DOCUMENT RESUME

ED 201 610 SP 017 950

AUTHOR Coladarci, Theodore; Gage, N. L.
TITLE Minimal Teacher Training Based on Correlational Findings: Effects on Teaching and Achievement.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE Apr 81
GRANT NIE-G-79-0014
NOTE 101p.; Paper presented at the Annual Meeting of the American Educational Research Association (Los Angeles, CA, April, 1981). Contains some light print.

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS Change Strategies; *Educational Innovation; Elementary Secondary Education; Inservice Teacher Education; *Teacher Attitudes; *Teacher Behavior; Teaching Methods; *Training Methods; *Training Objectives

ABSTRACT

Four classroom-based experiments in which teachers were trained to use a direct instruction model were analyzed to compare the results of intensive and minimal training methods. The direct instruction model involves extensive coverage of content, student time allocated to instructional tasks, and teacher time allocated to the encouragement of students. The observed results of those studies indicated that teachers could receive minimal training in a teaching method and successfully implement it in the classroom with significant improvement in student achievement. A fifth study, designed to test these findings, failed to corroborate the positive results obtained previously. After a year-long study, experimental-group teachers with a minimum of training did not evidence markedly greater conformity to the training recommendations than the control-group teachers did. The classes of these two groups of teachers were not appreciably different in end-of-year student academic achievement. Training-related behavior among experimental-group teachers was not modified enough to effect appreciable changes in subsequent student achievement. These results, and the results of the previous four studies, are analyzed, and recommendations are made for further research. (JD)

Minimal Teacher Training Based on Correlational Findings: Effects on Teaching and Achievement

Theodore Coladarci

University of Montana

N.L. Gage

Stanford University

Paper presented at the Annual Meeting
of the American Educational Research Association
Los Angeles, 1981

The research reported here was supported by a grant from the National Institute of Education, U.S. Department of Health, Education, and Welfare (Grant No. NIE-G-79-0014). The opinions expressed do not necessarily reflect the position, policy or endorsement of the National Institute of Education. Warm thanks to Gayle Zachariasen and Kathy Corbally for typing this paper.

Introduction

Direct instruction represents a constellation of teacher behaviors and classroom characteristics--"a convergence of results" (Rosenshine & Berliner, 1978, p. 3)--that has been identified in the accumulation of process-product research. Further, the direct instruction model has been regarded as foremost in explaining growth in conventionally measured achievement, especially at the elementary-grade level (Berliner, 1979; Berliner & Rosenshine, 1977; Good, 1979b; Powell, 1978). Powell (1978; also see Rosenshine, 1979, p. 38) offered perhaps the most succinct presentation of several key components of this model:

> The coverage of content is extensive, time is allocated to academic tasks, and the time is not broken by frequent interruptions or changes of task. Students spend a good portion of the time allocated to instruction actually engaged in instructional tasks, and the teacher monitors and encourages task engagement on the part of the students. . . .The atmosphere in the classroom is one in which academic work is both recognized to be important and performed. (p. 29)

Experimental Research

To date, four classroom-based experiments have been conducted that incorporated the direct instruction model: Anderson, Evertson, and Brophy (1979); Crawford, Gage, Corno, Stayrook, Mitman, Schunk, Stallings, Baskin, Harvey, Austin, Cronin, and Newman (1978); Good and Grouws

(1979); and Stallings, Needels, and Stayrook (1979). In each case, findings from previous correlational studies of process-product relationships were assembled into clear, concise reading material for teachers. Further, random assignment was employed in assigning the classes or schools to experimental conditions.

The training programs developed by Anderson et al. (1979), Good and Grouws (1979), and Stallings et al. (1979) were based largely on process-product relationships reported in Brophy and Evertson (1979), Good and Grouws (1977), and Stallings, Corey, Fairweather, and Needels (1978), respectively. Anderson et al. (1979) acknowledged the additional influence of Blank (1973) and the Southwest Educational Development Laboratory (1973). The training program developed by Crawford et al. (1978) involved the comprehensive examination and synthesis of the results of four large-scale correlational studies (Brophy & Evertson, 1974; McDonald & Elias, 1976; Soar, 1973; Stallings & Kaskowitz, 1974).

Because they have been conducted in regular classrooms, rather than specially contrived settings, these four experiments rate high in ecological validity. They have the realism of being concerned with teaching that has gone on over an extended time of several months or, more typically, the entire school year, rather than a few hours, days, or weeks. The teachers in these experiments have been practicing teachers, rather than student teachers or teachers specially selected and employed for the research project. And, because the teaching recommendations manipulated in these experiments were derived from studies of naturally occurring teaching behaviors, it was known in advance that this manipulation would call for no esoteric behavior that was alien to regular classrooms.

Although for the most part relying on different data bases for the development of the training programs, these experiments have in common the theme of direct instruction. In addition to demonstrating positive change in training-related teaching practices of experimental-group teachers, each experiment resulted in greater gains in student achievement for experimental-group classes when compared to control-group classes. Each experiment is described here.

Anderson et al. (1979). This experiment was conducted in White, middle socioeconomic-status (SES), first-grade classes. Schools were randomly assigned to treatments, after stratifying on school size and SES. The dependent measure was the total reading score on the Metropolitan Achievement Tests; the total readiness score on the Metropolitan Readiness Tests served as a covariate.

Experimental-group teachers received a manual presenting an instructional model, which set forth 22 principles.

> The treatment was minimal in cost and time. In October, the researchers met with teachers in the treatment schools and described the purpose of the study. The teachers who agreed to participate read the manual describing the instructional model and met again with the experimenters to discuss it. There was no further training, and no attempts were made during the year to boost the treatment. (p. 195)

Observations were conducted in all control-group classes and in 10 of the 17 experimental-group classes. Between November and May of the school year, each of these classes was observed on 15 to 20 occasions--roughly once a week. A specially constructed observation instrument was used that allowed the investigators to "measure implementation of the principles in the instructional model as well as other aspects of first-grade reading instruction that might be important in assessing students' achievement" (p. 198).

Because not all experimental-group classes were observed, the question could be addressed: Did the presence of observers moderate the effect of the treatment? That is, did experimental-group teachers who were observed have classes with greater achievement gains than the classes of the unobserved experimental-group teachers?

A series of between-class regression equations were employed to assess treatment effects, covariate-treatment interactions, and observation effects (i.e., observed versus unobserved experimental-group teachers). There was a significant treatment effect ($p < .05$): After the dependent variable was regressed on the covariate, an additional 10% of the variance was accounted for by entering the treatment term. From similar analyses, it was found that there was neither a statistically significant covariate-treatment interaction nor a statistically significant observation effect. The former finding indicated that, by conventional standards, homogeneity of regressions could be assumed; the latter indicated that the presence of classroom observers did not moderate the treatment effect on student achievement.

Crawford et al. (1978). Here, the context was 33 middle-SES, third-grade classes. Volunteer teachers, after their classes were stratified on mean academic achievement, were randomly assigned to three experimental conditions: observation only ($N = 10$), minimal training plus observation ($N = 11$), or maximal training plus observation ($N = 12$). Minimally trained teachers simply were mailed at weekly intervals a series of five training packets, which embodied 22 principles (only coincidentally the same number of principles used in the experiment conducted by Anderson et al. [1979]). The maximally trained teachers, in addition to receiving the weekly packets, attended weekly meetings in the five-week period during which the training packets were delivered. These meetings were devoted to review and discussion, along with videotape viewing and role playing.

As was noted above, all classes in the three groups were observed. The observations were performed for a total of approximately 16 full school days for each of the 33 teachers before, during, and after the training.

Ostensibly, the two modes of training delivery differed considerably with respect to the teacher engagement with the training material. Nevertheless, these two training conditions had equivalent effects on class achievement on a vocabulary posttest. And together, they were .69 of a standard deviation (SD) above the mean of the control-group classes ($p < .15$), although there was no comparable effect on a reading comprehension posttest. Interestingly, minimally trained teachers were found to implement more of the training recommendations than the maximally trained teachers. This difference, however, may be partly artifactual. The minimally trained teachers were initially higher than the maximally trained teachers on a measure of verbal fluency and a measure of of structuredness, both of which correlated positively with implementation. However, a difference in implementation--albeit a small one--remained after adjusting for these initial differences.

Good and Grouws (1979). Forty lower-SES, fourth-grade classes served as the context for this experiment. The dependent variable was performance on the mathematics subtest of a standardized achievement test administered in mid-December; the same test administered in September served as a covariate. The training procedures were similar to those reported by Anderson et al. (1979). An introductory meeting was held in September for all 40 volunteer teachers and their principals. At this meeting, the general nature of the study was outlined and, subsequently, schools were randomly assigned to treatments. The researchers then described the instructional model to the 21 experimental-group teachers for approximately 90 minutes, and the 45-page manual was distributed. Two weeks after treatment began, an additional 90-minute meeting was held to answer questions about the program. Almost all of the teachers were observed on six occasions between October and the end of January.

A class-level analysis of variance on residualized gain scores indicated a treatment effect ($p < .01$) favoring the experimental-group classes. Good (1979b) later reported that the experimental-group classes still held an advantage at the end of the school year when the district carried out its regular testing--roughly three months after formal observations were completed. That a treatment effect was detected early in the school year is noteworthy and, indeed, encouraging. Further, the persistence of this effect for three months after classroom observations were discontinued might be regarded tentatively as evidence of the stability of the treatment effect.

The findings of Crawford et al. (1978) regarding the minimally versus maximally trained teachers, the absence of an observation effect reported by Anderson et al. (1979), and the results of Good and Grouws (1979) led Good (1979b) to the following conclusion:

> Although more research on implementation is needed, two tentative conclusions are warranted: (a) elaborate delivery systems may not be necessary for effectively training inservice teachers to perform specifically identified classroom behaviors, and (b) observation of teachers does not necessarily have to be a part of the inservice training. (p. 57)

Stallings et al. (1979). This experiment differed from the other three in two critical respects. First, the context was junior and senior high school classes, rather than elementary-grade classes. Second, and perhaps more important, the training was accomplished through comparatively intensive workshops. Despite these differences, the study nevertheless is relevant to the present discussion in that, like the experiments discussed above, direct-instruction findings from previous correlational process-product research were put to experimental test in regular classes.

Volunteer teachers of 22 junior and 24 senior high school classes were randomly assigned to a training or no-training condition. Students represented a broad range of both ethnicity and locale. Each class was observed for three consecutive days in the fall, winter, and spring.

Four two-hour workshops were held for the 22 experimental-group teachers after the fall observations had been completed. In addition to extensive discussion and role playing pertaining to "the direct approach to teaching" (p. 6.10), observation-based feedback and recommendations were provided for each trained teacher. An additional two-hour workshop was held after the winter observations had been completed. Finally, a teacher-requested meeting was held in April for the experimental-group teachers from all districts. The meeting, which lasted a full day, provided teachers the opportunity to exchange information.

The dependent variable was gain on the Comprehensive Tests of Basic Skills (CTBS) from the end of one year to the end of the next. (Complete CTBS data were available for the classes of 14 control-group teachers and 15 experimental-group teachers.) The authors reported a standardized mean-difference of .52 SD in favor of the experimental-group classes. When calculated as recommended by Hedges (1980)[1] the standardized mean-difference is .43 _SD_--still, an encouraging value. Although no other analyses bearing on treatment effects were reported, one can compute a _t_ ratio using the mean gains and standard deviations provided (Stallings et al., 1979, Table 31). The resulting value is 1.19 ($p < .15$).

A note on implementation. In these four experiments, any treatment effect on student achievement clearly is mediated by the extent to which the teachers implemented the various instructional programs. That is, treatment implementation is a necessary condition for subsequent treatment effects on student achievement. Regardless of how thoroughly a teacher may have read the furnished materials, such treatment effects cannot be expected where teaching processes have remained practically unaltered. As Charters and Jones (1979) pointed out in the context of program evaluation, "it is the _use_ of new instructional packages. . . .that constitutes an innovation, not the mere presence of the packages in the classroom" (p. 6, emphasis in original).

[1] Hedges (1980) argued that a standardized mean-difference, or "effect size," is best computed by subtracting the control-group mean from the experimental-group mean, dividing the difference by the pooled within-group standard deviation, and multiplying by a correction factor based on the degrees of freedom represented in the denominator.

Assume that a given sample of teachers is reasonably motivated to conform to teaching recommendations--not an implausible assumption with volunteer teachers (e.g., Rosenthal & Rosnow, 1975). Then the question can be asked, What influences actual implementation? Is there inter-recommendation variation with respect to implementation?

The results on implementation from these four experiments lend support to what Doyle and Ponder (1977) called "the practicality ethic." In their discussion of teachers' reactions to change proposals regarding instructional practices, Doyle and Ponder (1977) held that "the study of the practicality ethic is the study of perceived attributes of messages and the way in which these perceptions determine the extent to which teachers will attempt to modify classroom practices" (p. 2).

A judgment concerning this ethic is shaped by three criteria: instrumentality, or the extent to which the change proposal is stated clearly and with "procedural specifications"; congruence, or the extent to which the proposal "is congruent with perceptions of [the teachers'] own situations"; and cost, or "the ease with which a procedure can be implemented and the potential return" (pp. 7-8).

Anderson et al. (1979) reported that successfully implemented recommendations tended to be those that described specific skills and focused on familiar, though not necessarily relied upon, behaviors. Similarly, Good and Grouws (1979) found that behaviors involving specific requests were more successfully implemented. And Crawford et al. (1978) reported that the less successfully implemented recommendations tended to be more global and non-specific. These findings seem to reflect operation of the instrumentality criterion.

When Stallings et al. (1979) concluded that "it is difficult to get teachers to try something they have opinions against" (p. 7.4), it would seem that these authors were addressing the congruence criterion. Similarly, Ebmeier and Good (1979), having analyzed further the data from the study conducted by Good and Grouws (1979), reported that teachers who already "believed" in an instructional model like the one being introduced through the training were characterized by greater implementation. Anderson et al. (1979) found that successfully implemented behaviors "had a rationale based on other classroom processes or student outcomes that made sense to teachers" (p. 219). It could be argued that this rationale provided congruency for the teachers. And to the extent that such a rationale furnishes information regarding the potential return of implementation vis-à-vis student outcomes, the teachers perhaps are able to make a judgment concerning the cost criterion.

Summary. In contrast to correlational process-product research, the results of these four experiments allow tentative statements about causality, rather than mere association. This experimental research indicates that training teachers to adopt a direct-instruction approach to teaching can result in positive change, when compared to untrained teachers, in both training-related teacher behavior and student achievement. Further, the findings of Anderson et al. (1979), Crawford et al. (1978), and Good and Grouws (1979) have suggested that such change does not necessarily require extensive investment on the part of the researcher--that positive results can be obtained through a minimal intervention. Referring to these three experiments, Good (1979a) held that "these studies illustrate that teachers can be taught direct instructional principles in relatively simple training programs that lead to changes in teachers' classroom behavior and student achievement" (p. 9).

The Role of the Present Study

The present study constituted a minimal intervention. Unlike the four experiments discussed above, this intervention was minimal in that (a) the treatment consisted solely of mailing training materials to the experimental-group teachers and (b) only a limited number of brief classroom observations were conducted.

As will be discussed in greater detail below, the teacher training in the present study took the form of the "minimal training" condition in Crawford et al., (1978). The latter study, however, involved frequent and lengthy classroom observations. And although Anderson et al. (1979) explicitly addressed the question of classroom observations as a moderator of treatment effects on student achievement, part of the training in that study involved attending two meetings with project staff to discuss the project and training materials. Good and Grouws (1979) provided a total of ten hours of workshops during the training period and conducted extensive observations, as well.

Thus, although encouraging claims have been made concerning the feasibility of a minimal intervention (e.g., Good, 1979a, 1979b; Good and Grouws, 1979), such an intervention had not yet been undertaken. That is, no intervention had been minimal with respect to both the delivery of training and the conduct of classroom observations. The present study was such an intervention.

Sample

The initial sample comprised 33 volunteer teachers and their fourth-, fifth-, and sixth-grade students in a large, urban school district in the San Francisco Bay Area. Because of subsequent complications, the number of classes on which the achievement-data analyses were based was reduced to 28. There were 966 students in the 28 classes, 631 of whom had both pretest and posttest achievement scores. All analyses bearing on academic achievement were based on these 631 students.

Most of the students (85%) in the sample were either Black (65%) or Caucasian (20%). As for occupational status, roughly 67% of the parents had "skilled" occupations or lower, with nearly one in five being unemployed or receiving AFDC. Approximately 20% had occupations rated as "professional" or "white collar." (See Coladarci [1980] for description of the occupational-rating instrument.)

Instruments and Procedures

The instruments in this study were the teacher education packets, the classroom observation schedule, and the Comprehensive Tests of Basic Skills.

The Teacher Education Packets

The teacher education packets (TEP) were developed and used by Crawford et al. (1978) in their study. As mentioned above, the TEP contained recommendations for teaching that were based on the large-scale correlational studies conducted by Brophy and Evertson (1974), McDonald and Elias (1976), Soar (1973), and Stallings and Kaskowitz (1974). The thousands of process-product correlations presented in the technical reports of these studies were examined and considered as the basis for prescriptive statements (see Crawford et al., 1978, Vol. I, pp. 25-31). There were several requirements for a particular process-product correlation coefficient: (a) The product variable had to be reading achievement; (b) the correlation coefficient had to be statistically significant ($p < .05$); and (c) the process variable had to be operationally defined.

Information sheets were prepared for each process-product correlation satisfying these conditions (see Crawford et al., 1978, Vol. I, Appendix A). Each sheet reported the process variable's operational definition, mean, standard deviation, and metric, along with the process-product correlation coefficient and interpretation of this coefficient. The operational definition of a variable was important in assessing the comparability of its meaning across studies. And the mean and standard deviation of a variable were necessary in estimating the desirable level of the variable in practice. Where a variable correlated positively with reading achievement, the desirable level was set at one standard deviation above its mean. Conversely, the desirable level was set at one standard deviation below its mean for any variable that correlated negatively with reading achievement.

The interpretations of the 125 qualifying correlation coefficients provided the basis for three packets of teaching recommendations, each packet corresponding to a general area of teaching: behavior management and classroom discipline, instructional methods, and questioning and feedback. Table 1 presents the number, and source, of variables represented in each of these three categories. The contents of each packet will be briefly discussed here.

Behavicr management and classroom discipline. This packet is based on the findings that classes characterized by a general unruliness and a poorly articulated system of rules are also characterized by frequent nonengagement in academic activities and student difficulty in attending to academic tasks. Teachers are informed of ways to manage their classes, largely in the light of Kounin (1970; also see Brophy & Putnam, 1979).

The packet cautions teachers against disciplinary errors that prolong or compound the problem--specifically,

Table 1

The Number and Source of Variables for Each of the Teacher Education Packets' Categories

Category	Number of Variables	Study
Behavior Management and Classroom Discipline	6	Brophy and Evertson
	3	Stallings and Kaskowitz
	2	McDonald and Elias
	2	Soar
Instructional Methods	16	Stallings and Kaskowitz
	11	McDonald and Elias
	6	Soar
	6	Brophy and Evertson
Questioning and Feedback Strategies	50	Brophy and Evertson
	17	Stallings and Kaskowitz
	4	McDonald and Elias
	2	Soar

Source: Crawford & Stallings (1978)

the disciplinary errors regarding "timing" and "target." Further, this packet encourages teachers to develop a "system of rules," which lets students know--without always having to consult the teacher--what they can and should do during a given period. Finally, to curb misbehavior as well as to identify and respond to students in need of assistance, teachers are encouraged to monitor activities when students are engaged in seatwork.

In short, teachers are encouraged to develop what Kounin (1970) called "withitness"; through monitoring and vigilance, teachers develop a keen awareness of their class--who is and is not academically engaged, who needs assistance, who is misbehaving (indeed, who is _about_ to misbehave), and so on.

Instructional methods. This packet highlights the importance of large-group instruction, frequent use of question-and-answer sessions, and use of visual aids and phonics exercises in reading activities. Additionally, with seatwork assignments, this packet informs teachers of the importance of assigning work of appropriate difficulty, using textbooks and workbooks (rather than games, toys, and machines), and minimizing through proactive planning the amount of time devoted to organizing and giving directions.

Questioning and feedback strategies. This packet pertains to the manner in which the teacher selects students to respond to questions, the difficulty level of the questions asked, and the provision of feedback subsequent to the student's response. A summary listing of TEP recommendations is presented in Table 2.

An introductory packet briefly discussed the TEP's rationale and provided a classroom vignette illustrating a teacher whose practices largely conformed to the TEP recommendations. A fifth packet reviewed and summarized the preceding packets. A sixth packet presented an additional classroom vignette, illustrating teaching practices that were both consistent and inconsistent with the TEP recommendations. To measure knowledge obtained, teachers were asked to respond to 24 different scenarios in terms of the extent to which the particular sequence of events conformed to the TEP recommendations. The six packets individually were mailed to experimental-group teachers in December and January.

Finally, the teachers received three refresher sheets corresponding to the second, third, and fourth packets, respectively. These sheets were intended to provide a succinct and accessible review of the contents of the three packets and were mailed, one per week, over a three-week period beginning in mid-February.

Table 2

Summary of TEP Recommendations

Behavior Management and Classroom Discipline

(1) Teachers should have a system of rules that allows pupils to attend to their personal and procedural needs without having to check with the teacher.

(2) Teachers should prevent misbehaviors from continuing long enough to increase in severity or spread to and affect other children.

(3) Teachers should attempt to direct disciplinary action accurately--that is, at the child who is the primary cause of a disruption.

(4) Teachers should keep "overreactions" to a minimum (even though overreactions are probably effective in stopping the misbehavior).

(5) Teachers (and aides, if present) should move around the room a lot, monitor pupils' seatwork, and communicate to the pupils an awareness of their behavior, while also attending to their academic needs.

Instructional Methods

(6) When pupils work independently, teachers should insure that the assignments are interesting and worthwhile and still easy enough to be completed by each pupil working without teacher direction.

(7) Teachers should keep to a minimum such activities as giving directions and organizing the class for instruction. They can do this by writing the daily schedule on the board, insuring that pupils know where to go and what to do, etc.

(8) Teachers should spend at least one-third to one-half of their time teaching larger groups of pupils (more than eight children). When they do teach smaller groups or individuals, they should take steps to make sure that the other pupils in the class have work to which they can attend.

(9) Teachers should make abundant use of textbooks, workbooks, and other pencil-and-paper activities. These have been found to be associated with higher pupil achievement. But the use of games, toys, and machines has not been found to be associated with higher pupil achievement.

(10) Teachers should provide visual demonstrations and phonics exercises in conjunction with reading activities.

(11) Teachers should frequently conduct public (i.e., addressed to a larger group or the whole class) question-and-answer sessions concerned with the academic subject matter at hand. With less academically oriented pupils, teachers may find it helpful to initiate some brief private discussions concerning personal matters.

Specific Methods for Asking Questions and Providing Feedback

(12) In selecting pupils to respond to questions, teachers should use the technique of calling on a child by name before asking the question, as a means of insuring that all pupils are given an equal number of opportunities to answer questions.

(13) Teachers should avoid calling on volunteers more than 10 or 15 percent of the time during question-and-answer sessions. It is also advisable to discourage pupil "call outs" to questions asked of other children (except possibly from less academically oriented children who may benefit from this type of activity).

(14) In the interest of promoting smooth, task-oriented discussions, teachers should not encourage large numbers of pupil-initiated questions and comments. It is also important for teachers to listen carefully to pupils' opinions and, if a disagreement is called for, to express such disagreement to the child.

(15) With less academically oriented pupils, teachers should ask easier questions--questions that can almost always be answered correctly. When questioning more academically oriented pupils, teachers should ask more difficult questions--questions that are answered incorrectly about one fourth of the time.

(16) Teachers should give praise only for really outstanding work; also, praise is likely to be more effective with less academically oriented pupils. Mild criticism is effective in communicating higher expectations ("you can do better") to more academically oriented pupils.

(17) With less academically oriented pupils, teachers should always aim at getting the child to give some kind of response to a question. Rephrasing, giving clues, or asking a new question can be useful techniques for bringing forth some answer from a previously silent pupil or one who says "I don't know" or answers incorrectly.

(18) With more academically oriented pupils who generally become actively involved in discussions, teachers should concentrate on getting the correct response. Therefore they should redirect questions to other pupils if the more academically oriented pupil answers incorrectly.

(19) Teachers should give the answer (to both more and less academically oriented pupils) if the response is at least partly correct. Teachers should not simply repeat the same question if any pupil (either more or less academically oriented) answers incorrectly, says "I don't know," or remains silent.

(20) With more academically oriented pupils, teachers should give brief feedback extensively (80% or more of the time) during private, one-to-one discussions. When dealing with less academically oriented pupils, teachers should use approximately equal amounts of brief and longer feedback, tailoring the duration of their reactions to the needs of the individual child in each situation.

(21) During reading-group instruction, teachers should give a maximal amount of brief feedback, and provide fast-paced activities of the "drill" type.

(22) During public question-and-answer sessions, teachers should occasionally give a detailed, "why" explanation in answer to a question.

Additional quizzes, general questions, and rating forms were included in the TEP, as well. Covering the main points and recommendations in the respective packets, the quizzes had either a multiple-choice or sentence-completion format. Answer keys were provided for all but one of the quizzes, the exception being a final, comprehensive quiz. General questions were structured in an open-questioned format and covered the teachers' TEP-related opinions, attitudes, and practices. The rating forms called upon the teachers to estimate the frequencies in their classes of various activities and events that were discussed in the packets. Although not included here, completed analyses of these data are discussed in Gage and Coladarci (1980) and Mohlman, Coladarci, and Gage (1980).

Classroom Observation Schedule

This instrument, adapted from a measure used by Crawford et al. (1978), is a 4-page record of observer judgments and estimates on both low-inference and high-inference variables. For example, it contains items pertaining to the number of times the "teacher teaches groups of 8 or more pupils at a time" and "teacher calls pupil by name before asking question" (low inference), as well as items regarding the degree to which there exists "effective use of system of rules by teacher" and "communication of awareness to pupils by teacher" (high inference).

Each of the 26 items in the observation record reflects components of the TEP (see Table 3). The alternatives for each item in the observation schedule were scored so the highest value represented the highest degree of conformity to the recommendations for that item and the lowest value represented the lowest degree of conformity. Thus, the observation records yielded a rough estimate of the extent to which teaching practices--of both experimental- and control-group teachers, before and after training--reflected the TEP recommendations.

Each teacher was observed on four two-hour occasions--twice in the fall and twice in the spring. Thus, in the present design, teachers and occasions were crossed. Each observer, however, did not observe all teachers; hence, observers were nested within teachers. Further, because not all observers observed on each of the four occasions, observers similarly were nested within occasions. The observation design is represented by the schematic in Figure 1.

The Comprehensive Tests of Basic Skills

The Comprehensive Tests of Basic Skills (CTBS), a nationally standardized test of academic achievement, served as the dependent measure. In 1976, the school district's committee on test selection chose the CTBS for regular use in the district. Of five standardized

Table 3

Summary Listing of Categories of the Classroom Observation Record

1. Target Errors

2. Timing Errors

3. Overreactions by teacher

4. Effective use of system of rules by teacher

5. Teacher awareness of behavior problems

6. Communication of awareness to pupils by teacher

7. a) Teacher wrote daily schedule on chalkboard

   b) Teacher made use of written schedule

8. Teacher calls pupil by name before asking question

9. Teacher accepts call-outs during question-answer sessions

10. Teacher encourages pupil-initiated questions and comments

11. Length of teacher's public feedback in reading group

12. Teacher monitors pupils' individual and small-group work

13. Teacher teaches groups of 8 or more pupils at a time

14. Teacher uses visual demonstrations in teaching reading or other academic subjects

15. Teacher uses phonics exercises in teaching reading

16. Students use textbooks, workbooks, paper-pencil activities, etc.

17. Teacher's amount of direction-giving and organizing

18. Positioning of reading or math group: pupils' backs toward the rest of class

19. Total time in private, personal matters with one child at a time

20. a) Number of times the pupils' academic answers were partly correct

    b) When there is a partly correct answer, estimate the number of times the teacher went ahead and gave the right answer

21. Teacher's task orientation towards the defined task

22. Teacher's positive affect toward one or more children

23. Teacher's negative affect toward one or more children

24. Attention of students

25. Noise level of classroom

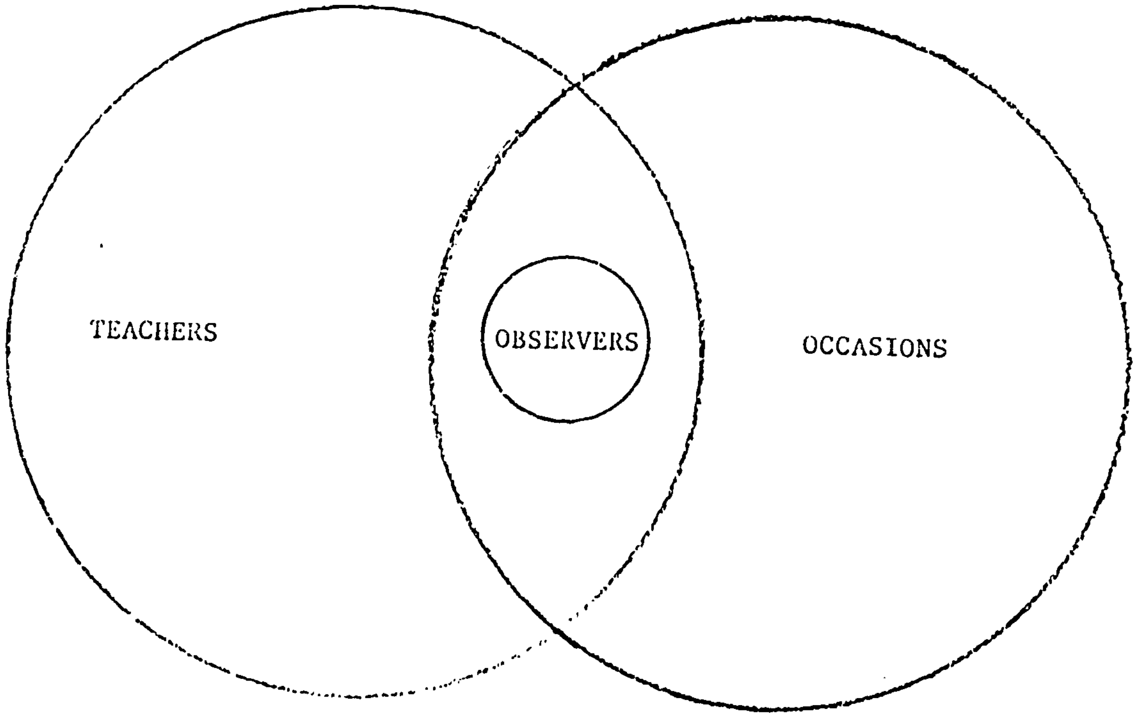

Figure 1. Schematic of Observation Design.

achievement tests that were evaluated for district use, this test was judged as providing the best match between test content and local curricula.

Test scores from the spring 1978 and spring 1979 administrations were recorded from computer printouts provided by the school district's research department. The 1978 data and the 1979 data served as the pretest and posttest, respectively. Each student's test scores were combined to yield a reading total score and a mathematics total score which, in turn, were combined as a total score.

Assigning Teachers to Experimental Conditions

Teachers were assigned to the control group or experimental group in the following manner: (a) Grade-equivalent means on the CTBS pretest were computed for each class; (b) scatterplots[2] were made, displaying the joint distribution of CTBS means and fall conformity-to-recommendations means; (c) teachers were paired in each scatterplot according to proximity; and (e) at a toss of a coin, one teacher in the pair was assigned to the experimental group, and the other teacher to the control group.

[2] These scatterplots were made separately for fourth-grade classes, fifth-grade classes, and fourth-fifth combination classes. The decision to add sixth-grade students to the sample was made after teachers were assigned to the experimental conditions.

Experimental-group teachers were asked to become familiar with the contents of the TEP and, further, to follow the various recommendations in their teaching. Additionally, these teachers were asked to complete the quizzes, open-ended questions, and rating forms associated with each of the packets, returning their responses by mail. Project staff did not meet with any teachers to discuss the training materials or to encourage implementation.

Results: Implementation of Training

The statistical analyses focused on three main questions. The first question concerns implementation of the training recommendations: Did the intervention appreciably alter the training-related teaching practices of experimental-group teachers? The second question concerns treatment effects on student achievement: Did the intervention produce significant increments in academic achievement for the students in the experimental-group classes? The third question addresses the study as a correlational, or process-product, one: Irrespective of experimental condition, was there a positive relationship between teachers' conformity-to-recommendations and student achievement? This section presents the analyses and results associated with the question of treatment implementation. The second and third questions are covered in the following sections.

As was noted above, each of the four observations was coded to yield a total score representing a teacher's general conformity-to-recommendations (CTR). The two fall CTRs were averaged for each teacher, as were the two spring CTRs. It was the CTR total, rather than the CTR item, that was emphasized in the analyses of the observation data. (For analyses of item data, see Coladarci [1980].) Analyses that employed the CTR total were considered more meaningful for several reasons. First, this study, as will be recalled, represented a minimal intervention. Here, one of the defining characteristics of such an intervention was the limited number of brief classroom observations. To propose a comprehensive item-level analysis of the observation data seems inconsistent with the stated purpose, as well as with the design, of the study. Second, as Crawford and Stallings (1978) pointed out, the most compelling and defensible analysis is one of the program as a whole (i.e., total CTR) simply because the discrete teaching recommendations were not independently manipulated. While analyses that focus on the discrete teacher behaviors may prove intriguing, the inevitable intercorrelation among these behaviors renders problematic any clear and meaningful interpretation. And, third, as the sum of _n_ positively correlated items has greater reliability than each item considered individually, larger and hence more meaningful differences are more likely to be found with the CTR total.

The first task was to explore the reliability-stability of the CTR. Then, group differences on CTR were examined to assess treatment implementation.

Reliability-Stability of CTR

The correlation between the two fall observations, the two spring observations, or the mean fall and mean spring observation represents at once (a) the reliability of the observers and (b) the stability of the teachers' behavior. That is, (a) what would be the agreement between the observations of two individuals if they had observed the same teacher on the same occasion? And (b) how similar would a teacher's observed behavior be over two occasions if observed by the same individual? With the observation design of the present study (see Figure 1), these two sources of variance are completely confounded.

Given the importance of establishing the reliability-stability of the CTR and, further, because it is somewhat independent of the question of treatment effects on student achievement, analyses assessing CTR reliability-stability were conducted on the original sample of 33 teachers as well as on the final sample of 28 teachers (i.e., for whom adequate CTBS data were available.)

Table 4 presents, for the full sample, the means, standard deviations, and intercorrelations for total CTR corresponding to each of the four observation occasions. The two fall CTRs are moderately correlated (_r_ = .27), as are the two spring CTRs (_r_ = .27), while the remaining correlations are much smaller and changing in sign.

The reliability of the sum of the two fall CTRs and of the two spring CTRs can be estimated by applying the Spearman-Brown formula. The estimated reliability of the fall sum (hereafter, "fall total CTR") becomes .43, which is the same estimate for the spring sum (hereafter, "spring total CTR"). These estimates are equivalent to generalizability coefficients (Cronbach, Gleser, Nanda, & Rajaratnam, 1972). As such, they represent the ratio of between-teacher variance to the total observed-score variance--the latter comprising both between-teacher variance and the nested combination of variance attributable to interactions involving teachers, occasions, and observers (see Figure 1).

The fall total CTR and spring total CTR are virtually uncorrelated (_r_ = -.01). Because this correlation was calculated with experimental conditions pooled, the zero correlation might suggest the influence of the training on the experimental-group teachers. The fall-spring correlations for control-group and experimental-group

Table 4

Conformity-to-Recommendations (CTR) Total Scores:
Means, Standard Deviations, and Intercorrelations,
Experimental Conditions Pooled
(N = 33)

				r			
		M	SD	1	2	3	4
1	Fall 1	70.5	10.0		.27	-.19	-.07
2	Fall 2	72.6	8.4			.07	.23
3	Spring 1	71.5	9.0				.27
4	Spring 2[a]	67.8	8.2				

[a] N = 32.

Table 5

Conformity-to-Recommendations (CTR) Total Scores:
Means, Standard Deviations, and Intercorrelations,
Experimental Conditions Pooled
(N = 28)

				r			
		M	SD	1	2	3	4
1	Fall 1	71.8	9.9		.18	-.36*	-.09
2	Fall 2	73.1	8.4			.03	.21
3	Spring 1	72.0	8.4				.16
4	Spring 2	67.4	8.2				

* $p < .10$

teachers are .21 and -.14, respectively.

Although the fall-spring correlation is essentially zero, the generalizability coefficient representing CTR scores across four occasions is .31. Because the former is based on two occasions whereas the latter is based on four occasions, it is not surprising that the latter is larger. The corresponding within-group ge eralizability coefficients are .47 and .16 for the control group and experimental group, respectively. Again, the difference between these two values suggests some influence of the training.

CTR means, standard deviations, and intercorrelations for the restricted sample are presented in Table 5. With the reduction in sample size, there is a concomitant reduction in the magnitude of the correlations between the two fall CTRs and, similarly, between the two spring CTRs. The former is reduced from .27 to .18, while the latter from .27 to .21. With the Spearman-Brown formula applied, the estimated reliabilities of the fall total CTR and spring total CTR are .31 and .28, respectively; these correlations are disappointingly low.

Additional information concerning reliability is obtained from the alpha coefficient (e.g., Cronbach, 1970), a measure of internal consistency. Table 6 presents the alpha coefficients associated with each of

Table 6

Conformity-to-Recommendations (CTR) Total Score Alpha Coefficients by Observation Occasion, For Full and Restricted Samples

Occasion	Full Sample (N = 33)	Restricted Sample (N = 28)
Fall 1	.76	.76
Fall 2	.67	.66
Spring 1	.69	.68
Spring 2	.69	.70

Note: In calculating the alpha coefficients, missing item-data were replaced with the item mean for the particular observation occasion.

the four observation occasions. Ranging from .66 to .76, these coefficients reflect respectable degrees of internal consistency.

Summary. When based on the full sample, total CTR evidences moderate reliability-stability both across the two fall occasions and across the two spring occasions. Indeed, when one considers the context in which reliability-stability was determined--the nature of the classroom observation instrument, the infrequency and brevity of classroom observations, the inherent variability of teacher behavior--the obtained correlations are almost impressive. When based on the restricted sample, however, these correlations are reduced substantially; the correlations apparently could not withstand a further reduction in size of an initially small sample. Finally, whether based on the full or restricted sample, within-occasion measures of internal consistency are relatively high.

Group Differences on CTR

Because classes were randomly assigned to the experimental conditions, the difference in fall CTR between the control group and experimental group was expected to be practically negligible. Inasmuch as classes were randomly assigned to experimental conditions and, further, pretraining and posttraining observations were conducted, the suitability of the analysis of covariance (ancova) was initially entertained. The utility of ancova in such a design would lie more in the consequent reduction of error variance than in the posttest adjustment for initial differences on the pretest (e.g., Linn & Slinde, 1977). Reducing the error term, of course, results in a more sensitive statistical test; but, because this reduction increases as the magnitude of the pretest-posttest correlation increases, ancova is of little use where this correlation is less than approximately .40 (Elashoff, 1969). Such a correlation was not expected (and, ultimately, not obtained) between the fall and spring CTR and, consequently, the use of ancova was considered unwarranted. Rather, $\underline{t}$ ratios were computed for the difference between the spring CTR means.

Table 7 presents the means and standard deviations of the CTR totals, by experimental condition. Although favoring the experimental group, the spring difference in CTR between the two groups is not statistically significant. These data indicate that, as a whole, treatment implementation was poor: Training-related teaching practices of the experimental-group teachers were not altered appreciably.

The mean differences presented in Table 7, however, can be examined at a descriptive level. For the full

Table 7

Conformity-to-Recommendations (CTR): Within-Group Means and Standard Deviations for the Fall, Spring, and Full-Year Observations for the Full Sample (N=33) and the Restricted Sample (N=28)

Full Sample

	Control (N = 16)		Experimental (N = 17)		
CTR	M	SD	M	SD	t[b]
Fall	72.38	7.14	70.71	7.67	- .65
Spring	69.19	5.85	70.56[a]	7.70	.57
Full Year	70.78	5.07	70.59[a]	5.11	- .11

Restricted Sample

	Control (N = 13)		Experimental (N = 15)		
Fall	74.15	5.39	70.97	8.07	-1.20
Spring	68.92	3.95	70.40	7.94	.61
Full Year	71.54	3.40	70.68	5.28	- .50

[a] N = 16.

[b] t was computed for the difference between uncorrelated means because the correlation between paired experimental- and control-group teachers' CTR scores was only .17 in the fall and .28 in the spring. The ts for correlated means had to be based on fewer teachers and were, in any case, essentially the same in magnitude and statistical significance as those reported here. No t reported here is statistically significant (α = .05).

sample, the experimental-group pretreatment CTR mean falls below the corresponding control-group mean (-1.67 raw-score points, or -.23 *SD*). After treatment, in contrast, the experimental-group mean is slightly above the mean of the control group (1.37 raw-score points, or .20 *SD*). These small differences are more pronounced in the restricted sample, where the standardized mean-differences are -.46 *SD* and .23 *SD*, respectively.

Also, although the experimental-group pretreatment CTR is slighlty lower than the corresponding control-group CTR and, further, the experimental-group posttreatment CTR is slighly higher than the corresponding control-group CTR, CTR in each group declines from fall to spring. This is considerably more marked for the control group, however: Raw change in CTR from fall to spring for control-group teachers is -5.23, whereas the corresponding figure for experimental-group teachers is -.57. This aspect of the mean differences in total CTR suggests that the effect of the teacher training may have been to retard a decline from fall to spring in the incidence of training-related teaching practices among experimental-group teachers.

While perhaps encouraging, these trends were not likely to have made any significant difference--statistically or practically--in the end-of-year academic achievement of the students in the control and experimental groups.

Results: Treatment Effects on Achievement

The impact of the TEP on student achievement was examined using the Johnson-Neyman technique (J-N), (e.g., Rogosa, 1980). This technique, an alternative to conventional analysis of covariance (ancova) for assessing treatment effects, is especially useful in comparing non-parallel regression lines. Because J-N is generally not as familiar as ancova, the two approaches will be briefly discussed side by side. The outline of the analyses that were performed will follow.

Basically, ancova is a combination of analysis of variance (anova) and regression analysis: The differences between posttest means are examined in conjunction with the posttest-on-pretest pooled regression. In a two-group pretest-posttest randomized design, for example, ancova evaluates the posttest mean-difference after taking into consideration between-group variance on the pretest, or covariate. Such a procedure has two important advantages over an anova on the posttest means, alone: (a) The posttest mean-difference is adjusted for any mean difference on the pretest, and (b) because pretest variance is removed from the error term and explicitly incorporated into the analysis, the reduced error variance results in more precision for the comparison of the within-group[3] regressions. This increased precision results in a greater probability of rejecting the null hypothesis (i.e., power) when, in the population, the null hypothesis does not hold.

A major assumption of the ancova model is that the covariate and treatment do not interact; that is, it is assumed that the population within-group regressions are parallel, or homogeneous. If significance tests indicate heterogeneity of regressions, ancova should not be used. But, because such statistical tests typically lack sufficient power to detect significant differences in slope (e.g., Cronbach & Snow, 1977), failure to reject the null hypothesis from a perfunctory test for homogeneity of regressions does not insure that the population within-group regressons are homogeneous. J-N makes no assumption regarding a covariate-treatment interaction and, according to Mendro (1975), "offers possibly the only satisfactory alternative to [ancova] when group regression coefficients are unequal" (quoted in Rogosa, 1977, p. 2). This technique, in contrast to ancova, establishes "regions of significance" on the covariate in which there is a statistically significant difference

[3] Here, the term "within-group" refers to the respective treatment group. Thus, a regression is computed separately within the experimental group and within the control group.

between treatments. Rather than asking the ancova question concerning the treatment effect, J-N asks the question: For what range of X, the covariate, does a significant treatment effect exist? The distinction between the two questions is nontrivial.

A brief example, in the context of the present study, may help clarify the distinction between these two methods. X and Y are the pretest and posttest, respectively, and there are two experimental conditions: Experimental-group teachers receive the TEP, and control-group teachers do not. The ancova model assesses the treatment effect by looking at the group differences on Y with X as a covariate. Essentially, Y is regressed on X and the residuals--what is not predicted by X--are examined for any treatment effect. Any obtained treatment effect is assumed to be constant over all levels of X--that is, the two slopes are assumed to be parallel (see Panel a of Figure 2).

The assumption is made, then, that the TEP has a relatively uniform impact on school achievement, regardless of whether the classes are low, medium, or high on entering ability.[4] If this is not the case--if, in fact, treatment and covariate interact--a major assumption

[4] Here, the term "ability" is used loosely, referring to the general achievement level of the class at the beginning of the school year.

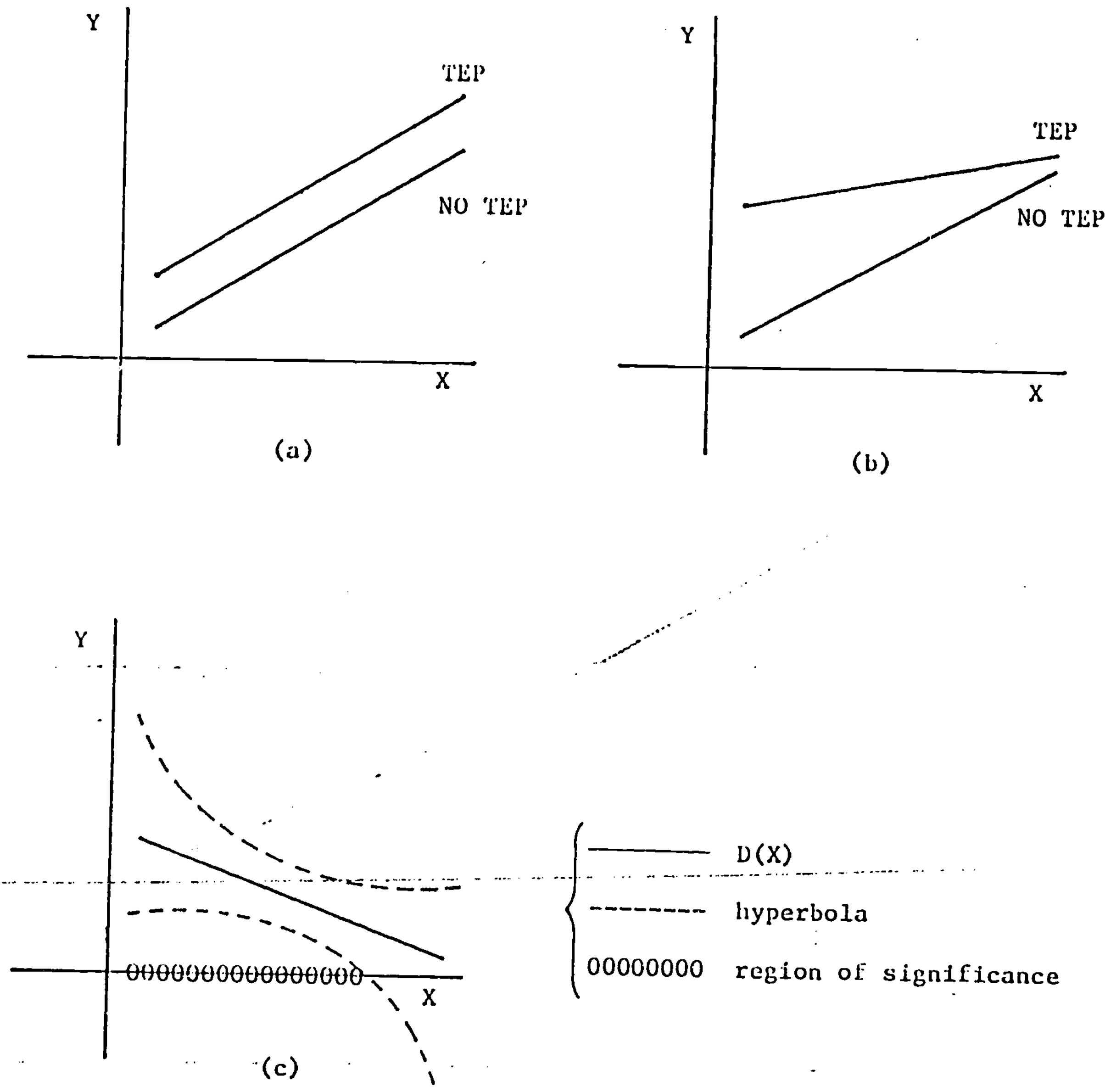

Figure 2. Comparing within-group regression lines: (a) parallel slopes, (b) treatment and covariate interact, (c) Johnson-Neyman technique.

underlying ancova is violated and alternative methods should be considered. If ancova _is_ used, the results may be very misleading, depending on the degree of interaction present.

Imagine that an interaction exists between treatment and covariate such that the TEP has a large effect in low-ability classes, but a virtually negligible effect in high-ability classes (see Panel _b_ in Figure 2). If ancova _is_ used, a single vertical distance between the two within-group regressions is assessed and, in turn, used as an estimate of the adjusted mean-difference on _Y_. This difference, however, is evaluted at the value of _X_ corresponding to the weighted average of the two group means; as such, this difference can be thought of as an "average" or "overall" treatment effect (Rogosa, 1980). In the context of the interaction presented in Panel _b_ of Figure 2, the reported treatment effect would be misleading, indeed: It would underestimate the treatment effect for low-ability classes and overestimate the treatment effect for high-ability classes. The problem is especially pronounced where the lines of a markedly disordinal interaction cross in the middle of the range of _X_: The adjusted mean-difference on _Y_ would be roughly zero (i.e., because this adjustment is made near the point at which the lines cross), while an examination of the vertical differences between the regression lines at low and high values of X would result in a drastically different impression. Clearly, with nonparallel straight lines, reports only of the vertical distance are neither very meaningful nor very compelling (Rogosa, 1980).

J-N, on the other hand, does not require the assumption of homogeneity of regressions. A line, D(X), is determined that represents the vertical distance (D), for a given range of X, between the two sample within-group Y-on-X regressions. "For the comparison of the within-group regressions, D(X) is the key summary of the data" (Rogosa, 1980). To assess statistical significance for D(X), a simultaneous confidence band is constructed for the difference of the population within-group regressions. This band is bounded by hyperbolae (see Panel c of Figure 2).

The simultaneous version of J-N identifies regions[5] on X in which there is a statistically significant difference between the two sample within-group regression lines. These regions are identified by the manner in which the confidence band intersects the X axis. Values of X that fall outside the confidence band are in the region of significance. In Pancel c of Figure 2, where

[5] Before constructing the simultaneous confidence band, one can conduct a preliminary test to see whether or not any regions exist. See Rogosa (1980) or, for an alternative formula, Serlin and Levin (1980).

D(X) is based on the interaction presented in Panel b, one sees that this region covers low-ability classes to classes moderately high in ability. The region of significance does not extend beyond this point. The conclusion would be that the teacher training had an appreciable impact on achievement for this particular range of X, but not beyond. It is in this manner that the results of J-N can be more revealing and meaningful than an ancova estimate of the "average" treatment effect.

Procedures and Outline of Analyses

Each class mean was weighted by the corresponding number of students on which the mean was based. Such weighting takes into account differences in precision of class means that results from different class sizes. Weighting is especially advisable when there is marked variability in this class characteristic (Cronbach, 1976, pp. 4.7-4.11; Cronbach & Webb, 1975).

First, the two within-group regressions were compared. The model for these regressions, computed separately within the experimental group (subscript E) and the control group (subscript C), is:

$$\begin{aligned} Y_j &= \alpha_E + \beta_E X_j + \epsilon_j \qquad \text{for } j=1,\ldots,n_E \\ Y_j &= \alpha_C + \beta_C X_j + \epsilon_j \qquad \text{for } j=n_E+1,\ldots,N \end{aligned} \qquad (1)$$

where α and β represent, respectively, the intercept constant and the slope associated with the posttest (*Y*) on pretest (*X*) regression. The within-group regressions can be combined as:

$$Y_j = \beta_1 + \beta_2 T_j + \beta_3 X_j + \beta_4 T_j X_j + \epsilon_j \quad \text{for } j=1,\ldots,N \tag{2}$$

where the new terms are the regression coefficient β_2 associated with the treatment *T*--the latter being a dummy variable coded 0 (control) or 1 (experimental)--and the regression coefficient β_4 associated with the interaction of treatment *T* and covariate *X*. β_4 is equivalent to the difference between the two within-group regression coefficients; in the context of Equation 1, $\beta_4 = \beta_E - \beta_C$. β_1 is the intercept constant and is equivalent to α_C in Equation 1. β_3 is the regression coefficient associated with the covariate, and is equal to β_C in Equation 1. Parameters in Equation 2 were estimated by ordinary least squares; analogous relations hold for the sample quantities. (Sample estimates are denoted by the lower-case "b.")

Again, J-N involves the calculation of a line, D(X). D(X) represents, for the range of *X*, the vertical distance between the two sample within-group regressions. Two sample estimates are needed to determine $D(X)$:

$$D(X) = b_2 + b_4X \qquad (3)$$

The line $D(X)$ is plotted against the $\underline{X}$ and $\underline{Y}$ axes. Here, b_2 is the point at which $D(X)$ intersects the $\underline{Y}$ axis and b_4 is the slope of $D(X)$. The point of intersection at the $\underline{X}$ axis is equal to $-b_2/b_4$. The $\underline{Y}$ axis, scaled in the units of the particular posttest, reflects the vertical distance between the two sample within-group regressions and intersects the $\underline{X}$ axis at $\underline{Y} = 0$. Thus, the difference between the two regression lines is zero at the point at which $D(X)$ intersects the $\underline{X}$ axis.

With this information, then, a plot such as that appearing in Panel c of Figure 2 can be constructed. From a plot like this, one can determine the vertical distance between the two sample within-group regressions for a given $\underline{X}$. Alternatively, this distance, for a given $\underline{X}$, can be assessed by solving for $D(X)$ in Equation 3. (The procedure for constructing these plots is outlined in Appendix A.)

Identifying the simultaneous region of significance involves few additional calculations (procedures are presented in Rogosa, 1980). The unstandardized regression coefficients b_2 and b_4 and elements of the corresponding variance-covariance matrix are required for these computations. A 100(1 - α) percent simultaneous confidence band for the difference of the two population within-group regressions is constructed. As noted above, the region of significance is identified by the intersection of the confidence band with the *X* axis. The region of significance comprises those values of *X* that fall outside the confidence band (see Panel *c* of Figure 2). If the confidence band does not intersect the *X* axis, there is no region of significance. (The procedure for constructing a 95% simultaneous confidence band is outlined in Appendix A).

If a "packaged" regression program (e.g., Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) is used and the terms in Equation 2 are entered stepwise, one additionally can examine the ancova estimate of the treatment effect, as well as test the homogeneity-of-regressions assumption. The training term (T) would be entered on the first step, followed by the covariate (X) on the second step, with the interaction term (TX) entered on the third and final step. Because it represents the difference of the two sample within-group regression coefficients, b_4 provides information bearing on the homogeneity-of-regressions assumption. When the two lines are parallel, the slopes are equal and $b_4 = 0$. Conversely, when the two lines are appreciably nonparallel--that is, when interaction exists--the slopes are different and b_4 is comparatively large. Thus, evaluating the magnitude of b_4 leads to a conclusion concerning the assumption of homogeneity of regressions.[6]

The treatment effect that would be provided by conventional ancova is obtained by examining the regression coefficient associated with the treatment term (T) at the second step of the regression procedure. This coefficient is identical to the treatment effect that would be obtained from a packaged ancova program and, as such, represents the adjusted mean-difference on the dependent variable. The pooled within-group regression coefficient is equivalent to the regression coefficient associated with the covariate at the second step of the regression procedure. The pooled slope is fundamental to ancova and, further, enables one to compute the adjusted means.

Thus, assessing treatment effects on student

[6] An interesting observation is that useful regions of significance can be obtained even where one fails to reject the null hypothesis $\beta_4 = 0$ (Rogosa, 1981).

achievement comprises three steps: For each dependent measure, (a) the sample within-group regressions were examined; (b) J-N was employed to identify possible regions on the covariate in which the two within-group regressions are significantly different; and (c) the ancova estimate of the treatment effect was examined--an estimate representing the treatment effect for the "average" individual over the range of _X_.

Pooling Grades

Table 8 presents the number of "quasi-classes," by grade and experimental condition. As used here, a quasi-class comprised only those students in a particular grade in a class. In the case where a teacher had, say, only fourth-grade students, the class and the quasi-class were identical. In contrast, the quasi-class merely was a subset of the class for the teacher who had a fourth-fifth combination.

It is clear from this table that the within-grade number of control-group and experimental-group classes is small--too small, in fact, to warrant a separate analysis of treatment effects on student achievement for each grade. Consequently, all analyses were conducted with different grade levels combined. This was accomplished through a linear transformation of the CTBS raw scores. First, each student's raw score was converted to a _T_ score (_M_ = 50, _SD_ = 10). This was done separately

Table 8

The Number of Quasi-Classes, by Grade and Experimental Condition

Experimental Condition	Grade		
	4	5	6
Control Group	7	10	4
Experimental Group	9	9	5

for each of the three grades. These T scores were then aggregated at the class level, yielding a mean T-score for each class.

As was noted above, each class mean was weighted by the corresponding number of students. This entailed assigning each student the mean of his class. Because the computer consequently treated the student as the unit of analysis, however, certain statistics (e.g., F ratio, standard error) needed to be adjusted to reflect the actual number of classes represented in the analyses. (The procedures for adjusting these statistics are outlined in Appendix B.)

Descriptive Statistics and Within-Group Regressions

The CTBS pretest and posttest means and standard deviations are presented in Table 9. Table 10 presents the intercorrelations among the measures; the pretest-posttest correlation for each measure is reported in the diagonal. As can be seen from Table 10, the pretest-posttest correlation for the total score is .84--a value representing the relative stability of performance over a 12-month period.

The sample within-group regression equations for the dependent measures are presented in Table 11. These within-group regressions were plotted and, accompanied by the corresponding within-group scatterplots, appear in Figures 3-8. The actual range of data for each group

Table 9

CTBS Pretest and Posttest Means and Standard Deviations

	Control (N = 13)		Experimental (N = 15)		Pooled[a] (N = 28)	
Test	M	SD	M	SD	M	SD
Pretest						
Reading Total	50.40	5.71	49.68	4.82	50.00	5.24
Mathematics Total	50.15	5.23	49.88	4.56	50.00	4.87
Total Score	50.29	5.78	49.76	4.87	50.00	5.30
Posttest						
Reading Total	50.98	5.32	49.21	4.83	50.00	5.13
Mathematics Total	50.37	4.54	49.70	4.07	50.00	4.29
Total Score	50.70	4.99	49.43	4.63	50.00	4.83

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50, SD = 10) of student level scores.

[a] Experimental conditions pooled.

Table 10

Intercorrelations of CTBS Subtests and Total Score

	1	2	3
1 Reading Total	(86)	91	98
2 Mathematics Total	86	(77)	98
3 Total Score	95	96	(84)

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50, SD = 10) of student-level scores. Weighted between-class correlations are reported; decimals have been omitted. Pretest correlations appear above the diagonal, posttest correlations appear below the diagonal, and the pretest-posttest correlation for each measure appears in the diagonal.

Table 11

Within-Group Regressions[a] of CTBS Posttest on Pretest

	Control (N = 13)				Experimental (N = 15)			
Test	r	a	b	SE(b)	r	a	b	SE(b)
Reading Total	.82	12.290	.768	.159	.91	3.995	.910	.117
Mathematics Total	.70	19.915	.607	.186	.85	11.983	.756	.131
Total Score	.79	16.463	.681	.160	.90	7.069	.851	.117

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50, SD = 10) of student-level scores.

[a] r = pretest-posttest correlation.
a = intercept constant.
b = unstandardized regression coefficient.
SE(b) = standard error of b.

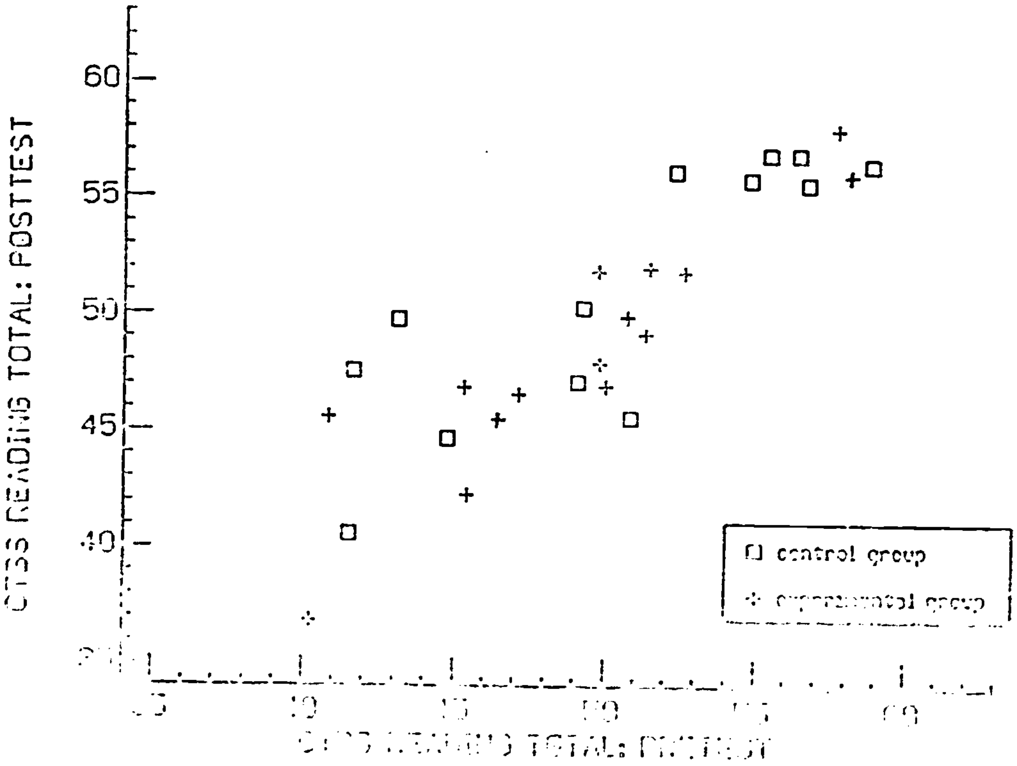

Figure 3. Within-group scatterplots: Reading Total.

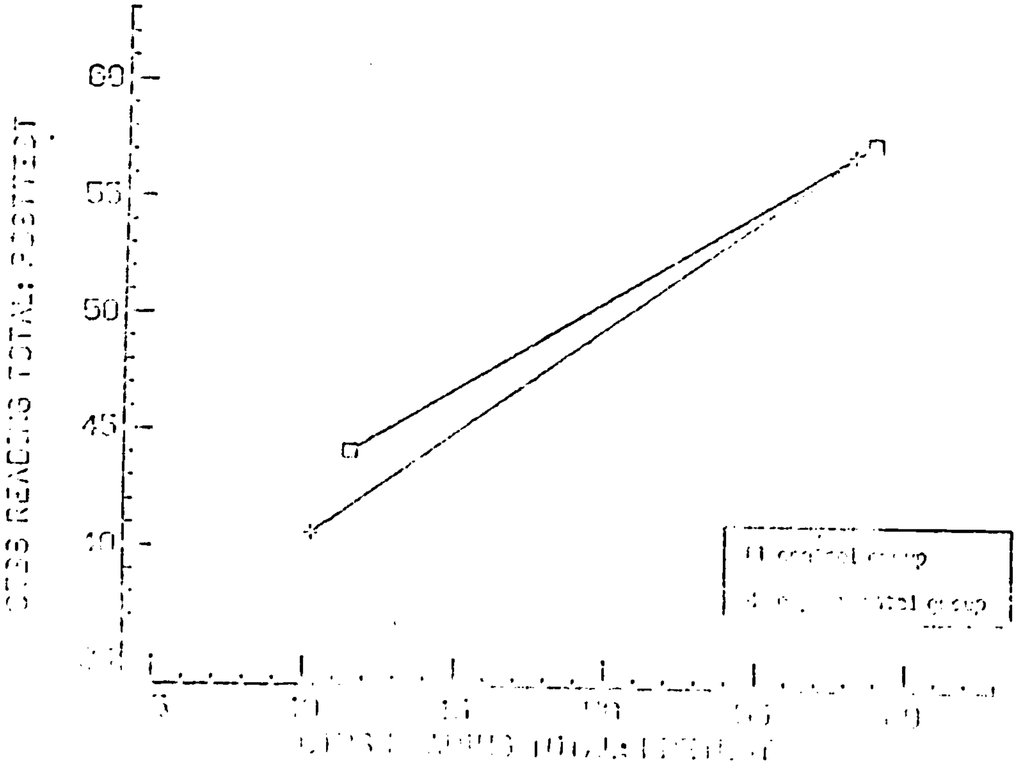

Figure 4. Within-group regressions: Reading Total.

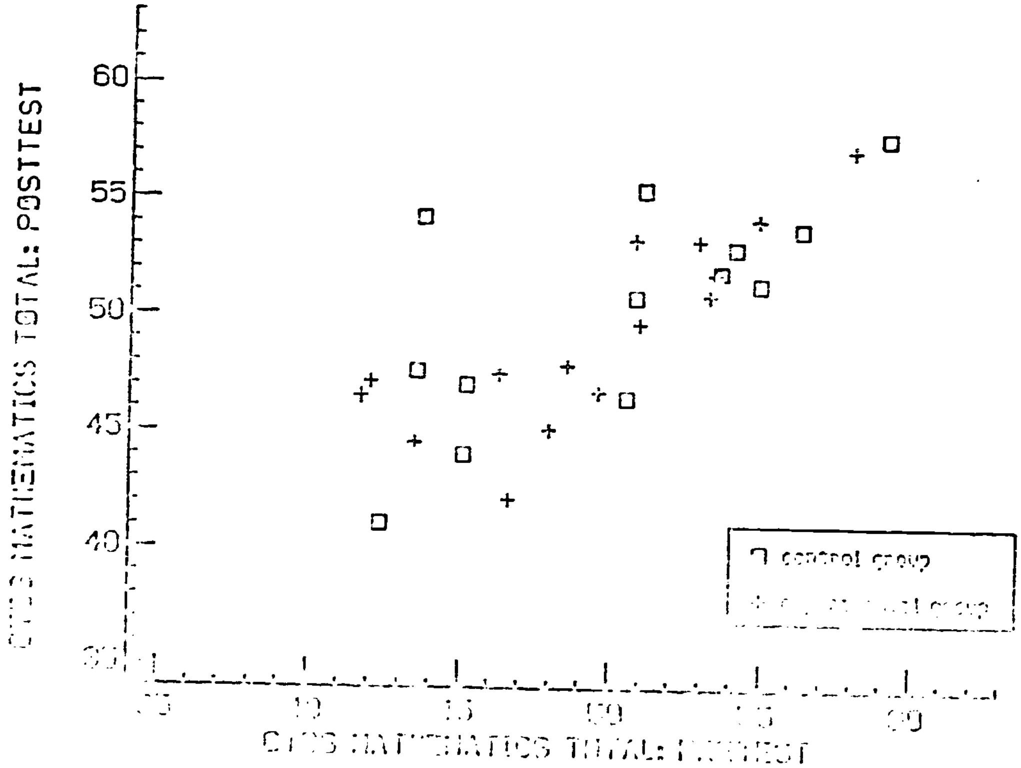

Figure 5. Within-group scatterplots: Mathematics Total.

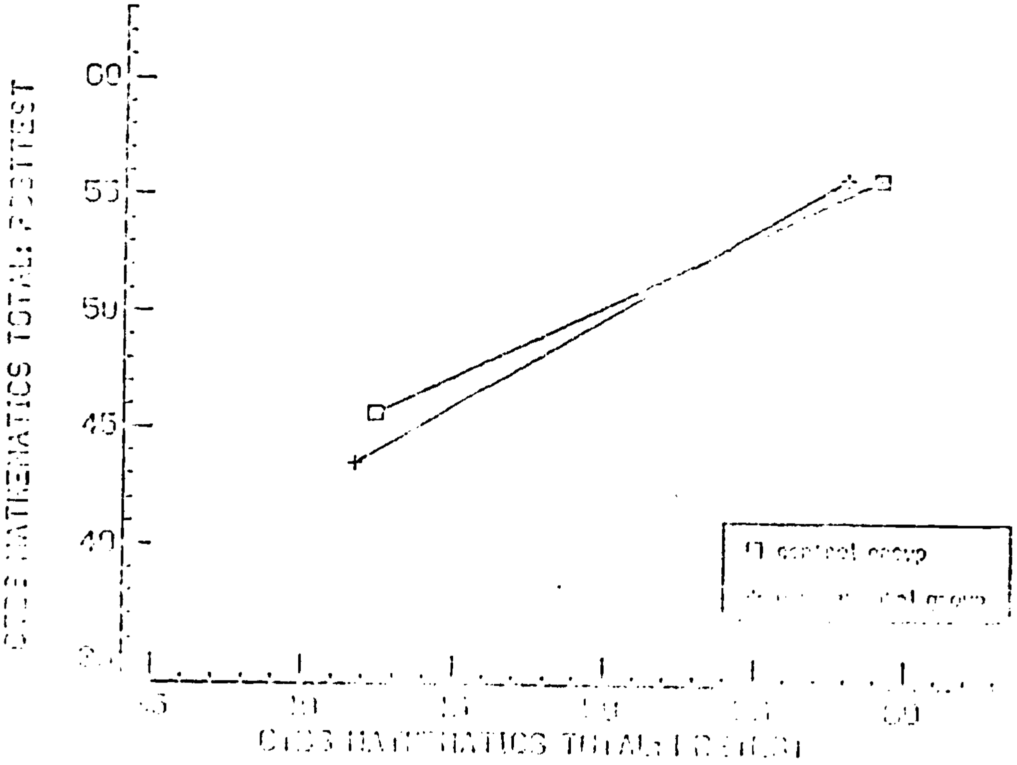

Figure 6. Within-group regressions: Mathematics Total.

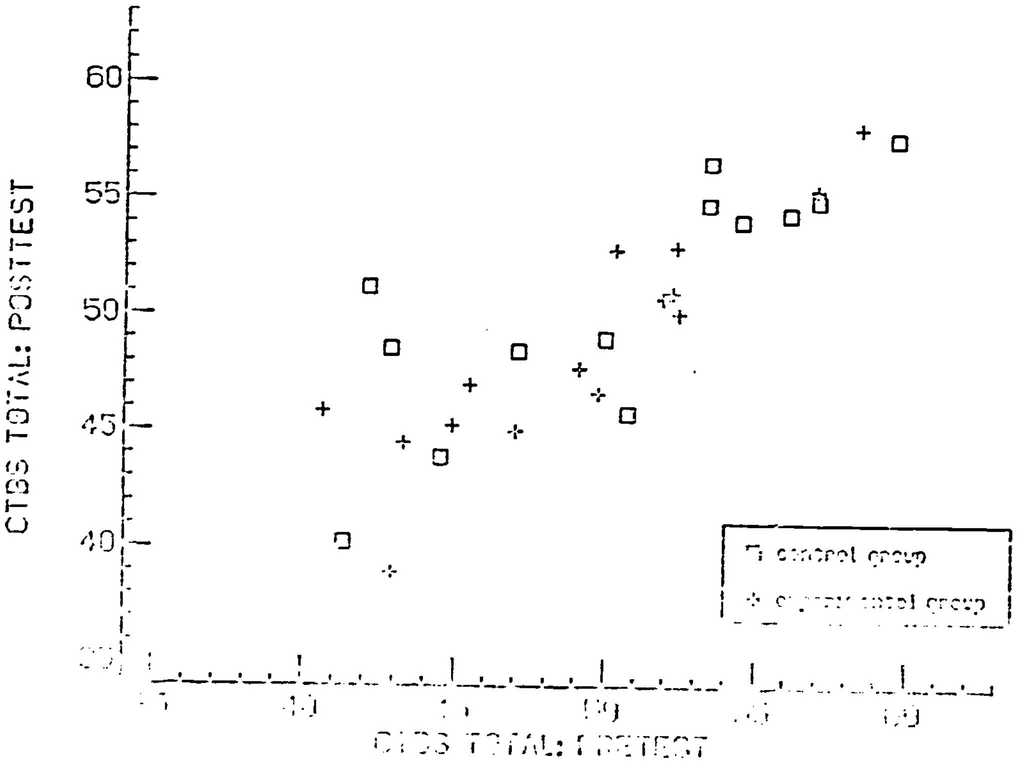

Figure 7. Within-group scatterplots: Total Score.

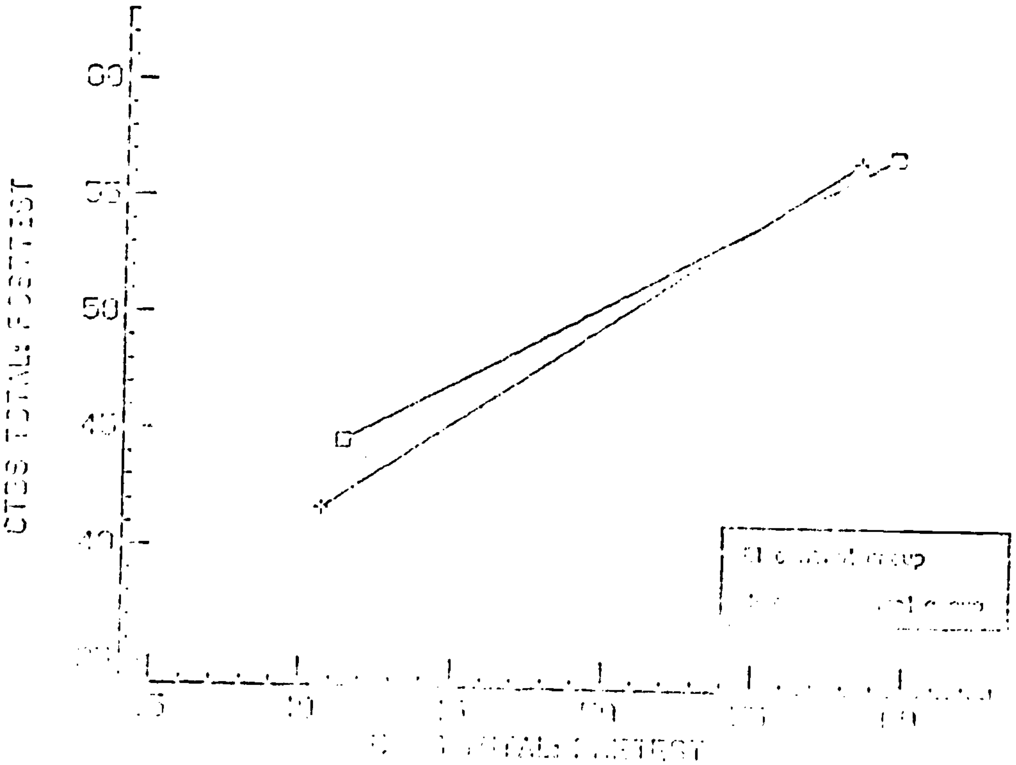

Figure 8. Within-group regressions: Total Score.

corresponds to the range of _X_ for which the particular regression was plotted.

As these figures illustrate, each pair of within-group regressions is similar: Vertical displacement at any point on _X_ is, at most, slight. From a mere visual inspection of these pairs of within-group regressions, then, it does not appear that the treatment had an appreciable effect on end-of-year student achievement.

Johnson-Neyman Analyses

Table 12 presents, for each dependent measure, the sample regression coefficients b_2 and b_4 corresponding to Equation 2 and reported in the final step of the regression procedure. Again, the regression coefficients b_2 and b_4 are associated with, respectively, the treatment term (T) and the term representing the interaction of treatment and covariate (TX). Additionally, three entries in the variance-covariance matrix[7] of these regression coefficients are needed for J-N: The variance of b_2 (denoted s_{22}), the covariance of b_2 and b_4 (denoted s_{24}), and the variance of b_4 (denoted s_{44}). Table 13 presents these three values for the dependent measures.

[7] Although there are several packaged programs that generate this matrix, the simplest by far is the SAS SYSREG procedure (SAS Institute Inc., 1979). All one does, in addition to specifying the regression model, is request option COVB. This option outputs the variance-covariance matrix of the unstandardized regression coefficients for the specified model.

Table 12

Selected Regression Coefficients[a] for Full J-N Equation

Test		*b*	SE(b)	*F*
Reading Total	b_2	-8.294	10.071	< 1
	b_4	.142	.200	< 1
Mathematics Total	b_2	-7.931	11.170	< 1
	b_4	.149	.222	< 1
Total Score	b_2	-9.394	10.125	< 1
	b_4	.171	.201	< 1

Note: *N* = 28. Grades were pooled through a within-grade transformation (*M* = 50, SD = 10) of student-level scores.

[a] b_2 = unstandardized regression coefficient associated with the treatment term (T).
b_4 = unstandardized regression coefficient associated with the interaction term (TX).
SE(b) = standard error of *b*.

Table 13

Elements[a] of the Variance-Covariance Matrix of Regression Coefficients for Full J-N Equation

Test	s_{22}	s_{24}	s_{44}
Reading Total	101.4323	-2.0059	.0401
Mathematics Total	124.7725	-2.4616	.0491
Total Score	102.5227	-2.0271	.0405

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50, SD = 10) of student-level scores.

[a] s_{22} = variance of b_2.

s_{24} = covariance of b_2 and b_4.

s_{44} = variance of b_4.

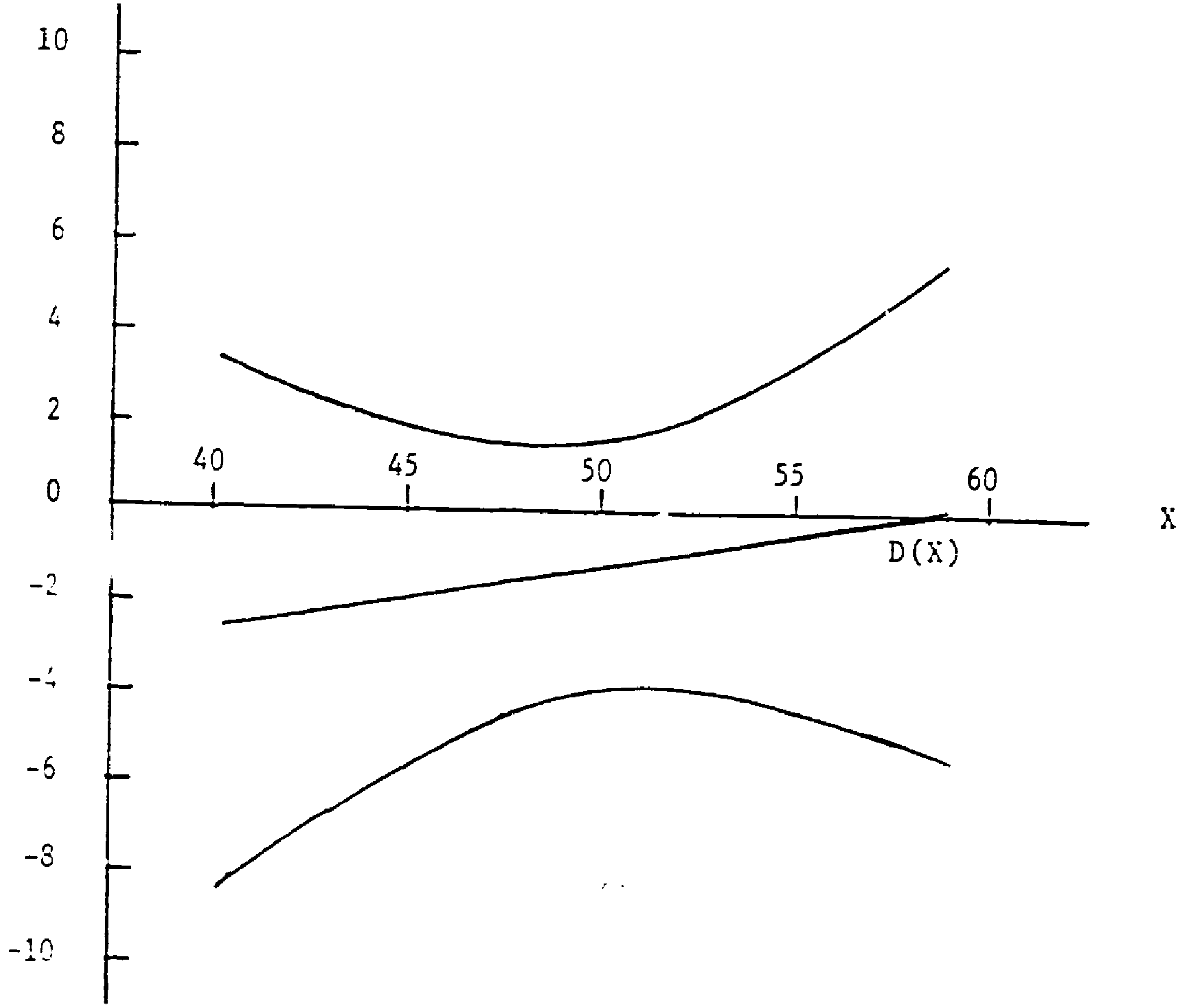

Figure 9. D(X) and 95% simultaneous confidence band: Reading total.

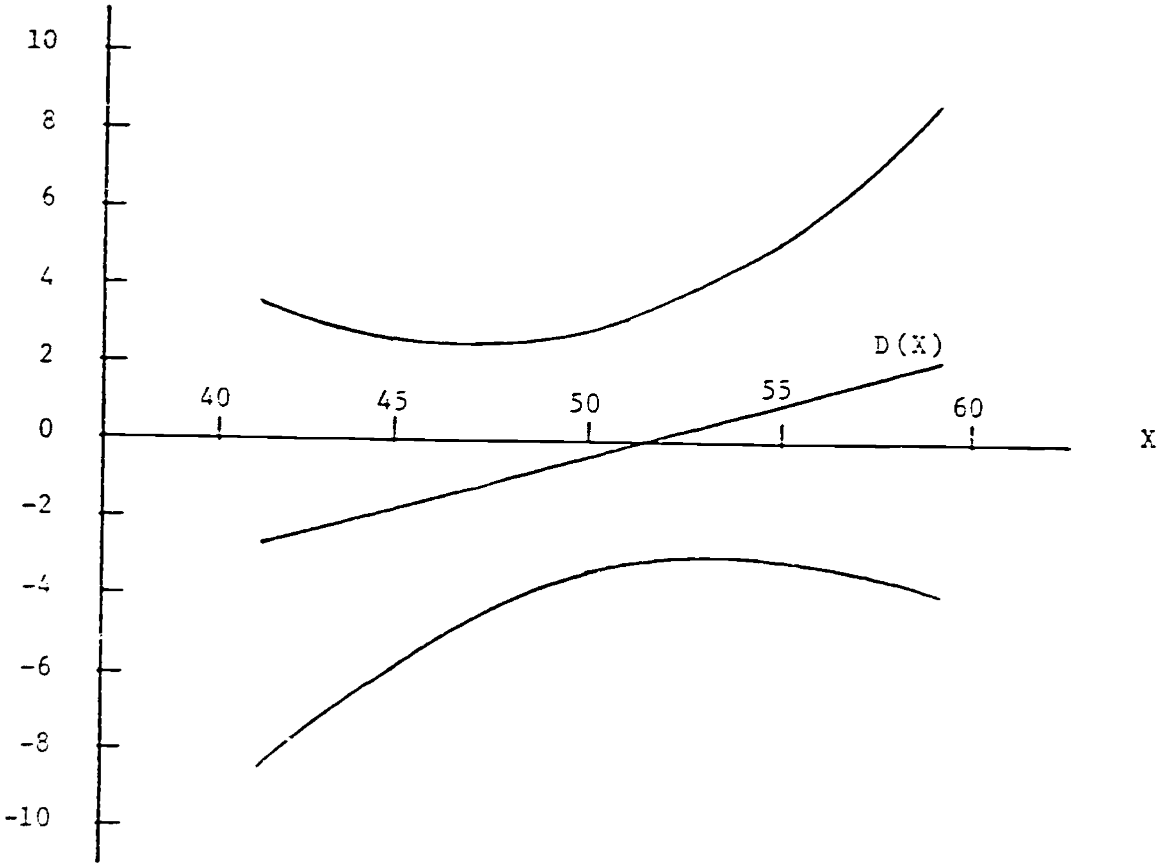

Figure 10. D(X) and 95% simultaneous confidence band: Mathematics total.

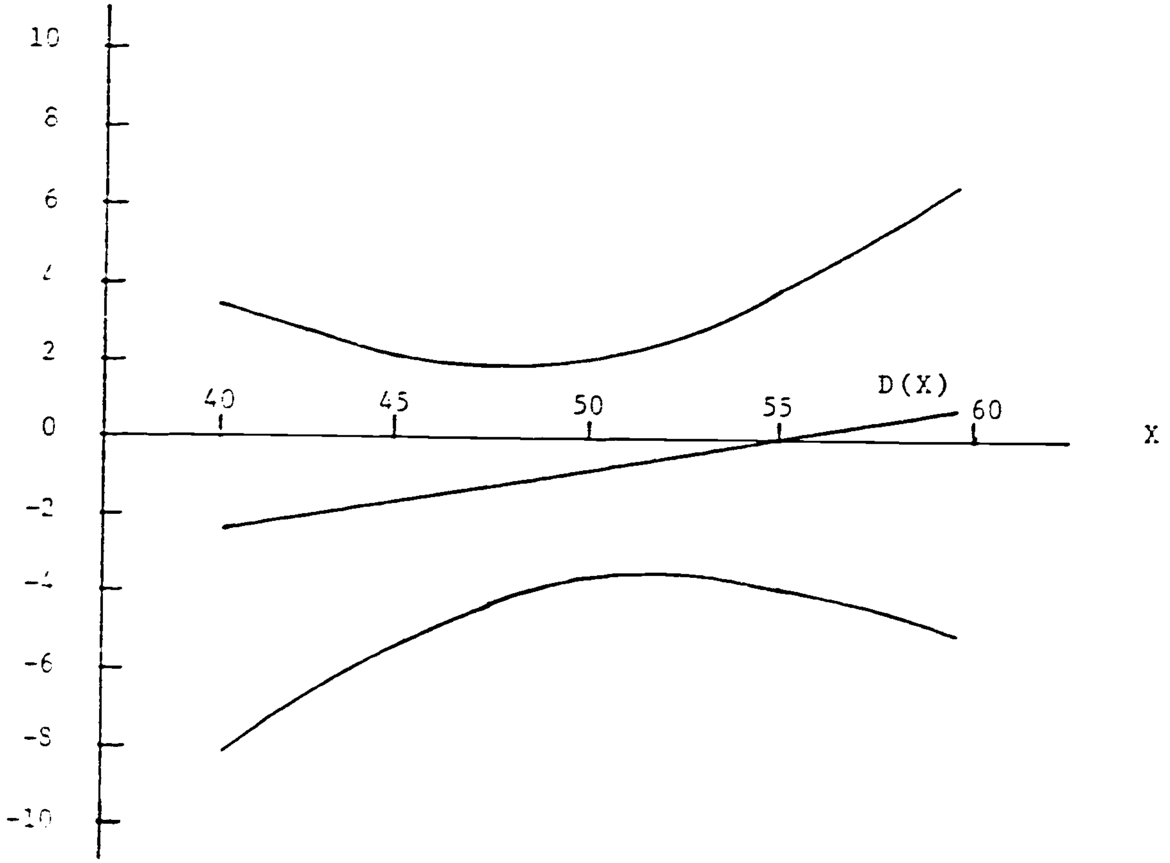

Figure 11. D(X) and 95% simultaneous confidence band: Total score.

procedure, it will be recalled, one is testing the assumption regarding homogeneity of regressions. The $\underline{F}$ ratios associated with b_4 (see Table 12) indicate, by conventional standards, that this assumption was not violated for any of the dependent measures. Thus, although some of the pairs of within-group regression lines are nonparallel, the degree of heterogeneity present is not sufficient to be statistically significant.

Table 14 presents the ancova estimate of the treatment effect and the corresponding 95% confidence interval for each measure. (The procedure for constructing these confidence intervals is outlined in Appendix C.) As can be seen from this table, these estimates of the "average" treatment effect are small and not significantly different from zero. Further, the 95% confidence interval in each case spans zero (which, of course, is expected with chance findings) and extend comparatively far into the positive region. It is conceivable, then, that with a replication study the obtained treatment effects all could be positive (though nevertheless statistically nonsignificant). The pooled within-group regression coefficients and the ancova-adjusted posttest means are presented in Table 15.

Summary

The teacher training was ineffective in improving student achievement. This conclusion held when (a) regions on the covariate were sought in which the difference

Table 14

Ancova Estimates of Treatment Effects,
With 95% Confidence Intervals

Test	Treatment Effect	F	Lower and Upper End-Points of a 95% Confidence Interval
Reading Total	-1.171	1.234	-3.333, .991
Mathematics Total	- .481	< 1	-2.895, 1.933
Total Score	- .866	< 1	-2.999, 1.267

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50, SD = 10) of student-level scores.

Table 15

Pooled Within-Group Regression Coefficients and Ancova-Adjusted Posttest Means

		Control (N = 13)	Experimental (N = 15)
Test	r (pooled)	M	M
Reading Total	.83	50.65	49.48
Mathematics Total	.68	50.27	49.78
Total Score	.76	50.48	49.61

Note: N = 28. Grades were pooled through a within-grade T-score transformation (M = 50. SD = 10) of student-level scores.

of the within-group regressions was statistically significant and (b) the ancova estimate of the "average" treatment effect was examined. This result is not surprising, of course, in view of the poor treatment implementation.

Results: The Relationship Between CTR and Achievement

In addition to examining the effects on student achievement of the intervention, one can pool experimental conditions and carry out process-product analyses. That is, correlations can be obtained between the teachers' conformity-to-recommendations (CTR) and student achievement--whether the former was naturally occurring or attributable to the training.

Thus it is acknowledged that, irrespective of experimental condition, there will be variability in CTR. To be sure, not all experimental-group teachers would be expected to demonstrate the same degree of CTR; teacher attitudes, beliefs, motivations, and so on, doubtless are operating here. And the assumption would not be made that, by virtue of their group assignment, control-group teachers would demonstrate no CTR whatsoever. On the contrary, one would expect natural variability in CTR here, as well.

Such an analysis is informative in the present context in that it yields additional evidence concerning the relevance of the teacher training to student achievement. In this sense, the "effects" of a program or treatment can be evaluated by examining all teachers, regardless of the experimental condition to which they initially had been assigned.

In the following analysis, a full-year measure of total CTR was obtained by averaging total CTR across the four occasions; this served as the "process" measure. "Product" was a residual score based on the CTBS posttest total. These residuals were obtained by regressing the CTBS posttest total on the pretest at the student level. (The difference between the obtained and the predicted score is the residual and represents performance on the posttest that is uncorrelated with pretest performance.) These residuals were then aggregated at the class level and, in turn, correlated with the process measure.[8] The result is a part correlation.

The resulting correlation between total CTR and residual achievement is $\underline{r} = .29$ ($\underline{p} > .10$). This correlation increases considerably, however, with the removal of one discrepant case ($\underline{r} = .40$, $\underline{p} < .05$). Figure 12 presents the scatterplot for this correlation, with the outlier identified. Thus the training as whole, derived from previous process-product research (Brophy & Evertson, 1974; McDonald & Elias, 1976; Soar, 1973; Stallings & Kaskowitz, 1974), ostensibly has some pedagogical value in the present context.

[8] Unlike the analyses presented in the previous chapter, here class mean-achievement was not weighted for differences in class size. Because it would have entailed similarly weighting CTR--a teacher variable, the weighting of which is inappropriate--weighting was deemed undesirable for this analysis.

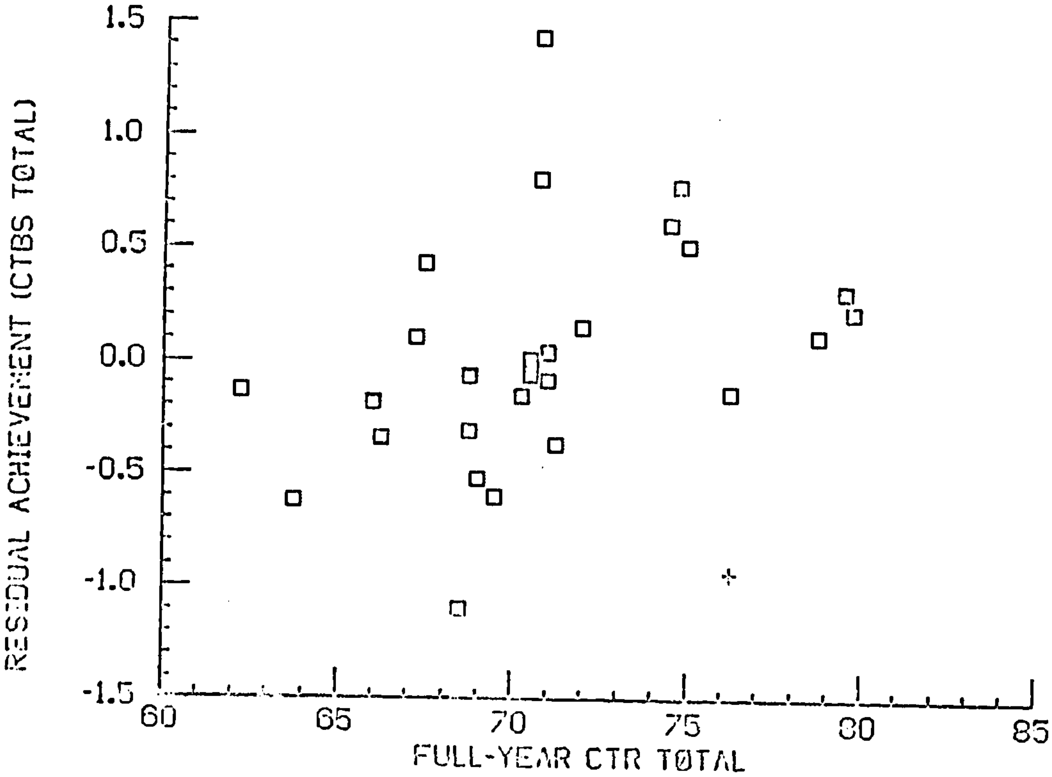

Figure 12. Scatterplot of full-year CTR and residual achievement. (Discrepant case is denoted by "+".)

Discussion and Recommendations

As an experiment, this study failed to corroborate the positive results obtained previously in similar classroom based experiments (Anderson et al., 1979; Crawford et al., 1978; Good & Grouws, 1979; Stallings et al., 1979). At the end of the school year, the experimental-group teachers did not evidence markedly greater conformity to the training recommendations than that exhibited by the control-group teachers. Further, the classes of these two groups of teachers were not appreciably different in end-of-year academic achievement.

Poor Treatment Implementation

There was a priori reason to expect the desired change in teaching practices among experimental-group teachers. After all, the intervention in the present study was the same as the "minimal training" condition in the study by Crawford et al. (1978). In that study, it will be recalled, the minimally and maximally trained teachers exhibited similar conformity to the recommendations, and both experimental groups evidenced marked superiority over the control group. Further, the experiments conducted by Anderson et al. (1979) and Good and Grouws (1979) similarly did not involve a comprehensive delivery system. And, as reported above, both studies resulted in positive change in training-related teaching practices among experimental-group teachers.

Here, it will be argued that poor treatment implementation in the present study was due in large part to several methodological and contextual differences between this study and those conducted by Crawford et al. (1978), Anderson et al. (1979), and Good and Grouws (1979).

Classroom observations. Classes in the present study were observed for a maximum of eight hours throughout the entire school year: two two-hour periods in both the fall and spring. Crawford et al. (1978), in contrast, obtained classroom observations for approximately 16 full days throughout the school year--before, during, and after treatment.

While the manifest function of classroom observations is to obtain information concerning classroom characteristics and events, the latent function of such observations may be to facilitate treatment implementation. Minimally trained teachers in the Crawford et al. (1978) study unwittingly may have come to regard the relatively frequent and lengthy classroom observations as a kind of supervision or monitoring. If so, the conduct of classroom observations likely would have enhanced the compliance of these experimental-group teachers with the training recommendations. The failure of experimental-group teachers in the present study to implement the training recommendations, then, may have resulted from the relatively infrequent and brief classroom observations. That is, if experimental-group teachers would perceive the conduct of classroom observations as a supervisory mechanism, then poor implementation might be attributed to their receiving much less of such supervision.

The plausibility of this conjecture must be evaluated in view of the finding reported by Anderson et al. (1979). As noted above, there were two experimental groups: trained and observed, and trained but unobserved. Analyses of end-of-year achievement data indicated that both groups were equally superior to the control group in improving reading achievement. If one assumes that equal effectiveness in improving achievement must have been accompanied by correspondingly equal conformity to the training recommendations, these results suggest that the absence of observers in the classes of experimental-group teachers does not reduce treatment implementation. Thus, it might be argued, the comparatively low amount of classroom observations in the present study cannot be held responsible for the ineffectiveness of its training in bringing about the desired changes in teaching practices among experimental-group teachers.

There remains, however, a fundamental difference between the present study and the one conducted by Anderson et al. (1979).

Initial meetings with teachers. In the present study, teachers never met with project staff for discussion, question-and-answer, and so on; the TEP simply were mailed to the experimental-group teachers. Anderson et al. (1979), in contrast, met twice with all experimental-group teachers (i.e., including those trained but unobserved)--once to describe the purpose of the study and distribute the training material, and the second time to discuss the instructional model presented in the training material.

These meetings likely fostered treatment implementation. First, the meetings doubtless were informative, facilitating understanding of the instructional model and its applicability. Second, by holding these meetings, the project staff were in a position to communicate enthusiasm for the training and personal concern for the teachers. The teachers' perception of that enthusiasm and concern could effect a favorable disposition of the teachers to the overall project and, in turn, enhance subsequent implementation. In short, these initial meetings with teachers may have served to address the three general factors affecting implementation of change proposals that were outlined by Doyle and Ponder (1977) and discussed above: instrumentality, congruence, and cost. The conduct of the two meetings in the Anderson et al. (1979) study, then, may have offset the absence of classroom observers for the trained-but-unobserved teachers.

Clearly, the relative effects on treatment implementation of classroom observations and treatment delivery remains an open question. The implication of the minimal-maximal parity reported by Crawford et al. (1978) is clouded by the comprehensive observations conducted in the classrooms of all teachers. Similarly, the implication of the trained-but-unobserved finding reported by Anderson et al. (1979) is obscured by the initial meetings attended by all teachers. And neither of these factors was manipulated in the study conducted by Good and Grouws (1979): All teachers attended initial meetings, and observations were conducted in all classrooms.

The SES context. A third factor possibly attenuating treatment implementation in the present study is the urban low-SES context in which it was conducted. The Crawford et al. (1978) study, it will be recalled, was conducted in a middle-SES school district, as was the study conducted by Anderson et al. (1979). Good and Grouws (1979) reported only that most of the participating schools were in low-SES "areas" (p. 356). (Without more information on this sample, it is difficult to disucss context effects on treatment implementation.

The assumption will be made nevertheless that the sample in the present study was lower in socioeconomic status than that in the Good and Grouws [1979] study.)

The context in the present study may have made it too difficult for the experimental-group teachers to respond positively and cooperatively to the training. The demands of teaching in an urban low-SES climate doubtless are quite different from those in other contexts. Indeed, Levy (1979) likened teaching in the urban school to engaging in combat. Further, students not uncommonly are ill-prepared for the motivational and cognitive demands of classroom processes, posing additional problems for teachers. A local newspaper, in fact, reported that this particular school district fell "at the bottom of the heap" (San Francisco Chronicle, November 9, 1979, p. 5) among the some 1,000 districts in California in average performance on a state-wide proficiency test administered during the school year in which the present study was conducted.

Treatment implementation, then, may have been attenuated by the context of the intervention. Although all teachers were volunteers and, hence, presumably disposed, initially at least, to cooperate in it, they may have been too distracted by their more difficult regular teaching activities to be able to comply with the training recommendations.

Proposition 13. In June 1978, Proposition 13 was passed in California and resulied in a number of measures to reduce property taxes and limit government spending. In addition to its large impact on school finance, Proposition 13 had other, perhaps equally serious, repercussions. The president of the California State Board of Education stated that

> Proposition 13. . . .had a significant effect on the morale of teachers and other public employees. Teachers feel that the state has abrogated all collective bargaining agreements--declaring them null and void and canceling all pay raises. Teachers realize that their jobs are tied to very uncertain revenue sources. . . .As uncertainty increases over the state bailout in future years, this morale problem may become worse. (Kirst, 1979, p. 431)

Empirical evidence supporting this appraisal was provided by Calfee and Pessirilo-Jurisic (1979), who interviewed 81 teachers, 197 principals and vice-principals, and six administrators to gain "insight into the nature of declining morale among public school teachers in California" (p. 4). The interviews were conducted in a large school district in the San Francisco Bay Area during the school year in which the present study was conducted, the first school year following the passing of this initiative.

The results indicated that many of these 104 educators perceived disconcerting changes in California education--changes they attributed primarily to Proposition 13.

By and large, "educators felt that they were working harder than ever, under worsening conditions and receiving fewer rewards--both psychologically and financially" (p. 20).

In retrospect, the state of affairs resulting from Proposition 13 may not have provided suitable conditions for a study that called upon teachers to expend additional time and energy. Thus, the "climate" of the California public school in the early Proposition 13 era may have attenuated treatment implementation in the present study. Although the teachers volunteered for the experiment in September 1978, three months after Proposition 13 was passed, its effects on teacher morale likely grew stronger after the school year got underway and teachers and administrators had further opportunity to consider its immediate and potential effects. Teacher morale, then, may have suffered *after* the teachers volunteered to participate in the study and *during* the school year in which it was conducted.

Absence of Treatment Effects on Student Achievement

Clearly, in the present study, treatment implementation was a necessary condition for treatment effects on student achievement. Analyses of the classroom observation data indicated that relevant teaching practices were not altered appreciably by the intervention. This, indeed, appears to be the most compelling reason for the absence of treatment effects on student achievement.

There are, however, additional, ostensibly plausible, reasons for the absence of such effects.

Irrelevant teaching recommendations. First, if the teaching recommendations contained in the TEP were actually unrelated to student achievement, one obviously would not expect treatment effects on this criterion in the present study. It is unlikely, however, that this is responsible for the obtained results. The recommendations, it will be recalled, were based on previous process-product correlations obtained from studies of teaching and learning in regular classrooms. Further, many of the process-product correlations that were incorporated into the TEP were originally obtained in a low-SES context. Thus, it seems fair to assume that the teaching recommendations were related to student achievement--achievement in a school district like the one in which the present study was conducted. And the obtained correlation between CTR and residual achievement lends support to this assumption.

Inappropriate dependent measure. Second, it is possible that there were not treatment effects on student achievement because, in part, the dependent measure--a standardized achievement test--was inappropriate for evaluating a treatment of this kind. To be sure, standardized achievement tests typically have high reliability and adequately reflect prevailing curricular trends (e.g., Sax, 1974). But, as Berliner (1977) has argued, these tests may have poor content validity at the classroom level. Further, because they often correlate substantially with measures of general intelligence, test items may not be very reactive to instruction. "Off-the-shelf standardized tests," Berliner (1977) contended, "make poor dependent variables for studies of teaching" (p. 148).

The principal investigators were prevented by the school district's central administration from administering a specially constructed achievement test to the students. Their concern, justifiably, was to avoid excessive testing. Consequently, the CTBS was used as the measure of achievement--a test that was routinely administered to the students as part of the school district's regular testing program.

While a specially constructed test may be more reactive to classroom instruction and, hence, a more appropriate dependent measure in research on teaching there is ample precedent for the fruitful use of standardized achievement tests in such research. For example, the large-scale correlational studies conducted by Brophy and Evertson (1974), McDonald and Elias (1976), Soar (1973), and Stallings and Kaskowitz (1974) employed standardized achievement tests as dependent measures. And each of these studies yielded substantive findings concerning process-product relationships. Further, three of the four classroom-based experiments employed standardized achievement tests as the dependent measure; each experiment obtained positive results.

An interesting finding related to this issue was reported by Good and Grouws (1979). In addition to administering a subtest of a standardized achievement test, they administered a "content test" specially constructed for the particular school district in which their study was conducted. While there was a strong and positive treatment effect on achievement as measured by the standardized test, there was not a comparable treatment effect on achievement as measured by the content test (although the mean difference was in favor of the experimental group). Ostensibly, the former was more sensitive, or reactive, to the treatment than the latter. (There was, however, a possibility of a ceiling effect on the content test; subsequent analyses should clarify this resu t.)

There is evidence, then, to support the use of standardized achievement tests in this kind of research. Perhaps, though, the issue ought not be phrased in an either-o" fashion; rather, the choice of a dependent measure should be made in view of the articulated goals of the intervention. If one hypothesizes that an intervention should improve student knowledge of the concepts, principles, and processes held in common by many curricula, a standardized achievement test would appear to be an appropriate criterion. If, in contrast, the hypothesized effects of an intervention calls for a measure that is much more sensitive to the instructional goings-on of the particular classroom, a specially constructed test would probably be more suitable. Both, however, can yield useful and complementary information concerning the effects of an intervention; perhaps an intervention is best evaluated by employing both, rather than one or the other (Sax, 1974, p. 261).

Summary. This section is concluded as it was begun. Although there are several ostensibly plausible reasons for the absence of treatment effects on student achievement, the most compelling reason for these results appears to be poor treatment implementation: Training-related behavior among experimental-group teachers simply was not modified enough to effect appreciable change in subsequent student achievement.

Recommendations for Subsequent Research

The results of the present study call into question the effectiveness of a minimal intervention. It would appear that, for an intervention to be successful, the project staff has to be "engaged" with the participating teachers in some fashion--for example, through meetings and frequent classroom observations.

The previous research, however, does not provide clear implicatons concerning the relative contributions of holding meetings and conducting classroom observations. Needed are studies that incorporate these features into the design systematically. That is, "meetings with teachers," and "classroom observations" would be design factors and independently manipulated. If these factors were dichotomous--e.g., one introductory meeting versus several workshops; a few brief observations versus frequent and lengthy observations--each experimental-group teacher would be randomly assigned to one of the four possible combinations. In addition to examining the main effect of each factor on outcome, one could examine possible interactions. Perhaps a small number of observations combined with several workshops produces maximum treatment implementation and, in turn, the largest increments in student achievement.

Further, one must consider contextual factors that may limit the feasibility of a minimal intervention. Some contextual factors could be incorporated into the design to examine possible main effects and interactions. It is possible, for example, that "class SES" and "meetings with teachers" interact in their effects on outcome. Teachers of lcw-SES classes may require relatively frequent and intensive meetings with project staff concerning the training program and its applicability to their particular teaching environment. Teachers of middle-SES classes, in contrast, might be able to profit from relatively few and brief meetings.

The Promise of the Minimal Intervention in Research on Teaching

Good and Grouws (1979) argued that their findings, along with those of Anderson et al. (1979) and Crawford et al. (1978), indicated that classroom-based experiments

> are capable of yielding improvements in student learning that are practically as well as statistically significant. Such data are an important contradiction to the frequently expressed attitudes that . . . brief, inexpensive treatments cannot hope to bring about significant results. (p. 361)

The results of this study should serve to temper such optimism concerning the promise of the minimal intervention in research on teaching.

APPENDIX A

Procedures for Plotting the Line D(X) And Constructing a 95% Simultaneous Confidence Band

Procedures are outlined here for plotting the line D(X) and constructing a 95% simultaneous confidence band. (These procedures are adapted from Rogosa [1980].) The pretest (X) and posttest (Y) of the CTBS total score are used as an example.

Plotting the Line D(X)

The first task in plotting the line D(X), of course, is to define the line. The line D(X) is defined as $D(X) = b_2 + b_4X$ or, in the case of the CTBS total score, $D(X) = -9.394 + .171(X)$ (see Table 12).

This line is plotted against the X and Y axes. The range of data for X, the pretest, is 40.7 to 59.6. The Y axis, scaled in the units of the posttest, reflects the vertical distance of the two sample within-group regression lines and intersects the X axis at Y = 0. Thus, the difference between the two regression lines is zero at the point at which the line D(X) intersects the X axis.

By using the equation $D(X) = -9.394 + .171(X)$, one can determine D(X) for the minimum (40.7) and maximum (59.6) values obtained for X. The two resulting points--(40.7, -2.43) and (59.6, .80)--are plotted and joined (see Figure 11).

Constructing a 95% Simultaneous Confidence Band

Additional statistics are needed to construct such a band. The weighted average of the two group means is denoted C_a. (C_a is the point at which the ancova estimate of the treatment effect is evaluated.) $C_a = -s_{24}/s_{44}$, where s_{24} is the covariance of b_2 and b_4 and s_{44} is the variance of b_4 (see Table 13). For the CTBS total score, $C_a = 2.0271/.0405 = 50.052$.

$D(C_a)$ is the difference of the two sample within-group regression lines evaluated at C_a. That is, $D(C_a)$ is the $D(X)$ at C_a (which is equivalent to the ancova estimate of the treatment effect, or the adjusted mean-difference on $\underline{Y}$.) The estimated variance of $D(C_a)$ is denoted $s^2_{D(C_a)}$ and is equal to $s_{22} + s_{24}C_a$. The new term, s_{22}, is the variance of b_2 (see Table 13). Thus, in the present example, $s^2_{D(C_a)} = 102.5227 - 2.0271(50.052) = 1.062$.

Also needed to construct a simultaneous confidence band for the line $D(X)$ is the estimated variance of $D(X)$, or $s^2_{D(X)}$: $s^2_{D(X)} = s^2_{D(C_a)} + s_{44}(X - C_a)^2$. For $\underline{X} = 40.7$, for example, $s^2_{D(X)} = 1.062 + .0405(40.7 - 50.052)^2 = 4.604$. This variance is calculated similarly for other values of $\underline{X}$.

A $100(1 - \alpha)$ percent simultaneous confidence band for the line $D(X)$ consists of the area in the X,Y plane that is enclosed by the upper and lower hyperbolae

$$D(X) \pm \sqrt{2F^{\alpha}_{2,N-4}\, s^2_{D(X)}} \quad . \tag{5}$$

These hyperbolae can be constructed by using Equation 5 for successive values of $\underline{X}$. Here, $F^{.05}_{2,23} = 3.44$. (A degree of freedom was subtracted to "adjust" for a team-taught class.) The interval for $D(X)$ at $X = 40.7$, then, is $D(X) \pm \sqrt{2(3.44)(4.60)} = -2.43 \pm 5.63$. Equation 5 is employed for as many values of X necessary to detect the shape of the hyperbolae (see Figure 11). (Thus, $s^2_{D(X)}$ need only be determined for values of X for which Equation 5 is employed.)

APPENDIX B

Adjusting Statistics for Weighting Class Means

The following ratio was used to adjust certain statistics for weighting the class means:

$$\frac{M - k - 1}{N - k - 1}$$

where the actual number of classes is denoted by $\underline{M}$, $\underline{k}$ refers to the number of predictors in the regression, and $\underline{N}$ represents the total number of students. For any adjustment in the present context, $\underline{M} = 27$ and $\underline{N} = 631$.

$\underline{k} = 1$ for adjusting statistics corresponding to the within-group regressions (Table 11), because there is one predictor (i.e., the pretest, $\underline{X}$). For adjusting statistics associated with the J-N regressions (Tables 12 and 13), $\underline{k} = 3$ (i.e., $\underline{T}$, $\underline{X}$, and $\underline{TX}$).

This ratio was used to adjust three statistics: the standard error for a regression coefficient (which was divided by the square root of the adjusting ratio), the $\underline{F}$ ratio (multiplied by the adjusting ratio), and the elements of the variance-covariance matrix for the J-N regression (divided by the adjusting ratio).

APPENDIX C

The Procedure for Constructing a
95% Confidence Interval for the
Ancova Estimate of the Treatment Effect

Outlined here is the procedure for deriving a 95% confidence interval for the ancova estimate of the treatment effect. (This procedure is discussed in greater detail in Rogosa [1980].) The CTBS total score is used as an example.

The ancova estimate of the treatment effect is equivalent to the difference of the two sample within-group regression lines evaluated at C_a, where C_a is the point on $\underline{X}$ corresponding to the weighted average of the two group means. Thus, the ancova estimate is the D(X) at C_a or, equivalently, $D(C_a)$. A 100(1 - α) percent confidence interval for $D(C_a)$ is bounded by the endpoints

$$D(C_a) \pm \sqrt{F^{\alpha}_{1,N-4}\ s^2_{D(C_a)}}$$

where $s^2_{D(C_a)}$ is the estimated variance of $D(C_a)$. (See Appendix A for further discussion of $s^2_{D(C_a)}$.)

As reported in Table 14, $D(C_a)$ for the CTBS total score is -.866. $s^2_{D(C_a)}$ = 1.062 (see Appendix A) and $F^{.05}_{1,23}$ = 4.28. (A degree of freedom was subtracted to "adjust" for a team-taught class.) Thus, the 95% confidence interval for $D(C_a)$ is $-.866 \pm \sqrt{(4.28)(1.062)} = -.866 \pm 2.132$.

In contrast to the conventional procedure for constructing a confidence interval for the ancova treatment effect, the procedure outlined here does not require the assumption that $\beta_4 = 0$. Consequently, when the test statistic for the null hypothesis $\beta_4 = 0$ (i.e., $b_4^2\ /s_{44}$) is greater than 1, the latter procedure results in a narrower confidence interval.

References

Anderson, L., Evertson, C., & Brophy, J. An experimental study of effective teaching in first-grade reading groups. Elementary School Journal, 1979, 79, 193-223.

Berliner, D. Impediments to measuring teacher effectiveness. In G. Borich, The appraisal of teaching: Concepts and process. Reading, MA: Addison-Wesley, 1977, pp. 146-161.

Berliner, D. Tempus educare. In P. Peterson & H. Walberg (Eds.), Research on teaching: Concepts, findings, and implications. Berkeley, CA: McCutchan, 1979, pp. 120-135.

Berliner, D., & Rosenshine, B. The acquisition of knowledge in the classroom. In R. Anderson & W. Montague (Eds.), Schooling and the acquisition of knowledge. Hillsdale, NJ: Erlbaum, 1977, pp. 375-396.

Blank, M. Teaching and learning in the preschool: A dialogue approach. Columbus, OH: Charles E. Merrill, 1973.

Brophy, J., & Evertson, C. Process-product correlations in the Texas Teacher Effectiveness Study: Final Report. Research Report No. 74-4. Austin, TX: Research and Development Center for Teacher Education, University of Texas, 1974.

Brophy, J., & Putnam, J. Classroom management in the elementary grades. In D. Duke (Ed.), *Classroom management*. The 78th yearbook of the National Society for the Study of Education, Part 2. Chicago: University of Chicago Press, 1979, pp. 182-216.

Calfee, R. & Pessirilo-Juristic, G. *Perceived changes in California schools and classrooms*. Stanford, CA: Institute for Research on Educational Finance and Governance, 1979.

Charters, W., & Jones, J. On the risk of appraising non-events in program evaluation. *Educational Researcher*, 1973, *11*(2), 5-7.

Coladarci, T. *A classroom-based experiment assessing the impact on teacher behavior and student achievement of a direct-instruction approach to teaching.* Unpublished doctoral dissertation, Stanford University, 1980.

Crawford, J., Gage, N., Corno, L., Stayrook, N., Mitman, A., Schunk, D., Stallings, J., Baskin, E., Harvey, P., Austin, D., Cronin, D., & Newman, R. *An experiment on teacher effectiveness and parent-assisted instruction in the third grade* (3 vols.). Stanford, CA: Center for Educational Research at Stanford, 1978.

Crawford, J., & Stallings, J. *Experimental effects of in-service teacher training derived from process-product correlations in the primary grades*. Paper presented at the annual meeting of the American Educational Research Association, Toronto, 1978.

Cronbach, L. Essentials of psychological testing (3rd ed.). New York: Harper & Row, 1970.

Cronbach, L. Research on classrooms and schools: Formulation of questions, design, and analysis. Stanford, CA: Stanford Evaluation Consortium, July 1976.

Cronbach, L., Gleser, G., Nanda, H., & Rajaratnam, N. The dependability of behavioral measures: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Cronbach, L., & Snow, R. Aptitudes and instructional methods: A handbook for research on interactions. New York: Irvington, 1977.

Cronbach, L., & Webb, N. Between-class and within-class effects in a reported Aptitude x Teacher interaction: Reanalysis of a study by G. L. Anderson. Journal of Educational Psychology, 1975, 67, 712-727.

Doyle, W., & Ponder, G. The practicality ethic and teacher decision-making. Interchange, 1977, 8(3), 1-12.

Ebmeir, H., & Good, T. The effects of instructing teachers about good teaching on the mathematics achievement of fourth grade students. American Educational Research Journal, 1979, 16, 1-16.

Elashoff, J. Analysis of covariance: A delicate instrument.

American Educational Research Journal, 1969, 6, 383-402.

Gage, N., & Coladarci, T. Replication of an experiment with a field-based in-service teacher education program: Final Report. Stanford, CA: Center for Educational Research at Stanford, January 1980. (ERIC Document Reproduction Service No. ED181023)

Good, T. Research on teaching. Paper presented at the national invitational conference, "Exploring Issues in Teacher Education: Questions for Future Research." University of Texas, Austin, 1979. (a)

Good, T. Teacher effectiveness in the elementary school. Journal of Teacher Education, 1979, 30, 52-64. (b)

Good, T., & Grouws, D. Teacher effects: A process-product study in fourth-grade mathematics classrooms. Journal of Teacher Education, 1977, 28, 49-54.

Good, T., & Grouws, D. The Missouri Mathematics Effectiveness Project: An experimental study in fourth-grade classrooms. Journal of Educational Psychology, 1979, 71, 335-362.

Hedges, L. Combining the results of experiments using different scales of measurement. Unpublished doctoral dissertation, Stanford University, 1980.

Kirst, M. The new politics of state education finance. Phi Delta Kappan, 1979, 60, 427-432.

Kounin, J. Discipline and group management in classrooms.

New York: Holt, Rinehart & Winston, 1970.

Levy, G. Ghetto school. New York: Pegasus Books, 1970.

Linn, R., & Slinde, J. The determination of the significance of change between pre- and posttesting periods. Review of Educational Research, 1977, 47, 121-150.

McDonald, F., & Elias, P. The effects of teaching performance on pupil learning. Beginning Teacher Evaluation Study: Phase II, Final Report (Vol. 1). Princeton, NJ: Educational Testing Service, 1976.

Mendro, R. A Monte-Carlo study of the robustness of the Johnson-Neyman technique. Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

Mohlman, G., Coladarci, T., & Gage, N. Comprehension and attitude as predictors of implementation of teacher training. Paper presented at the annual meeting of the American Educational Research Association, Boston, 1980.

Nie, N., Hull, C., Jenkins, J., Steinbrenner, K., & Bent, D. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Powell, M. Research on teaching. The Educational Forum, 1978, 43, 27-37.

Rogosa, D. Some results of the Johnson-Neyman technique. Unpublished doctoral dissertation, Stanford University, 1977.

Rogosa, D. Comparing nonparallel regression lines. Psychological Bulletin, 1980, 87, 307-321.

Rogosa, D. On the relationship between the Johnson-Neyman region of significance and statistical tests of parallel within-group regressions. Educational and Psychological Measurement, 1981, in press.

Rosenshine, B. The third cycle of research on teacher effects: Content covered, academic engaged time, and direct instruction. In P. Peterson & H. Walberg (Eds.), Research on teaching: Concepts, findings, and implications. Berkeley, CA: McCutchan, 1979, pp. 28-56.

Rosenshine, B., & Berliner, D. Academic engaged time. British Journal of Teacher Education, 1978, 4, 3-16.

Rosenthal, R., & Rosnow, R. The volunteer subject. New York: Wiley, 1975.

SAS Institute Inc. SAS user's guide (1979 ed.). Raleigh, NC: SAS Institute, Inc., 1979.

Sax, G. The use of standardized tests in evaluation. In W. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974, pp. 243-308.

Serlin, R., & Levin, J. Identifying regions of significance in aptitude-by-treatment-interaction research. American Educational Research Journal, 1980, 17, 389-399.

Soar, R. Final Report: Follow Through classroom process measurement and pupil growth (1970-1971). Gainesville, FL: Institute for the Development of Human Resources, University of Florida, 1973.

Southwest Educational Development Laboratory. Bilingual Kindergarten Program Inservice Manual (Vol. I). Austin, TX: National Education Publishers, 1973.

Stallings, J., Corey, R., Fairweather, J., & Needels, M. The study of basic reading skills taught in secondary schools. Menlo Park, CA: SRI International, 1978.

Stallings, J., & Kaskowitz, D. Follow Through classroom observation evaluation (1972-1973). Menlo Park, CA: SRI International, 1974.

Stallings, J., Needels, M., & Stayrook, N. How to change the process of teaching basic reading skills in secondary schools: Phase II and Phase III, Final Report. Menlo Park, CA: SRI International, 1979.