann and Becker's Distar Program; BSCS; First Year Communication Skills Program; or Harvard Project Physics).

2. Use observational systems to describe instructional variables which are considered specific to the program and most emphasized by the curriculum planners and which are also considered to have general educational importance (and may or may not be emphasized by the curriculum designers).

3. Study the relationship between instructional activities and behavioral change in the students in a variety of outcomes. Problems and suggestions for selecting measures of instructional behaviors and student growth on outcomes of interest, and for data analysis and design are presented elsewhere (Medley and Mitzel, 1963; Gage, 1969; Flanders, 1970; Rosenshine, 1970, 1971; Rosenshine and Furst, 1971, 1973; Tatsuoka, 1972).

4. Modify the training procedures and/or materials on the basis of the studies completed in phases two and three.

5. Conduct new studies with appropriate control groups to determine the effects of the modifications and to determine the new relationships between instructional activities and student growth. By recycling through phases one through four, the curriculum designer, publisher, and researcher successively approximate optimum training procedures, thus affecting gains in student achievement or other outcomes of interest.

Although Rosenshine and Furst's "descriptive-correlational-experimental-loop" design for curriculum research and evaluation is not unique (see Tatsuoka, 1972), no study was found which included all phases of the design. Research studies which include part of the "loop" exist. However, even this type of instructional research

within curriculum programs is rare. In fact, only two studies were found which included the training, descriptive, and correlational phases and also used program-specific variables: Kochendorfer (1967) and Baker (1969).

Background of Experimental Study

Description of the Distar Instructional System

One of the most successful (Science Research Associates, 1971a) and controversial of all the early childhood curriculum materials programs is Distar Reading, Language, and Arithmetic (Engelmann and Bruner, 1969, 1970; Engelmann and Carnine, 1969, 1970, 1972; Engelmann, Osborn, and Engelmann, 1969; Engelmann and Osborn, 1970, 1972; Engelmann and Stern, 1972), a commercial model of the Engelmann-Becker (Bereiter-Engelmann) Follow Through program. Unlike other programmed materials, the Distar program is not a self-instructional program. Instead, the teacher follows a carefully structured and logically sequenced teaching program. The presentation books provide the teacher with a script, a series of demonstrations and tasks to be presented word for word. The teacher's role thus changes from one of designing instruction to one of teaching a particular format to criterion, involving all of the children in the instruction, correcting mistakes, providing feedback, and reinforcing the children's responses.[3/]

Teacher Implementation Variables

Five areas of teacher behavior are emphasized throughout teacher guides and training manuals: (1) following the format--using the exact wording provided in the materials, not producing additional statements or asking questions unless the format calls for them; (2) signals--using clear signals for the children to respond so that they all respond at the same time and not imitate other children's responses; (3) corrections and criterion teaching--correcting children's mistakes as they occur

3/ For a more complete outline of the philosophy and methods used in the Engelmann-Becker program the reader is referred to Engelmann (1969a, 1969b) and Maccoby and Zellner (1970).

and requiring that they return to the beginning of the task so that they may recognize each exercise as a series of steps which relates to a goal and to certain rules; (4) praise and feedback--reinforcing the children who are on task and relating their performance to the rules of the task; and (5) pacing--moving at varying speeds through the lesson so that the children will understand the point of each task and so that their interest and enthusiasm will be maintained.

According to the Distar curriculum authors, these are the basic implementation variables. It is assumed that if a teacher behaves in these ways the children will achieve the academic objectives of the Distar program. That is, the Distar curriculum developers hypothesize that the above teacher behaviors are directly related to student achievement.

Experimental Study

Purpose

The purpose of the experimental study was to determine (1) the feasibility of modifying the implementation level of a randomly selected group of Distar teachers, and (2) the effects of such training upon their students' achievement level. Two related aspects of teacher behavior were chosen for experimental manipulation: correction procedures and criterion teaching. These variables were selected because (1) they are unique to the Distar program and have received consistent and statistically significant support in previous correlational studies (Siegel and Rosenshine, 1973); and (2) they are the most difficult behaviors for most Distar teachers to implement appropriately. Establishing the importance or lack of importance for these variables would greatly affect the design of future Distar training programs.

Procedure

Teachers and program. Fifty teachers from a large Southwestern school district received two days of in-service training before they began teaching Distar Language I (Engelmann, Osborn, and Engelmann, 1969). The program emphasizes the language of

instruction--that is, the actual language the teacher uses in the classroom--and systematically takes a child from identification of familiar objects to the description and classification of those objects as well as concepts of logical reasoning. The training program, (SRA, 1971b) focused on teaching selected formats, analyzing taks, correcting mistakes, and general procedures for implementing the Distar system.

Students. The children in each classroom were divided into three "homogeneous" groups based on a criterion-referenced, individually administered pretest of language skills taught in the first 80 lessons of the program. One group from each classroom was randomly selected for further study. Each group consisted of five to ten first-grade children who had no previous experience with Distar. Forty-six percent of the children in the study were Mexican-Americans, 37 percent Black, and 17 percent Anglo.

Collection and coding of classroom observation data. Each of the 50 teachers was audiotaped while teaching one lesson during each of six one-week segments. The six taping segments were equally spaced throughout a seven-month time period. The teachers did not know that they were to be taped until about five minutes prior to the teaching of a lesson. The research assistant turned on the cassette recorder, adjusted the volume, and left the room. She returned in 30 minutes to collect the tape. The teacher was never permitted to listen to the recording. For the most part, the content of the lessons for each taping occasion was rather uniform across teachers.

The audiotapes were coded by three teams of two graduate students each. An observational system developed especially for coding any lesson in the Distar Instructional System was used. The instrument was developed by Siegel and a team of Distar authors and trainers.[4/]

4/ S. Engelmann, J. Osborn, E. Bruner, P. Mahan, L. Meyers, D. Granat, and B. Rosenshine (consultant).

Interrater agreement. Although all team members had previous experience with the Distar Instructional System Observation Instrument (DISOI), a system was devised to establish and to maintain consistency within teams, since each team coded separate categories. Initially, members of each team coded five randomly selected tapes separately and then together to establish their consensus in identifying and rating the behaviors specified by the instrument. Each team member then received a set of ten tapes to code. Upon completing the set, each team member randomly selected one tape for the other team member to recode. After the recoded tapes were independently coded for a second time, the team met to agree upon their ratings and, if necessary, to modify or extend the coding rules. This procedure was repeated until all 300 tapes were coded, thus providing 30 tapes which were used to determine interjudge correlations. For every category a judge's score (rating or counting) for each task of a lesson was correlated with the other judge's score for each task of the same lesson. Table 1 presents the range of interjudge correlations for each of the categories.

Observation instrument. Although DISOI was designed to be used with all these Distar programs (Language, Reading, and Arithmetic), a series of "ground rules" were developed and explicitly stated so that consistent decisions could be made among coders while listening to audiotapes of teachers presenting the Language program. The following is a description of two categories of the Distar Instructional System Observation Instrument. The entire instrument is presented elsewhere (Siegel, 1973b).

Corrections

The paradigm for correcting basic mistakes (according to the Distar Orientation Manual--Revised edition) is as follows: (Note: Repeating the entire task from the beginning will be considered a part of "criterion teaching").

Type I Mistake--Lacks Information

1. Teacher gives the answer (A) or provides additional information (Ai).

2. Teacher tests the child by repeating the segment missed (T).

Type II Mistake--Motor/Speech Problem

1. Teacher gives the answer (A).

2. Teacher repeats the signal (R).

3. Teacher leads the child (L_n).

4. Teacher tests the child by repeating the segment missed (T).

Type III Mistake--Does Not Understand Signal

1. Teacher repeats the signal or calls attention to the signal (R).

2. Teacher or another child models the response (M).

3. Teacher tests the child by repeating the segment missed (T).

Definitions

a. Giving the answer (A) is simply telling the child the correct response.

b. Providing additional information (Ai) is not telling the child the entire answer but merely providing extra information so that the child can "come up with" the correct response.

c. Testing the child or children (T) is asking the question again or requiring the child(ren) to respond.

d. Leading (L_n) is responding with the child. The teacher and the child simultaneously say the response. The "n" indicates how many times the teacher says the response with the child. Ideally the teacher should lead two or more times.

e. Repeating the signal (R) is identical to testing the child. The teacher asks the question again or repeats the command to respond. The difference in labelling is a function of when the question is reasked. If it appears as the first step in the paradigm, then it is (R); if the question is repeated as the last step in the paradigm, then it is (T). Normally the teacher does not expect the child to correctly answer when she "repeats the signal." The purpose of this step is merely to call attention to what the children should be responding to.

f. Modeling (M) is performing the teacher's part and the child's part. This is done to demonstrate to the child how the two parts are related. The teacher asks the question (or requires a response) and then answers the question (or performs the response). Note that in (A) the teacher only answers the question (or performs the correct response).

This category is coded as follows:

A. Each mistake is classified as either Type I, II, or III.

B. The teacher's correction procedure is indicated by using the above key.

Example: "R M T" would indicate that the teacher repeated the signal, modeled the response, and tested the child on the segment missed.

Example: "R L_3" would indicate that the teacher repeated the signal and led the response three times.

C. Each correction is rated according to the following eight-point rating scheme:

8: Teacher corrects the mistake immediately after it occurs.

If the mistake is Type I (lacks information), teacher gives the answer (A) or provides information (Ai);

If the mistake is Type II (motor/speech), teacher gives repetition activities (L_n) where n is greater than or equal to two. (The teacher may proceed the leading with giving the answer (A) and/or repeating the signal (R).);

If the mistake is Type III (does not understand the signal), teacher repeats the signal or calls attention to the signal (R) and teacher or another student provides a model (M);

then the teacher tests the child (or group) by repeating the segment of the task that was missed (T).

Example for Type I mistakes: AT

Examples for Type II mistakes: L_2T; L_5T; AL_3T; RAL_4T

Examples for Type III mistakes: RMT

7: Teacher pairs the type of mistake with the correct procedure as indicated above but adds additional procedures and tests.

Examples for Type I mistakes: MT (note that A is included in M); AE (note that T is included in E--repeating the entire task from the beginning); ML_1T; RAT

Examples for Type II mistakes: ML_2T; $RAML_3T$

Examples for Type III mistakes: $ARML_3T$; RL_1MT

6: Teacher pairs with type of mistake an incorrect procedure (including omitting steps) and tests.

Examples for Type I mistakes: L_2T; RL_1T

Examples for Type II mistakes: AT; RMT

Examples for Type III mistakes: MT; AT; L_3T; AL_2T

5: Tests only (T) or repeats the entire task only (E).

Examples for Type I mistakes: T; E

Examples for Type II mistakes: E; T

Examples for Type III mistakes: T; E

4: Like 8 but with no test.

3: Like 7 but with no test.

2: Like 6 but with no test.

1: No correction at all or giving the wrong answer or information.

Criterion Teaching

This behavior is counted whenever it occurs. Notice that the behavior may or may not follow an appropriate correction procedure. That is, this category is independent of the rating received for category II, corrections.

A. For Type I mistakes, the teacher repeats the entire task from the beginning (E1).

B. For Type II mistakes, the teacher repeats the entire task from the beginning (E2).

C. For Type III mistakes, the teacher repeats the entire task from the beginning (E3).

D. For Type I, II, or III mistakes, the teacher repeats a segment of the task but not the entire task (S). A segment of a task consists of repeating one or more signals prior to the signal missed--but not from the beginning.

E. The teacher recycles through (repeats) a segment of the task or the complete task (RC). This behavior differs from A-D in that the recycling does not immediately follow a mistake. The sequence is as follows: (1) a child makes a mistake; (2) the teacher may or may not correct the mistake (appropriately); (3) the teacher proceeds with the next part of the task; and (4) the teacher recycles through a portion of the task or the entire task.

Although the primary interest of this study was to determine the relationships between correction procedures, criterion teaching and student achievement (variables 4-7; 9-15) data was collected for other variables as well. This enabled an examination of possible changes in other behaviors as a function of training in correction procedure and criterion teaching behaviors. Rating sheets were used to code the data. For each of the 27 variables the total score was divided by the total number of tasks (or total number of mistakes, in the case of corrections and criterion teaching) which occurred in a given 30-minute taping.

TABLE 1

List of Variables and Range of Interjudge Correlations for Each Category

Variable number	Abbreviation	Type of variable	Brief description of variable	Range of interjudge correlations in coding audiotapes
1	FG	Rating-5 point scale	Following the format	.90 - .96
2	FI	Rating-5 point scale	Following the format-individual tasks	.92 - .95
3	FT	Rating-5 point scale	Following the format-total tasks (sum of variables 1 and 2)	
†4	C1	Rating-8 point scale	Corrections-Type I mistakes (lacks information)	.93 - .97
†5	C2	Rating-8 point scale	Corrections-Type II mistakes (motor/speech problem)	.86 - .94
†6	C3	Rating-8 point scale	Corrections-Type III mistakes (does not understand signal)	.90 - .95
†7	CT	Rating-8 point scale	Corrections-total mistakes (sum of variables 4-6)	
8	#MIS	Counting	Total number of mistakes by students	
†*9	E1	Counting	Repeating the entire task from the beginning-Type I mistakes	.96 - .98
†*10	E2	Counting	Repeating the entire task from the beginning-Type II mistakes	.94 - .96
†*11	E3	Counting	Repeating the entire task from the beginning-Type III mistakes	.96 - .97
†*12	ET	Counting	Repeating the entire task from the beginning-total mistakes (sum of variables 9-11)	
†*13	S	Counting	Repeating a segment of the task-total mistakes	.94 - .95
†*14	RC	Counting	Recycling through the task but not following a mistake (recycling through a segment or the entire task)	.94 - .95
†*15	ET+S+RC	Counting	Criterion teaching-total (sum of variables 12-14)	

TABLE 1 (continued)

Variable number	Abbreviation	Type of variable	Brief description of variable	Range of interjudge correlations in coding audiotapes
**16	NP	Counting	Number of signals that the teacher does not pause a sufficient amount of time for the students to respond	.89 - .92
17	NU	Counting	Number of student nonunison responses to a signal	.93 - .96
*18	A/NU	Counting	Ratio of number of attempts to obtain a unison response to the number of nonunison responses	.93 - .96
19	LNI	Counting	Leading which is not indicated by the format	.91 - .91
20	LI	Counting	Leading which is indicated by the format	.91 - .94
21	RA	Counting	Number of times the teacher repeats the correct answer	.92 - .94
22	GTP	Counting	Number of times the teacher provides general praise after a correct task response	.93 - .94
**23	STP	Counting	Number of times the teacher provides specific praise after a correct task response	.91 - .91
**24	GBP	Counting	Number of times the teacher provides general praise after an appropriate nontask behavior response	.87 - .92
**25	SBP	Counting	Number of times the teacher provides specific praise after an appropriate nontask behavior response	.88 - .93
**26	IP	Counting	Number of inappropriate responses that the teacher praises (general or specific)	.89 - .96
**27	IC	Counting	Number of appropriate responses that the teacher corrected	.90 - .93

* An angular or inverse sine transformation was used with this variable: $X' = 2 \arcsin \sqrt{X}$, where X is expressed as a proportion. 1/ (2n) was substituted for X = zero.

** A square-root transformation was used with this variable: $X' = \sqrt{X + .5}$.

† Experimental (retraining) variables.

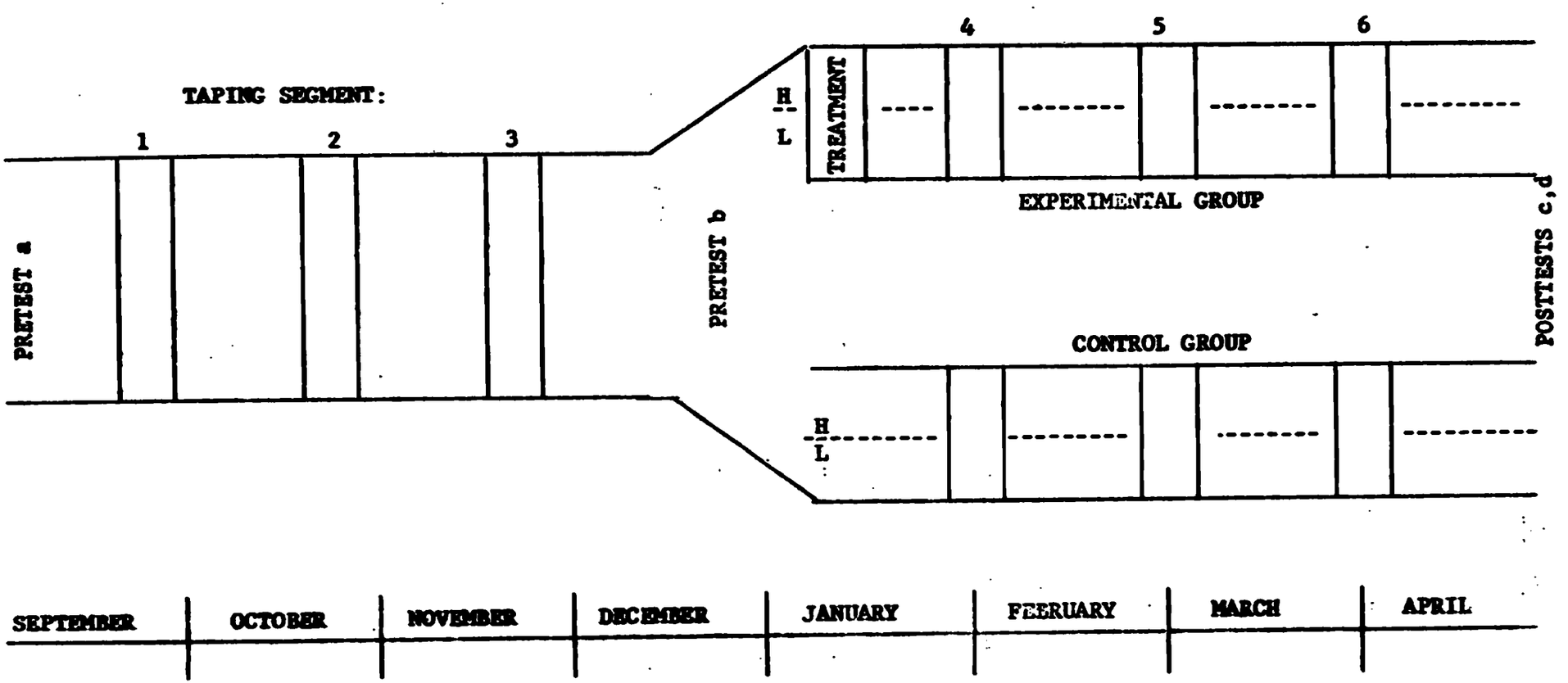
Selection of experimental and control group teachers. After the first three observations were coded, the teachers were divided into "high" and "low" implementors according to their average scores on variable 7 (CT) and 15 (ET + S + RC)--correction-total mistakes and criterion teaching, respectively.[5/]

The high and low implementors were then randomly divided into two groups of 25 teachers each. (See Figure 1) One group was randonly designated the experimental group, and these teachers were invited to attend a one-day retraining workshop. Unfortunately only 23 of the 25 invited teachers attended the retraining session. The two teachers that did not attend were high implementors. Their principals explained that they were already excellent teachers and did not need retraining. Therefore, they were added to the control group (27 teachers). This probably biased any possible outcomes in favor of the control group.

Second pretesting of students. Prior to the retraining workshop, a mid-year pretest was administered by a trained research assistant to all children in the experimental and control groups. The measure used was the Continuous Test for Language One (CTL) (Engelmann-Becker Corporation, 1971). This test is used by the Engelmann-Becker Follow Through Model to measure pupil performance in the Distar Language program. The oral test is criterion-referenced and individually administered.

The CTL consists of several subtests which measure mastery of each major skill taught in the level I Language program: object concepts, action statements, parts, sentence repetition, prepositions, categories, plurals, pronouns, function words, verb tense, comparatives-superlatives, if-then, before-after, only. For the purposes of the pretest, one-third of the 114 items was administered. These items

5/ More specifically, principal-components analysis was used to determine two linear composite scores for each teacher. The rank order of each teacher's sum of composite scores then indicated her degree of implementation.



PRETESTS:

a. Distar Language Pretest (DLP)
b. Continuous Test for Language One (PCTL)

POSTTESTS:

c. Continuous Test for Language One (CTL)
d. APELL Test (APELL)

Figure 1. Scope and sequence of experimental study.

included concepts from lessons 24 to 93. (There are 180 lessons in the Language I program.)[6/]

An example of two items appears in Figure 2. If an item had several parts, all parts must have been correct for the item to be scored "pass."

Description of treatment. The experimental teachers were provided with one day of retraining on the behaviors of "correction procedures" (variables 4-7) and "criterion teaching" (variables 9-15).[7/] During the week following the workshop, each teacher also received 30 minutes of classroom consultation. The trainer concentrated on the above two sets of behaviors while observing the teacher teach a lesson. The control teachers received no additional training, nor did they receive classroom consultation.

The one day retraining workshop (Siegel, 1973a) focused on several of the teaching problems which were noted during the coding of each teacher's three pretreatment audiotapes. These were as follows:

a. distinguishing an appropriate response from a mistake;

b. distinguishing different types of mistakes;

c. correcting a mistake according to the type of mistake;

d. correcting a mistake even though several children in the group responded appropriately;

e. testing the child or children on the segment of the task that was incorrect; and

f. returning to the beginning of a task after a mistake was corrected.

A simplified three-step correction paradigm was presented during the retraining session (see Figure 3).

6/ The reliability of the 38 items for all children in the experimental and control groups was found to be .89, using KR20.

7/ In a strict sense, criterion teaching--returning to the beginning of a task after a mistake has been corrected--is reflected only in variables 9-12 (E1, E2, E3, ET, respectively). However, a less conservative interpretation of "criterion teaching" would also include variable 13 (returning to a segment of a task after a mistake has been corrected), variable 14 (recycling through the task for "firm-up" purposes) and variable 15 (the total of variables 12-14).

58. Note: There is no illustration for this item.

(a) (Present pencil and glass.) Put the pencil under your chair.
(b) (If child places pencil under chair) Where is the pencil? (wait)
(c) Say the whole thing. (wait)
(d) Put the glass on your chair.
(e) (If child places glass on chair) Where is the glass?
(f) Say the whole thing. (wait)

Answers:

(a) Child puts pencil under chair.
(b) Under the chair.
(c) The pencil is under the chair.
(d) Child puts glass on chair.
(e) On the chair.
(f) (The) glass is on (the) chair.

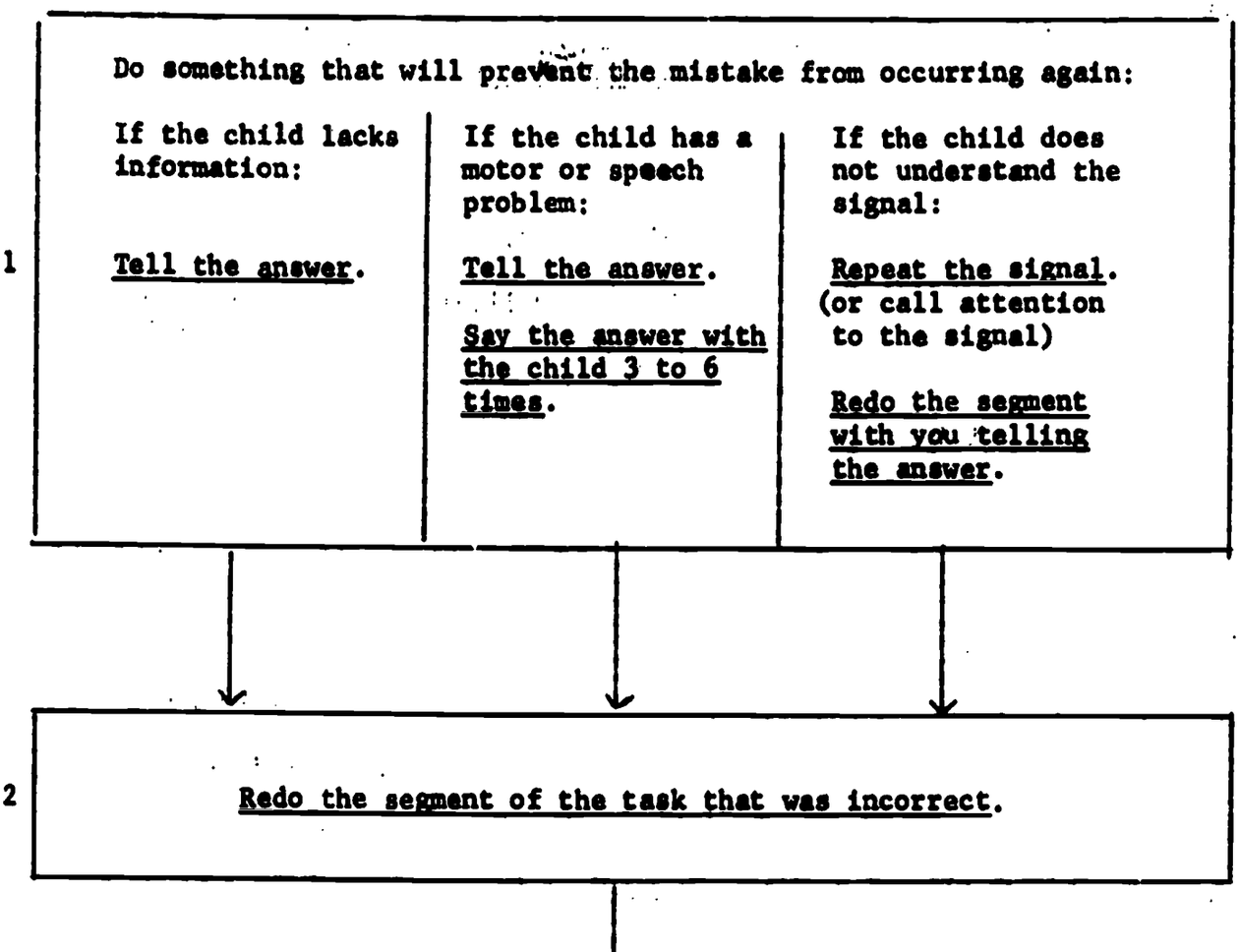
82. (Picture of two girls: one sitting and with short hair; one running and with long hair; both girls are wearing dresses.)

(a) (Present picture.) Who could say, "You are running?" Show me. (wait)
(b) (Point to sitting girl.) She could say, "You are running." Who would she say that to? (wait)
(c) Who could say, "My hair is long." (wait)
(d) (Point to girl with long hair.) What could she say? (wait)
(e) Who could say, "We are wearing dresses." (wait)
(f) (Point to both girls.) What could they say? (wait)

Answers:

(a) Child points to sitting girl.
(b) Child points to running girl.
(c) Child points to running girl.
(d) My hair is long.
(e) Child points to both girls.
(f) We are wearing dresses.

Figure 2. Example of two items from the Continuous Test for Language One.



Note: If the mistake was a motor or speech problem, it will be impossible to teach to criterion in one day. At step three, we therefore expect only incremental improvement. The goal in correcting other types of mistakes is that at step 3, all the children in the group can perform the entire task--from start to finish--without making a mistake.

Figure 3. Three-Step Correction Paradigm used in Retraining Workshop. (Siegel, 1973, p. 33)

Posttesting of students. All children were posttested on the following measures:

1. The Engelmann-Becker Continuous Test for Language One (Engelmann-Becker Corporation, 1971); and
2. The APELL Test (Edcodyne Corporation, 1971).

Both measures were individually administered by a trained research assistant. Fifty-two items were selected from the larger 114-item Continuous Test, spanning lessons 44 through 133. The APELL Test consisted of 25 items and all were given. The reliability of these measures (KR20) was .92 for the CTL and .66 for the APELL.

The APELL Test represented a measure of language achievement which was not specifically developed to test outcomes of the Distar Language program (although there were overlapping objectives between the test and the program). The measure was criterion-referenced and assessed six areas of language development: nouns, pronouns, verbs, adjectives, plurals, and prepositions. Each item of the APELL Test consisted of three pictures. The child was asked to touch one of the three pictures. For example, item 22 displayed a picture of a cup under a table, a picture of two cups on a table, and a picture of a cup on a table. The tester said, "Look at the pictures. Put your finger where there is nothing on the table."

Results

There were actually two main parts to the question, "Did the treatment work?." The first was whether there were significant differences between the experimental and control group of teachers in behavior on the variables of correction procedures and/or criterion teaching. The second was whether there were significant differences in adjusted achievement on the students' posttests between the experimental and control group of teachers. Each part was further subdivided into two additional questions: were there differential and interaction effects for the high and the low implementing teachers?

Analysis of Teacher Behavior

Multivariate analysis of the retraining variables. A "mixed" or "split-plot" MANOVA design was used to answer the question of significant modification of (changes in) behavior. The design had two between-block treatments (A) and (C) and one within-block treatment (B). Treatment A had two levels--experimental and control; treatment C had two levels--high and low implementation on correction procedures and criterion teaching. Treatment B also had two levels--the mean score of audiotaping observations 1-3 and the mean score of audiotaping observations 4-6. That is, both sets of observations were treated as dependent variables obtained at two different levels of time factor--one prior to the treatment and one subsequent to the treatment.

Table 2 presents the repeated measure MANOVA for the 11 retraining variables. The siginificatn ($p < .001$) Treatment by Occasion (AB) interaction and inspection of the group centroids indicated that the experimental group after the retraining session performed significantly higher than the control (no retraining) group. There were no significant differences ($p > .05$) in implementation before the retraining. Discriminant analysis furthermore revealed that the criterion teaching (ET) variable--repeating the entire task after a mistake has been corrected--contributed most to the dimension along which the difference between the experimental and control group's "post minus pretreatment" difference was maximized.

Univariate analyses of retraining and nonretraining variables. For each of the correction procedure and criterion teaching variables, the treatment group by occasion interaction (AB) effect was significant ($p < .05$). Furthermore, a posteriori comparisons among group means, using Tukey's HSD test, revealed that for each of the 11 retraining variables, the experimental group teachers after retraining were rated significantly ($p < .05$) higher implementors than the control

TABLE 2

Repeated Measures MANOVA for the 11 Retraining Variables

Source of Variation	Degrees of Freedom	F-Ratio for Multivariate Test of Equality of Mean Vectors
Treatment (A)	11/36	5.290***
Implementation (C)	11/36	3.395**
A X C	11/36	0.750
Occasion (B)	11/36	13.588***
A X B	11/36	9.186***
B X C	11/36	1.955
A X B X C	11/36	1.241

*$p < .05$

**$p < .01$

***$p < .001$

(no retraining) group teachers, although there were no significant differences between the ratings of the two groups prior to the treatment. Table 3 presents the F - ratios and probabilities for the 11 retraining variables. Figure 4 illustrates for the total criterion teaching score (variable 12) the treatment by occasion interaction at each level of implementation.

Pairwise comparisons of the AB group means also indicated that the teachers in the experimental group significantly ($p < .05$) improved in performance from the first three taping occasions to the second three taping occasions. This result was found with each of the 11 variables, whereas for the control group teachers, significant improvement in performance over the two sets of taping occasions occurred only with variables 4, 7, 13, and 15 (see Table 3). However, even though the control group teachers significantly improved over time on four of the variables, the experimental teachers not only made significant improvements over time, but these improvements were also significantly greater than the control teachers on all variables!

Contrary to what might be expected, the retraining program did not differentially affect the high and low implementors of the experimental group. That is, the high and low experimental teachers made "equal" gains in implementation (see for example Figure 3).

Although only the correction procedure and criterion teaching variables were specifically emphasized during the retraining of the experimental teachers, analyses for nonretraining variables are presented in Table 4. Of course, caution should be taken in interpreting the results of a series of separate univariate analyses.

Perhaps the most interesting result of the analysis of nonretraining variables was the experimental group's significant increase over time, as compared to the control group's performance, on variables "repeating the correct response

TABLE 3

F-Ratios and Probabilities for the Retraining Variables

Variable		Source of Variation					
		Treatment (A)		Implementation (C)		A X C	
Number	Description	F	p<	F	p<	F	p<
4	Corrections--Type I Mistakes	8.762	.005	23.176	.000	.126	.724
5	Corrections--Type II Mistakes	.122	.729	11.986	.001	2.329	.134
6	Corrections--Type III Mistakes	.815	.371	14.026	.000	.085	.772
7	Corrections--Total Mistakes	6.700	.013	28.622	.000	.000	.986
9	Repeating Entire Task-I	47.898	.000	4.403	.041	.263	.611
10	Repeating Entire Task-II	5.284	.026	3.695	.061	.207	.651
11	Repeating Entire Task-III	8.702	.005	.760	.388	1.866	.179
12	Repeating Entire Task-Total	35.021	.000	5.779	.020	.638	.428
13	Repeating Segment-Total	10.383	.002	17.076	.000	.509	.479
14	Recycling Through Task	7.137	.010	4.155	.047	1.968	.167
15	Sum of Variables 12, 13, and 14	33.016	.000	16.719	.000	.072	.790

TABLE 3 (continued)

Variable		Source of Variation							
		Occasion (B)		A X B		B X C		A X B X C	
Number	Description	F	p <	F	p <	F	p <	F	p <
4	Corrections--Type I Mistakes	44.289	.000	11.236	.002	11.942	.001	1.065	.308
5	Corrections--Type II Mistakes	7.878	.007	7.589	.008	3.267	.077	4.031	.051
6	Corrections--Type III Mistakes	5.579	.022	10.396	.002	8.996	.004	5.821	.020
7	Corrections--Total Mistakes	40.854	.000	18.344	.000	14.018	.000	2.719	.106
9	Repeating Entire Task-I	87.600	.000	81.793	.000	.585	.448	.003	.956
10	Repeating Entire Task-II	7.398	.009	4.053	.050	2.833	.099	.609	.439
11	Repeating Entire Task-III	19.375	.000	25.957	.000	1.433	.237	.876	.354
12	Repeating Entire Task--Total	100.303	.000	74.411	.000	.001	.980	.315	.578
13	Repeating Segment--Total	17.363	.000	4.067	.049	.014	.908	.359	.552
14	Recycling Through Task	18.534	.000	8.513	.005	.241	.626	.273	.604
15	Sum of Variables 12, 13, and 14	126.793	.000	57.969	.000	.286	.596	.411	.524

Note: Degrees of Freedom: 1 and 46.

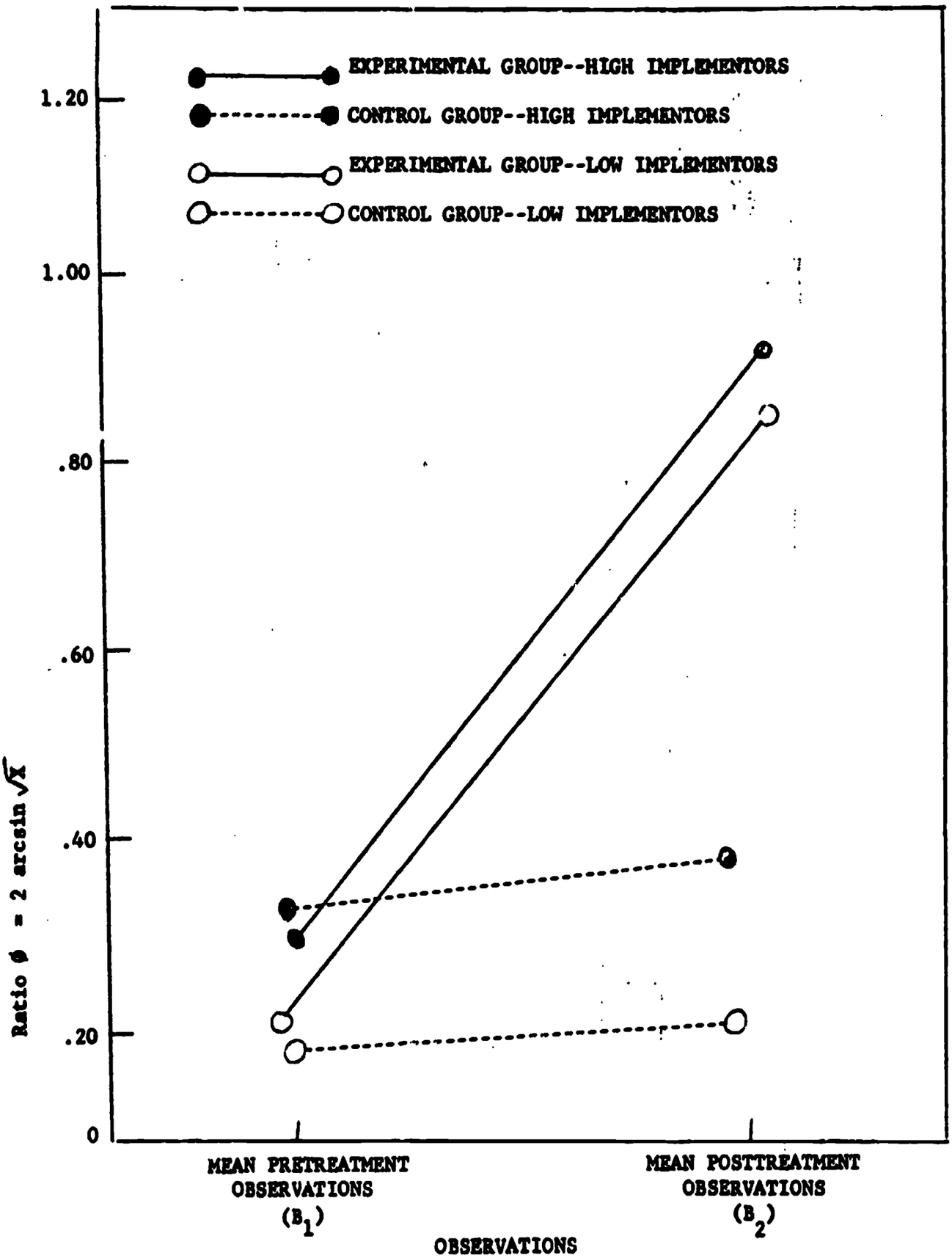


Figure 4. Treatment (A) by Observation (B) interaction at each level of implementation (C) for variable 12 (criterion teaching following all mistakes).

TABLE 4

Summary of Repeated Measures ANOVA for the Nonretraining Variables

Variable[a]		Source of Variation--Probability						
Number	Abbreviation	Treatment (A)	Implementation (C)	A X C	Occasion (B)	A X B	B X C	A X B X C
1	FPG	NS[b]	.002	NS	.010	.001	NS	NS
2	FFI	NS	.025	NS	NS	NS	NS	NS
3	FFT	NS	.004	NS	.022	.002	NS	NS
8	#MIS	NS	NS	NS	.015	NS	NS	NS
16	NP	NS	NS	NS	.001	NS	NS	NS
17	NU	NS	NS	NS	.013	.004	NS	NS
18	A/NU	NS	NS	NS	NS	NS	NS	NS
19	LNI	NS	NS	NS	NS	NS	.013	NS
20	LI	NS	NS	NS	.049	NS	NS	NS
21	RA	.050	NS	NS	.035	.045	NS	NS
22	GT	.036	NS	NS	.023	.004	NS	NS
23	ST	NS	NS	NS	.033	NS	NS	NS
24	GB	NS	NS	NS	.010	NS	NS	NS
25	SB	NS	NS	NS	.001	.019	NS	NS
26	IP	NS	NS	NS	NS	NS	NS	NS
27	IC	NS	NS	NS	NS	NS	NS	NS

Note: Degrees of Freedom = 1/46.

[a]See Table 1 for a complete description of the variables.

[b]$p > .05$

(21)," "general task praise (22)," and "specific behavior praise (25)." This result provides important counter-evidence to those teachers and administrators who believe that teachers who teach to criterion or require mastery performance from their students tend to focus on negative aspects of student performance--thus making the instructional setting aversive to the students. Quite to the contrary, the results of the present study suggest that the experimental (criterion teaching) teachers praised more often (with general and specific words of praise) than control (noncriterion teaching) teachers.

Analysis of Student Achievement

The second major question was whether there were significant differences in achievement on the students' posttest scores after they were adjusted statistically for the differences on the pretest score. The technique of multivariate analysis of covariance and discriminant analysis served to answer this question.

A two-way MANCOVA design was used. Factor A had two levels (experimental and control) and Factor B also had two levels (high and low implementors on correction procedures and criterion teaching). The dependent variables were (1) the Continuous Test for Language One (CTL) and (2) the APELL Test. Together these tests reflected "classroom and instructional language competency." The covariate used in the analysis was the Pre-Continuous Test for Language One (PCTL). This measure was administered one week prior to the treatment. Mean scores for the raw score and adjusted dependent achievement measures and covariate are presented in Table 5.

In addition to the usual multivariate normality assumptions, the conditions which the data must satisfy if the results of the analysis of covariance are to have suitable generality are (1) the regression plane of the posttest score on the pretest scores in the experimental group must be parallel to that in the control group, and (2) residual variation about the regression planes for the two groups

TABLE 5

Unadjusted and Adjusted Mean Pretest and Posttest Scores
Used in the Multivariate Analysis of Covariance

			High Implementors		Low Implementors		All Implementors
PCTL	Experimental Group		20.916		16.449		18.585
	Control Group		20.619		21.456		21.022
CTL	Experimental Group		27.897		21.820		24.726
		Adj:	26.574	Adj:	25.670	Adj:	26.122
	Control Group		24.750		24.908		24.826
		Adj:	23.891	Adj:	22.988	Adj:	23.444
APELL	Experimental Group		22.808		21.488		22.119
		Adj:	22.478	Adj:	22.328	Adj:	22.400
	Control Group		21.680		21.845		21.759
		Adj:	21.552	Adj:	21.402	Adj:	21.480

Note: Standard deviations for the adjusted CTL and APELL are 4.708 and 1.318, respectively.

must be homogeneous. Tests of these hypotheses indicated that the assumptions were tenable ($p > .50$).

The results of the multivariate analysis of covariance for the main effect due to treatment (retraining or no retraining) are presented in Table 6. The significant F-test ($p < .05$) leads us to the conclusion that there was information in the language measures (CTL and APELL) that was not in the pretest measure (PCTL) regarding differences between the retrained teachers (experimental group) and the nonretrained teachers (control group). The correlations between the adjusted criterion variables and the discriminant function (factor structure) indicate how much each dependent variable is contributing to the discrimination among the levels of the treatment factor. Since the APELL Test has the highest correlation (.94) with the discriminant function, it appears that the discrimination among the groups is--to a somewhat greater extent--due to differences in scores on the APELL (the program-general test) rather than to the CTL (the program-specific test). As can be seen from the mean discriminant scores of the two treatment groups on the discriminant function, the experimental group received the higher mean discriminant score (26.088).

The results of the multivariate analysis of covariance for the main effects due to implementation indicated that the effect was not significant (F-ratio for the multivariate test of equality of mean vectors = 0.172, $p > .84$). This means that the treatment was equally effective for the high and low implementors. Likewise, the interaction effect was not significant ($F = 0.062$, $p > .94$).

TABLE 6

Results of Multivariate Analysis of Covariance
for Main Effects Due to Treatment

	F-Ratio for Multivariate Test of Equality of Mean Vectors
df hypothesis	2
df error	44
F	3.309
P <	.046

	Standardized Discriminant Function Coefficients
CTL	.405
APELL	.742

	Factor Structure for the Discriminant Function
CTL	.76
APELL	.94

	Mean Discriminant Scores
Experimental Group	26.088
Control Group	24.774

CONCLUSIONS

Summary of Results

One "pass" through the descriptive-correlational-experimental-loop research paradigm has been completed with the Distar program. An observational instrument was developed which reflected the specific teacher behaviors that the program developers believed to be important for student cognitive gain. In two small studies (Siegel and Rosenshine, 1973) it was determined that teacher behaviors that were considered important for successful program implementation (following the format, appropriate correction procedures, teaching a format to criterion, requiring unison responding to signals) were significantly related to student achievement. Furthermore, the present study provided experimental support for specific correction procedures and criterion teaching--two categories of behaviors which were characteristic of the Distar Instructional System. A randomly selected group of high and low implementing teachers were retrained in techniques of correcting students' mistakes according to a prespecified paradigm and recycling through an instructional task until all of the children in the group respond without error. As a result of retraining, the experimental teachers performed at a significantly higher level of implementation than the control teachers. In addition there were significant differences (favoring the experimental group) in language achievement on the students' posttest scores after they were statistically adjusted for the differences on the pretest scores. Thus, significant changes in teacher behavior (and particularly along the dimension of criterion teaching--the behavior of repeating the entire task after a mistake has been corrected) apparently caused significant changes in student achievement (on both a program-specific and program-general measure).

However, even though the differences in achievement were significant, the differences were not as great as might be expected from the substantial modifications

in teacher behavior. Large changes in teacher implementation did not cause large changes in student achievement. One could therefore reasonably ask whether it is worth the effort (time, money, inconvenience) to retrain teachers with the result of small (although statistically significant) gains in student achievement. Unless one is to be misled, the answer must go beyond the normative or comparative results of training. For example, the experimental group's mean criterion teaching (ET) score was 2.4 standard deviation units above the control group's mean score for the same variable. Nevertheless, the experimental group repeated the task from the beginning, after a mistake was appropriately or inappropriately corrected, only 20 percent of the time; the control group recycled to the beginning of the task following 2 percent of the corrected mistakes. In an absolute sense the experimental group was teaching to criterion at a low-moderate level of implementation, at best. Statistically significant differences which account for a large percentage of the variance may not be educationally significant. Justification of retraining, therefore, should be viewed in terms of absolute standards of performance (as well as normative standards of performance).

A General Research Strategy

The descriptive-correlational-experimental-loop paradigm is perhaps most readily understood within the context of a behaviorally oriented or structured curriculum program such as the Distar Instructional System. Yet it is proposed that this research strategy would be applicable to less structured and more "open" curricula as well.

Table 7 illustrates this application and suggests various program specific (important) process variables and possible outcome variables for three early childhood programs.[8/] The program implementation (process) variables and outcome

8/ Although the description of variables for each program was inferred from the literature, program authors for each curriculum may disagree with the wording or emphasis in this table. However, the purpose here is to illustrate the versatility of the descriptive-correlational-experimental-loop research paradigm and not necessarily to give an accurate description of process and outcome variables for specific programs.

TABLE 7

Examples of Process and Outcome Variables for Three Early Childhood Education Programs

Program	Example of Process Variables	Example of Outcome Variables
The Engelmann-Becker Program (Engelmann and Becker, sponsors)	Following the format for group and individual activities; Correcting all mistakes according to a pre-specified paradigm; Teaching all formats to criterion (mastery); Requiring unison responses to signals.	Reading, language, and arithmetic achievement measures--criterion-referenced and norm-referenced.
The Cognitively Oriented Approach (Weikart, sponsor)	Arranging the environment so that the child can actively interact with it; Allowing the child to establish the learning pace; Arranging the curriculum so that it progresses from the simple to the complex--from the concrete to the abstract.	Rating scales to measure effects on socioeconomic development; Measures of language development; IQ; and Academic achievement measures.
EDC Program (Armington, sponsor)	Teacher flexibility; Ability to pose questions; Helping child to verbalize and to formulate his ideas and thoughts; Emotional climate established by the teacher; Arranging the environment for corrective feedback; Allowing the child to engage and disengage from a task at the child's will.	Measures to assess student initiative and self-direction, curiosity, and imagination, openness to change, self-respect and respect for others; Concentration, listening, and remembering.

variables are primarily determined by the curriculum developers and may or may not be stated explicitly. Furthermore, the behaviors which are emphasized during preservice and in-service teacher training may vary from what is expressed in the writings (journal articles, books, teacher guides, etc.) of the curriculum designers. Nevertheless, the researcher and program developer should ultimately concur on the implementation variables and on the procedures and instruments used for collecting teacher and pupil data. If this were not the case, then the program developer could argue--justifiably--that the researcher's study did not test the program's implementation variables. It would be rather difficult for a researcher to justify the inclusion of a particular "program-specific (important)" variable on a program-specific observation instrument when this claim is denied by the program authors. A compromise solution, however, is possible. The observation instrument could reflect three types of variables: those varibles which the curriculum developers and researchers hypothesize to be important for the success of the program, those variables which only the curriculum developers hypothesize to be important, and those variables which only the researchers hypothesize to be important. Although this would not obviate the above situation, the research would reflect each group's biases. This procedure could be expanded to include implementation variables which are specific or important to other curricula, as well as variables (especially outcome variables) which are important to different groups (parents, educators, legislators, students, etc.). For example, Armington would be concerned with measures of curiosity and imagination for children in the EDC Program whereas many parents may be concerned with measures of reading and arithmetic achievement. Both sets of variables may be collected--not only for the EDC Program but for other programs as well.

The research paradigm, furthermore, does not specify the manner in which implementation and outcome variables are to be collected. Perhaps observation

instruments would be used--or possibly student ratings, teacher or parent questionnaires, video tapings, audiotapings, surveys, or a series of "unobtrusive" measures. In fact, certain programs will typically value one form of data collection over another. For example, the Engelmann-Becker program would be satisfied with norm-referenced and criterion-referenced tests of achievement. The EDC Program, on the other hand, would probably value more indirect measures of student behavior and attitude. Again, many types of measures could be used for each program if more generalizable relationships are of interest.

Some Implications for Teacher Education

A major concern, previously implied in the introduction to this study, is whether or not generalized teaching behaviors are of importance to the implementation of curriculum programs.[9/] It was suggested that teacher behaviors which are specific to a program may be more important in affecting student outcomes than teacher behaviors which apply to a wide range of programs. Indeed, no teacher teaches the curriculum "first-grade reading;" she teaches the SWRL Reading Program, the Bank Street Readers, IPI, Addison-Wesley, Distar Reading, McGraw-Hill, or SRA Reading Labs, for example. Teacher behaviors which are critical to the success of one program may not be very important to the success of another--or may even be detrimental.

Granted, there are similarities between programs. For example, a teacher surely would not have to learn a new set of teaching skills when she teaches Distar Reading after having taught Distar Language. Likewise, there are certain similarities between programs oriented toward cognitive growth, between curricula oriented toward behavior modification, and so on. However, it is hypothesized that as the classification becomes more encompassing--goes beyond a specific curriculum program to include programs of a general type and ultimately, models of instruction--the probability decreases that any teacher behavior that applies

9/ Curriculum program is here defined in the broadest sense--ranging from a published set of materials to a philosophy or model of instruction.

to all elements (programs) of the larger set will be a powerful variable (that is, will account for a large percentage of the variance in the outcome measures) for all elements of one or more subsets.[10/]

Traditionally, researchers have sought those variables which cut across all possible instructional settings. These variables would apply at the same time to all programs in a discovery learning model, a behavior control model, and other models as well. This overindulgence in parsimony may well cause the bypassing of those specific variables which contribute most to significant outcomes of a particular program.

An alternative approach to research in teacher education is proposed. Choose several programs from a prespecified group of instructional programs. The set may be defined in broad or narrow terms (for example, the set of first-grade reading programs, the set of objectives-based reading programs, the set of linguistic reading curricula for disadvantaged children, the set of inquiry-type science programs for primary grade students, the set of Piagetian Follow-Through programs, etc.). Apply the descriptive-correlational-experimental-loop research paradigm to individual programs in the set. The results of such a research program would reveal those variables--for each program--which are critical to affecting student gain in outcomes of interest. Moreover, the results would reveal those dimensions which apply to all programs in the set.

The implications for teacher education of this inductive approach are clear. Teacher training programs which emphasize general strategies for teaching may be providing information for the school teacher which is less than useful. Rather, would-be teachers could profit more from learning to teach a sample of program types and the behaviors crucial to each. For example, a student teacher interested in early childhood education could learn to implement three or four curriculum

10/ No matter what the criteria chosen for the grouping and classification of programs--age, grade level, subject matter, model of teaching, psychological orientation, etc.--it is hypothesized that this holds true.

programs which sample a wide range of instructional strategies: say, Gilkeson and Zimiles' Bank Street Program, Bushell's Behavior Analysis Program, Gordon's Florida Project, and Weikart's Cognitively Oriented Approach. Or rather, if a more specialized training were desired, the set would include only programs of a certain type: for example, Engelmann's Distar Program; Bushell's Behavior Analysis Program, and Resnick's Primary Education Project. Thus, teacher behaviors which are specific or important to a particular program may be learned as well as those implementation behaviors or instructional activities which are generally important for the success of several programs.

LIST OF REFERENCES

Baker, E. L. Relationship between learner achievement and instructional principles stressed during teacher preparation. Journal of Educational Research, 1969, 63, 99-102.

Cronbach, L. J., and Furby, L. How we should measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Edcodyne Corporation, APELL test. Orange, Calif.: Edcodyne Corporation, 1971.

Engelmann, S. Conceptual learning. San Rafael, Calif.: Dimensions Publishing Co., 1969. (a)

Engelmann, S. Preventing failure in the primary grades. Chicago: Science Research Associates, 1969. (b)

Engelmann, S., and Bruner, E. Distar reading I. Chicago: Science Research Associates, 1969.

Engelmann, S., and Bruner, E. Distar reading II. Chicago: Science Research Associates, 1970.

Engelmann, S., and Carnine, D. Distar arithmetic I. Chicago: Science Research Associates, 1969.

Engelmann, S., and Carnine, D. Distar arithmetic II. Chicago: Science Research Associates, 1970.

Engelmann, S., and Carnine, D. Distar arithmetic III. Chicago: Science Research Associates, 1972.

Engelmann, S., Osborn, J., and Engelmann, T. Distar language I. Chicago: Science Research Associates, 1969.

Engelmann, S., and Osborn, J. Distar language II. Chicago: Science Research Associates, 1970.

Engelmann, S., and Osborn, J. Distar language III. Chicago: Science Research Associates, 1972.

Engelmann, S., and Stern, S. Distar reading III. Chicago: Science Research Associates, 1972.

Engelmann-Becker Corporation. Continuous test for language one. Eugene, Oregon: Engelmann-Becker Corporation, 1971.

Flanders, N. A. Analyzing teaching behavior. Reading, Mass.: Addison-Wesley, 1970.

Gage, N. L. Teaching methods. In R. L. Ebel (Ed.), Encyclopedia of educational research (4th ed.). London: Macmillan, 1969. Pp. 1446-1458.

Gallagher, J. J. Teacher variation in concept presentation of the BSCS curriculum program. Urbana: Institute for Research on Exceptional Children, University of Illinois, 1966. (ERIC: ED 023 206)

Gallagher, J. J. Analyses of teacher classroom strategies associated with student cognitive and affective performance. Cooperative Research Project No. 3325. Urbana: University of Illinois, 1968. (ERIC: ED 021 808)

Katz, L. G. Children and teachers in two types of head start classes. Young Children, 1968, 24, 342-349.

Kochendorfer, L. H. The development of a student checklist to determine classroom teaching practices in high school biology; Classroom practices of high school biology teachers using different curriculum materials; and Use of the biology classroom activity checklist in identifying specific classroom practices of individual teachers and students. In A. E. Lee (Ed.), Research and curriculum development in science education; I. The new programs in high school biology. University of Texas at Austin, 1967, 71-78; 79-84; 85-89.

LaShier, W. S., Jr. The use of interaction analyses in BSCS laboratory block classrooms. Journal of Teacher Education, 1967, 18, 439-446.

Lindvall, C. M., and Cox, R. C. The IPI evaluation program: AERA evaluation monograph no. 5. Chicago: Rand McNally, 1970.

Maccoby, E. E., and Zellner, M. Experiments in primary education; Aspects of project Follow-Through. New York: Harcourt Brace Jovanovich, Inc., 1970.

Medley, D. M., and Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand McNally, 1963.

Niedermeyer, F. C., and Dalrymple, M. C. The development and evaluation of an exportable teacher training package for a kindergarten reading curriculum. Paper presented to the meeting of the American Educational Research Association, February 1970.

Olivero, J. L. Developing the oral language program. Albuquerque, New Mexico: Southwestern Cooperative Educational Laboratory, undated.

Rosenshine, B. Evaluation of classroom instruction. Review of Educational Research, 1970, 40, 279-301.

Rosenshine, B. New directions for research on teaching. Paper presented at the conference on How Teachers Make a Difference, sponsored by the Bureau of Education Personnel Development, U. S. Office of Education, Washington, D. C., April 1971.

Rosenshine, B., and Furst, N. Research on teacher performance criteria. In B. O. Smit (Ed.), Research on teacher education: A symposium. Englewood Cliffs, N. J.: Prentice-Hall, 1971. Pp. 37-72.

Rosenshine, B., and Furst, N. The use of direct observation to study teaching. In R. M. W. Travers, Second Handbook of Research on Teaching. Chicago: Rand McNally and Co., 1973. Pp. 122-183.

Science Research Associates. Summary of twenty-one case studies on the effectiveness of the Distar Instructional System. Chicago: Science Research Associates, 1971. (a)

Science Research Associates. Distar participant's manual: Level I. Chicago: Science Research Associates, 1971. (b)

Siegel, M. A. The relationship between teacher behaviors stressed during training and student outcomes within a curriculum materials package. Unpublished Master's thesis, University of Illinois, 1972.

Siegel, M. A. Correcting basic mistakes: A training program for Distar Language I teachers. Chicago: Science Research Associates, 1973. (a)

Siegel, M. A. An experimental investigation of teacher behavior and student achievement in the Distar Instructional System. Unpublished dissertation, University of Illinois, 1973. (b)

Siegel, M. A., and Rosenshine, B. Teacher behavior and student achievement in the Bereiter-Engelmann Follow-Through program. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, February 1973.

Siegel, M. A., Rosenshine, B., Marzano, W., and Walsh, P. A multivariate comparison of two program-specific observation instruments. Unpublished manuscript, Bureau of Educational Research, University of Illinois, 1972.

Simon, A., and Boyer, E. G. (Eds.) Mirrors for behavior: An anthology of classroom observation instruments. Vols. 1-14 and Summary and Supplementary Vols. A and B. Philadelphia: Research for better schools, 1967, 1970a, 1970b.

Soar, R. S. Advantages of multiple systems over a single system. Paper presented to the American Educational Research Association, 1971. Gainesville: Institute for Development of Human Resources, College of Education, University of Florida.

Soar, R. S., Soar, R. M., and Ragosta, M. The validation of an observation system for classroom management. Paper presented at the meeting of the American Educational Research Association, February 1971. Gainesville: University of Florida.

Tatsuoka, M. M. Nationwide evaluation and experimental design. Paper presented at the meeting of the American Educational Research Association, April 1972, University of Illinois.

Walberg, H. J. Predicting class learning: An approach to the class as a social system. American Educational Research Journal, 1969, 6, 529-542.

MAS:mjc