DOCUMENT RESUME

ED 129 914 95 TM 005 788

AUTHOR Solomon, Warren; And Others
TITLE The Development, Use, and Importance of Instruments that Validly and Reliably Assess the Degrees to Which Experimental Programs Are Implemented.
INSTITUTION Central Midwestern Regional Educational Lab., St. Ann, Mo.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
NOTE 26p.

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Classroom Observation Techniques; Comparative Analysis; *Experimental Programs; Formative Evaluation; Preschool Education; *Preschool Programs; *Program Development; Program Effectiveness; *Program Evaluation; Rating Scales; Test Construction; Test Reliability; Test Validity
IDENTIFIERS *DARCEE Classroom Assessment Scale; DARCEE Program; Demonstration and Research Center Early Education

ABSTRACT

This study develops and tests an instrument to assess the fidelity of the intended program, i.e., experimental treatment in the evaluation of a preschool program. During the school year (1972-73) CEMREL (Central Midwestern Educational Lab) investigated the consequences of different levels of training on implementation of the Demonstration and Research Center for Early Education (DARCEE) program,. Part of this investigation involved three separate ratings of the pilot test classrooms with the assessment scale. These ratings were given at the beginning, middle, and end of the school year. Classrooms with maximum training scored on the average approximately 10 per cent higher on each of the essentials than did the classes with materials only. With comparison classes, however, that consistency was lacking. On the essentials of physical setting, unit use, and parent involvement the comparison classrooms actually scored higher than the DARCEE group with the maximum training, on two other essentials (reinforcement and behavior management and attitude development) and on student involvement they scored higher than DARCEE classrooms with the minimum training, whereas on the essentials of skill development, organization and use of time, grouping, teacher roles and responsibilities, and teacher preparation these classes scored lower than both DARCEE classroom treatments. (Author)

THE DEVELOPMENT, USE, AND IMPORTANCE OF INSTRUMENTS THAT VALIDLY AND RELIABLY ASSESS THE DEGREE TO WHICH EXPERIMENTAL PROGRAMS ARE IMPLEMENTED[1]

Warren Solomon, Daniel Ferritor,
Joseph Haern, Edwin Myers
CEMREL, Inc.

Over the past several years we have witnessed an almost exponential rise in intervention programs, curriculum materials, and special training programs designed to facilitate cognitive, perceptual, psychomotor, and social-emotional development in home and school settings. Simultaneously, there has been a similar rise in the quantity as well as the quality of educational program evaluations. In many of these evaluations we find increased attention focused on the assessment of specific child outcomes targeted by the program, or materials within them.

When evaluations deal primarily with the sets of expected child outcomes derived from the program objectives, the interpretations appear to be relatively straightforward. That is, the evaluator can state that the

---

[1]Prepared under the auspices of CEMREL, Inc., a private non-profit corporation supported in part as an educational laboratory by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed in this publication do not necessarily reflect the position or policy of the National Institute of Education. No official endorsement should be inferred. The authors wish to express their appreciation to Dr. Thomas Johnson for his help throughout the development of the instrument and in the preparation of this manuscript. We also wish to express our appreciation to Dr. Paul Dokecki and the DARCEE training staff for their help in the development of the instrument.

program's materials, strategies, and training procedures were carred out in an effort to attain a given set of objectives and, in fact, attained a certain percentage of those objectives. Based on statements such as these, one is tempted to draw conclusions on whether or not the treatment was efficacious. "Efficacious treatments" could be defined, for example, as ones in which 50, 60, or possibly 70 per cent of the program's objectives were attained. In point of fact, this hypothetical evaluation strategy may lead to erroneous conclusions (Gross, Giaquinta, & Bernstein, 1971, pp. 3-7). The critical factor may be the "were carried out" or implementation dimension. The fact that materials and strategies were prescribed does not guarantee that the teacher actually engaged children in the intended way with the program's set of curriculum materials. If throughout the year the teacher did not implement a particular aspect of the program, it is misleading to say that the program is one that is not able to attain its objectives. Perhaps, one is equally justified in suggesting that the program's training was carried out poorly.

This is not to say that the development and use of tests based on program objectives are unimportant in program evaluations. The argument, rather, is that such program evaluations are incomplete (Stake, 1967, p. 5) and may suggest unwarranted causal relationships between treatment and hypothesized outcomes. What is also necessary is an examination of the extent to which the program actually was implemented.

The objective of this study was to develop and test an instrument to assess the fidelity of the intended program, i.e., experimental treatment

in the evaluation of a preschool program.[2] With such an instrument the evaluator could look carefully at the program as it is installed in different sites to examine the degree to which each of the independent or regulating variables defined by the developer as major components are present. With the knowledge of the presence or absence of these variables, it is possible to conduct a thorough and fair evaluation of the program's ability to attain the outcomes it seeks, as well as to evaluate the training component of the program. With such data one might find that a program that apparently attained only 50 per cent of its prescribed objectives, in fact attained 90 per cent of the objectives that teachers actually chose and attempted to attain. Such information has relevance not only for summative evaluations and for comparative analyses of the program effectiveness, but for formative evaluations, to provide data to program developers an possible revisions in program specifications and training procedures.

Why Develop a Degree of Implementation Instrument?

Some argue that the most economical way to assess how well a program is implemented would be to make use of an existing instrument. If we imagine, for example, that a program has as a component the prescription that the teacher teach indirectly, then the interaction analysis system of Flanders (1960, pp. 257-265) could be used to help determine the degree of implementation. Or, if the program prescribes that teachers ask many questions that call for divergent thinking responses, the interaction analysis system of

[2] The particular preschool program we were evaluating was one developed at Peabody College by the Demonstration and Research Center for Early Education (DARCEE). In this document the program will be called "The DARCEE Program."

Gallagher and Aschner (1968, pp. 219-133) would seem appropriate.

Unfortunately, as we attempted to assess degree to which teaching teams in classrooms in selected sites were implementing the major independent variables of the DARCEE preschool program,[3] none of the existing observational systems served our needs. While they might have provided some interesting research data, they would not have shed light on the implementation questions felt to be critical in the evaluation. Our solution was to develop a new instrument that would answer these questions. In particular, we wanted an instrument that could assess to what extent teachers were implementing the entire preschool program.

Measuring the total program implementation allows the evaluator to gain information relevant to a number of issues. First, he can determine the extent to which the objectives of the teacher training materials and procedures are realized. Second, he can use the instrument to serve the formative evaluation role of helping trainers use data to examine their training priorities. Third, he can determine which of the program variables are harder to implement than others. And, finally, he can determine which program variables are most important in attaining child outcomes. That

---

[3]DARCEE's program process variables have been named by DARCEE the program's "essentials." The essentials are sets of prescriptions organized about themes that specify how the teacher is to organize the space, time, groupings, and content and specifies how the teacher is to interact with team members and children. One set of prescriptions focused on organization of space is called "the physical setting"; another set of prescriptions focused on how the teacher is to interact with children is called "behavior management and positive reinforcement," and so on. DARCEE specified ten essentials in the 1971-72 school year and eleven in the 1972-73 school year. The latter set includes "physical setting," "organization and use of time," "grouping," "teacher roles and responsibilities," "teacher preparation," "materials use, " "attitude development," "behavior management and positive reinforcement," "skill development," "unit use," and "parent involvement."

is, he can distinguish between those independent variables having major effects as opposed to minor or negligible effects on the desired child outcomes.

Our decision to develop a new instrument was not unique. Gross, Giaquinta, and Bernstein (1971) developed such an instrument in a study of the installation of an innovative program in an elementary school, as did Oliver and Shaver (1966), when they investigated two styles of teaching (socratic analysis and recitation analysis) in their social studies curriculum project.

Development of the Instrument

There were several phases in the development of "The DARCEE Classroom Assessment Scale." First, we became familiar with the DARCEE program by reading DARCEE documents describing the program, by observing the program in operation in many sites, by participating in DARCEE training workshops, and by discussing the DARCEE program at length with DARCEE developers and trainers. This phase may be called the "program familiarization phase."

During the second phase, the "instrument development phase," we used the description of DARCEE's ten "essentials" (Brown, Dokecki, O'Connor, & Stinson, 1971) and sorted the 55 items of a classroom checklist previously developed by DARCEE using each DARCEE essential as a category. New items were then written, and vague items of the original checklist were clarified to make possible reliable scoring.

The third phase of the instrument development could be called the "instrument refinement phase." After the first version was drafted, a meeting was held for the development staff to examine and critique the

instrument. Then, following completion of a second draft of the instrument, members of the CEMREL staff and DARCEE training staff observed the development site classroom, as well as classrooms in Mille Lacs, Minnesota and Macon, Georgia. At least two observers scored the same classroom at the same time in an effort to determine the interscorer agreement and further specify items to make them more reliable. Interscorer agreement scores (percentages of agreement) were 70.5 per cent and 83.3 per cent in two Mille Lacs classrooms (February 1972) and were 94.4 per cent and 89.3 per cent in two Macon classrooms (April 1972).

Later, in the spring of 1972, after formulas for computing subscores that correspond to DARCEE essentials were developed, all DARCEE classes and four non-DARCEE classes were visited and scored using the assessment scale. The findings, summarized below, reveal that the instrument was sensitive to differences between DARCEE and non-DARCEE classrooms on many of the subscores.

The assessment scale was further revised during the summer of 1972 to make the instrument one that raters unfamiliar with the DARCEE preschool program could use. In this revised version, the items include much more descriptive information, terms are defined more precisely, and scoring instructions are detailed.

From the above description of how the assessment scale was developed, it is clear that (a) the instrument was developed after evaluators studied the program, (b) that the instrument has been refined using recommendations of program developers as inputs in an effort to get content validation, and (c) the instrument has been modified based on field tests of its use to increase the reliability of items.

Nature of the Implementation Instrument

The implementation instrument was designed to assess the extent to which various preschool classrooms resemble the ideal DARCEE classroom as defined by DARCEE developers. The instrument consists of 95 items which utilize three different measurement strategies: (a) some items are scored by observing each of the teachers as they interact with the children or by observing the displays and physical arrangements of the classroom, (b) some items are scored by examining documents written by the teachers, and (c) some items are scored by rating responses made by the teachers when interviewed. Figures 1, 2, and 3 show examples of these three scoring techniques.

Whatever form of measurement was employed on any given item, each item is scored on a three-point scale ranging from 0 to 2 with 0 representing non-correspondence with the ideal DARCEE classroom, and 2 representing correspondence with the DARCEE classroom. Scores on each item contribute to one or more subscores which correspond to specific DARCEE essentials. By collapsing the 95 items into subscores, one may examine each classroom with regard to the extent to which each DARCEE essential is being implemented in the classroom. The subscores are then summarized on a chart showing the classroom profile. Figure 4 is an example of the summarizing profile.

The assessment scale requires one full day of classroom observation starting before the beginning of class and ending only after the teachers have completed their daily planning and evaluation meeting, which usually occurs after the children have departed for the day. To assure the content validity of the subscores, the items and subscore formulas were

Figure 1 Sample Observation Item

Completeness of the Schedule

The DARCEE classroom schedule includes a number of specified kinds of activities that are to recur each day. Since the schedules are usually posted on the wall, you will usually need only look at the posted schedule to score this item. If the schedule is not posted, you could simply take note of activities that occur as they occur and check them off on the score sheet. The activities that should recur daily are:

a. At least one large-group activity. [In large group, the entire classroom of children sit together to receive instrumtion conducted usually by the lead teacher.]

b. One small-group activity. [Small-group activities are conducted by teachers teaching groups of four to ten children.]

c. A second small-group activity. [The description for "b" applies here.]

d. Structured free choice. [Children are given a period of time to participate in an activity or activities they have chosen from a limited number of options.]

e. Meals and/or snacks.

f. Toileting and washing hands.

g. Outdoor activity. [Weather permitting, children have some time during the daily session to go outside to play. If there is inclement weather, they have some substitute activity, usually active games.]

h. A second large-group meeting near the end of the day.

On the score sheet, check which of the above items are part of the daily schedule. Then, score:

(0) If TWO or MORE of the above ITEMS are OMITTED.

(1) If ONE of the above ITEMS is OMITTED.

(2) If NONE of the above ITEMS is OMITTED.

Figure 2 Sample Item Scored by Analyzing Written Records

The Number of Lesson Plans

In the DARCEE classroom prior to the daily session the Lead Teacher should have prepared a lesson plan for large group, and she and her Assistants should have prepared lesson plans for all of their small groups. The definition of "lesson plan" for this item is as follows: The lesson plan must be a statement in writing of (a)at least one objective and (b) at least one material and strategy to be used for the samll- or large-group activity session.

To score this item, collect all lesson plans and eliminate those that do not meet criteria (a) or (b). Then, score:

(0) If 2 OR FEWER LESSON PLANS WERE WRITTEN prior to teaching by all of the teachers.

(1) For situations between (0) and (2).

(2) If ALL OR ALL BUT ONE of the POSSIBLE LESSON PLANS WERE WRITTEN PRIOR TO TEACHING. [To figure out how many lesson plans are possible, assume that each teacher should have one lesson plan for each small group he or she teaches and that in addition the Lead Teacher should have a lesson plan for her large-group session. For example, in Mrs. Keller's room there are two small-group activity sessions and two teachers, including Mrs. Keller. Under those circumstances there should be four small-group activity lesson plans plus one large-group activity lesson plan, or a total of five lesson plans. If there were an additional Assistant Teacher as part of Mrs. Keller's team, two additional small-group lesson plans should be prepared, making a total of seven lesson plans.]

Figure 3 Sample Item Scored from Teacher Interview

Criteria for Grouping and Regrouping Children in Small Groups

In the DACCEE classroom each child is placed in a small group for daily instructional purposes, meals and/or snacks, and other reasons. Children are to be grouped and regrouped in their particular small groups on the basis of two principles: (a) ability (children are to be placed in groups with children having similar levels of skills) and (b) social factors (children are to be placed in groups of children with whom they are compatible. Some children high in certain behavior patterns, like following directions, may be placed in groups as role models for others to follow. Some children are placed in groups to separate them from children whose influence on their behavior is negative.)

To score this item ask the question in the box below:

> How did you place children in their small groups? [If the answer is too general to score, ask specific questions such as, "Why did you place Johnny in Miss Smith's group instead of Mr. Kelso's? Why did you place Annette in the group she is in?" etc.] Do you regroup your children? [If so] how do you decide which children to regroup?

Score:

(0) If the TEACHER indicated she CONSIDERED NEITHER (a) ABILITY FACTORS consistent with the DACCEE program or (b) SOCIAL FACTORS consistent with the DACCEE program (see the paragraph describing DARCEE grouping rules above).

(1) If she CONSIDERED REASON (a) OR (b) BUT NOT BOTH (a) and (b), OR [illegible] situations between (0) and (2).

(2) If she CONSIDERED (a) AND (b) both for grouping and re-

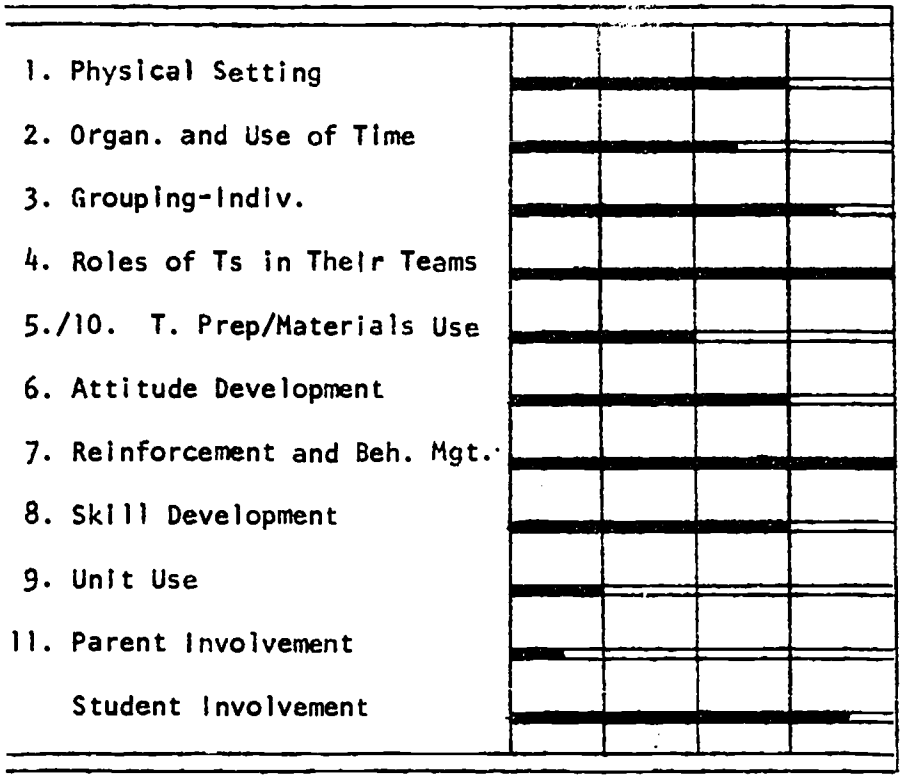
Figure 4 Sample Summary Profile for Classroom

Date of Observation 1/1/73 Time of Day: 8:30-2:30

Site: St. Louis

Teacher: Mr. Jones Rater: B. Stone

RATING FORM PROFILE



Subscore	Subscore Average	Proportion of Agreement Score
1. Physical Setting	1. 1.5	1. .75
2. Organ. and Use of Time	2. 1.25	2. .62
3. Grouping-Indiv.	3. 1.75	3. .87
4. Roles of Ts in Their Teams	4. 2.0	4. ___
5./10. T. Prep/Materials Use	5./10/ 1.0	5./10. 5.0
6. Attitude Development	6. 1.5	6. .75
7. Reinforcement and Beh. Mgt.	7. 2.0	7. 1.0
8. Skill Development	8. 1.5	8. .75
9. Unit Use	9. .5	9. .25
11. Parent Involvement	11. .25	11. .12
Student Involvement	SI 1.75	SI .87

examined and modified by the DARCEE preschool development and training staff.[4]

Before they are ready to use the instrument independently, raters need approximately a day and one-half of training, which includes supervised use of the instrument.

---

[4]It should be noted that the assessment scale itself is an instrument which measures the degree to which teachers behave in accordance with DARCEE teaching principles and as a result the instrument reflects DARCEE's assumption that teaching process variables are more important independent variables than content so far as child outcomes are concerned. As evaluators, we were not able to accept only measures of these process variables as the sole measurement of degree of implementation. That is, since there was an entire set of child objectives in the cognitive and skill domain, there should also be a degree of implementation measure on whether the teachers actually taught the content implied in the objectives. Therefore, the CEMREL evaluation staff not only developed the degree of implementation measure discussed in this paper, it also sought to determine the degree to which the teachers actually attempted to attain the program's child outcomes.

To measure this dimension, we, in conjunction with the DARCEE developers, designed and produced late in the first year of the field test an instrument which the teachers marked daily for each child with regard to whether they had attempted to teach particular objectives and whether they had been successful. This instrument was developed too late in the year to assess the degree to which the teachers attempted to teach the specified child objectives. Therefore, at the end of the year a questionnaire was designed and administered to each teacher focusing on whether or not she had attempted to teach each of the DARCEE behavioral objectives. In our causal model we felt that the DARCEE essentials would probably be major independent variables for attitude outcomes in the children and minor independent variables for skill objectives, whereas the work put in on the objectives themselves should be the major independent variables for the skill objectives.

The Reliability of the Instrument

Three types of reliability were obtained for the observation instrument. The first type, which we will call interscorer agreement or interrater reliability, concerns the agreements of different raters observing at the same classroom at the same time. This reliability is utilized to estimate the effectiveness of training raters to use the instrument. The instrument must be reliable in the sense that each rater will score similar events in the same way.

Table 1 presents the design utilized to obtain these reliabilities. The coefficients for each of the interscorer agreements in the design are presented in Table 2 as proportion of interscorer item by item agreements out of total possible agreements. These coefficients range from 68.4 to 97.8 per cent agreement and average to 85.05 per cent agreement.

The second type of reliability ~~is also~~ represented in Table 1. Denoted simply as reliability coefficient, this reliability measure refers to the consistency of the classroom over a short period of time. Two raters rated the same classrooms but on different days in close proximity. Thus, this coefficient assesses not only interrater reliability, but also the approximate representativeness of a given classroom day with any other day within the same time frame, such as a week.

The results of this reliability are presented in Table 3. Most of the subscores do not appear to be subject to daily classroom variation with the exception of the Organization and Use of Time subscore and the Teacher Preparation--Materials Usage subscore. Since in our evaluation

Table 1

DESIGN TO DETERMINE RELIABILITIES OF THE DARCEE CLASSROOM ASSESSMENT SCALE IN THE FALL OF 1972

Classroom[a]	Raters					Kind of Reliability
	1	2	3	4	5	
A	11/21	11/21	11/21	11/21		Interscorer Agreement
B	11/27	11/27	11/27			Interscorer Agreement
C	12/5	12/5	12/5			Interscorer Agreement
D	12/4	12/6				Reliability Coefficient
E	12/6	12/4				Reliability Coefficient
F				11/13	11/13	Interscorer Agreement
G				11/14	11/14	Interscorer Agreement
H				11/13	11/13	Interscorer Agreement
I				11/14	11/14	Interscorer Agreement

[a]Classrooms A-E are located in Louisville, Kentucky. Raters 1-3 are local residents trained by Rater 4, a CEMREL employee. Classrooms F-I are located in Macon, Georgia. Raters 4 and 5 are CEMREL employees who reside in St. Louis

Table 2

PROPORTION OF INTERSCORER ITEM AGREEMENTS USING THE DARCEE CLASSROOM ASSESSMENT SCALE DURING THE FALL OF 1972[a]

	Raters						
Classroom	1-2	1-3	1-4	2-3	2-4	3-5	4-5
A	70.5	68.4	69.5	77.9	82.1	84.2	
B	89.5	92.6		92.6			
C[b]	96.8	96.8		97.8			
F							87.4
G							[illegible]
H							78.9
I							82.1

[a]Scores represent proportion of items on which there was agreement to total possible agreements. Disagreements on a three-point scale could be one- or two-point disagreements. In no case were there more than three point disagreements out of the 95 possible chances.

[b]Classroom C is a non-DARCEE comparison classroom.

Table 3

RELIABILITY COEFFICIENTS FOR INSTANCES IN WHICH TWO RATERS EACH SCORED THE SAME CLASSROOM ON DIFFERENT DAYS IN THE SAME WEEK

Subscore	Classroom 1	Classroom 2
1. Physical Setting	---a	---a
2. Organization and Use of Time	.612	.802
3. Grouping	.988	.716
4. Roles of Teachers in Their Teams	.892	.870
5. Teacher Preparation-Materials Use	.364	.716
6. Attitude Development	.629	.780
7. Reinforcement and Behavior Mgt.	.693	.641
8. Skill Development	.969	.286
9. Unit Use	---a	---a
10. Parent Involvement	---a	.000
11. Student Involvement	.895	.847

[a]No correlations could be calculated when all items of this subscore had identical rating and, hence, no variance.

Table 4

AVERAGE PRODUCT MOVEMENT-COEFFICIENTS OF INTERSCORER AGREEMENTS BY SUBSCORES[a]

Subscore	Mean Correlation[b]	Range of Correlations	S.D. of Correlations
1. Physical Setting	.88	.58-1.0	.18
2. Organization and Use of Time	.94	.76-1.0	.10
3. Grouping	.92	.61-1.0	.11
4. Roles of the Teachers in Their Teams	.98	.91-1.0	.03
5. Teacher Preparation-Materials Use	.97	.88-1.0	.04
6. Attitude Development	.86	.59-.99	.14
7. Reinforcement and Behavior Mgt.	.65	.00-1.0	.36
8. Skill Development	.97	.87-1.0	.05
9. Unit Use	1.00	1.0-1.0	.00
10. Parent Involvement	.84	.00-1.0	.30
11. Student Involvement and Attentiveness	.88	.39-1.0	.19

[a]Coefficients of interscorer agreements on subscores are defined as the correlation coefficients between the items which constitute each subscore.

[b]The number of correlations averaged to form the mean correlation is 15.

strategy we plan to use the assessment scale at the beginning, middle, and end of the school year to determine the extent of implementation variation over the year, It is important to estimate which subscores are less susceptible to actual teaching content and other daily variations. Additional estimates of this variation will be obtained during the school year.

In computing the third type of reliability we looked at the reliability of items within each subscore. Periodically, CEMREL sends score sheets to the DARCEE training staff so that trainers might determine noncorrespondence of individual classrooms to DARCEE principles and practices. In order to determine the internal consistency of subscores, item by item correlations were computed for each possible pair of observers who rated the same classroom at the same time. The results of this analysis are presented in Table 4.

The only mean correlation below a readily acceptable level of .80 was for the subscore on Reinforcement and Behavior Management. This subscore is the most difficult to rate, especially on the items dealing with setting standards and with reinforcement tallies. Apparently, raters see and rate approximately the same thing for each of the other subscores.

Results

Analysis of 1971-72 DARCEE Classrooms Using the Assessment Scale

In spring of 1971-72 each DARCEE classroom and four non-DARCEE classrooms were observed using the assessment scale. Since each class was observed only one time and since the assessment scale was still being revised the results in Table 5 should be regarded tentatively. They are

Table 5

COMPARISON OF SUBSCORES FOR DARCEE AND NON-DARCEE CLASSROOMS OBSERVED DURING THE SPRING OF 1972

Subscore	DARCEE Classrooms (N=15)		NON-DARCEE Classrooms (N=4)	
	Mean Subscore	Standard Deviation	Mean Subscore	Standard Deviation
1. Physical Setting	.80[a]	.16	.43	.08
3. Grouping	.84[a]	.12	.63	.14
3. Planning and Evaluation	.46[a]	.16	.13	.18
4. Teacher Roles and Responsibilities	.66[a]	.10	.16	.07
5. Organization and Use of Time	.85	.23	.69	.23
6. Unit Approach	.55	.28	.58	.41
7. Teaching Techniques	.71[a]	.13	.54	.16
8. Parent Involvement	.29	.28	.02	.07
9. Student Participation	.86	.12	.84	.21
Mean of Subscores	.67[a]		.44	

NOTE: Subscores are reported here as proportions of agreement scores with 0 representing non-correspondence with DARCEE and 1 representing correspondence. These scores were obtained simply by dividing the actual subscores by 2.

[a]On a 1-tailed t test for groups with independent means DARCEE classes scored significantly higher than non-DARCEE classes on these scores ($p < .05$).

shown mainly to indicate how the instruments could be used to analyze the degree of implementation.

As Table 5 indicates the DARCEE classes did show greater correspondence with DARCEE principles and procedures. Examination of subscores reveals, however, that the DARCEE classes did not significantly exceed non-DARCEE classes on all subscores. They did exceed the non-DARCEE classes on five subscores ($p. < .05$) (physical setting, grouping, planning and evaluation, teacher roles and responsibilities, and teaching techniques). On the parent involvement subscore the DARCEE classrooms were rated higher than the non-DARCEE class but this difference was not significant. Of the three other subscores (organization and use of time, unit approach, and student participation) there appeared to be only small differences between the DARCEE and non-DARCEE classes. As Table 5 shows, four subscores were implemented at a level of .80 or higher, whereas only two subscores were implemented at a level lower than .50. Examination of the particular subscores involved reveals that the DARCEE teachers had greatest success in obtaining student participation, setting up daily schedules, organizing their classroom space, and grouping their children in a manner consistent with DARCEE's prescriptions, they had moderate success in assuming appropriate DARCEE roles and responsibilities, in using DARCEE teaching techniques and the unit approach, and they had least success in planning and evaluation and parent involvement.

Preliminary Analysis of Training

During the current school year (1972-73) CEMREL is investigating the consequences of different levels of training on implementation of the

DARCEE program (Johnson, 1973). Part of this investigation involves three separate ratings of the pilot test classrooms with the assessment scale. These ratings are given at the beginning, middle, and end of the school year. Figure 5 shows the beginning of year average ratings for four classes with maximum DARCEE training (preservice, inservice, home visitor, and training materials), for five classes with minimum training (materials only), and for eight comparison classes with no DARCEE training. The data in Figure 5 were collected during the first administration of the assessment scale.

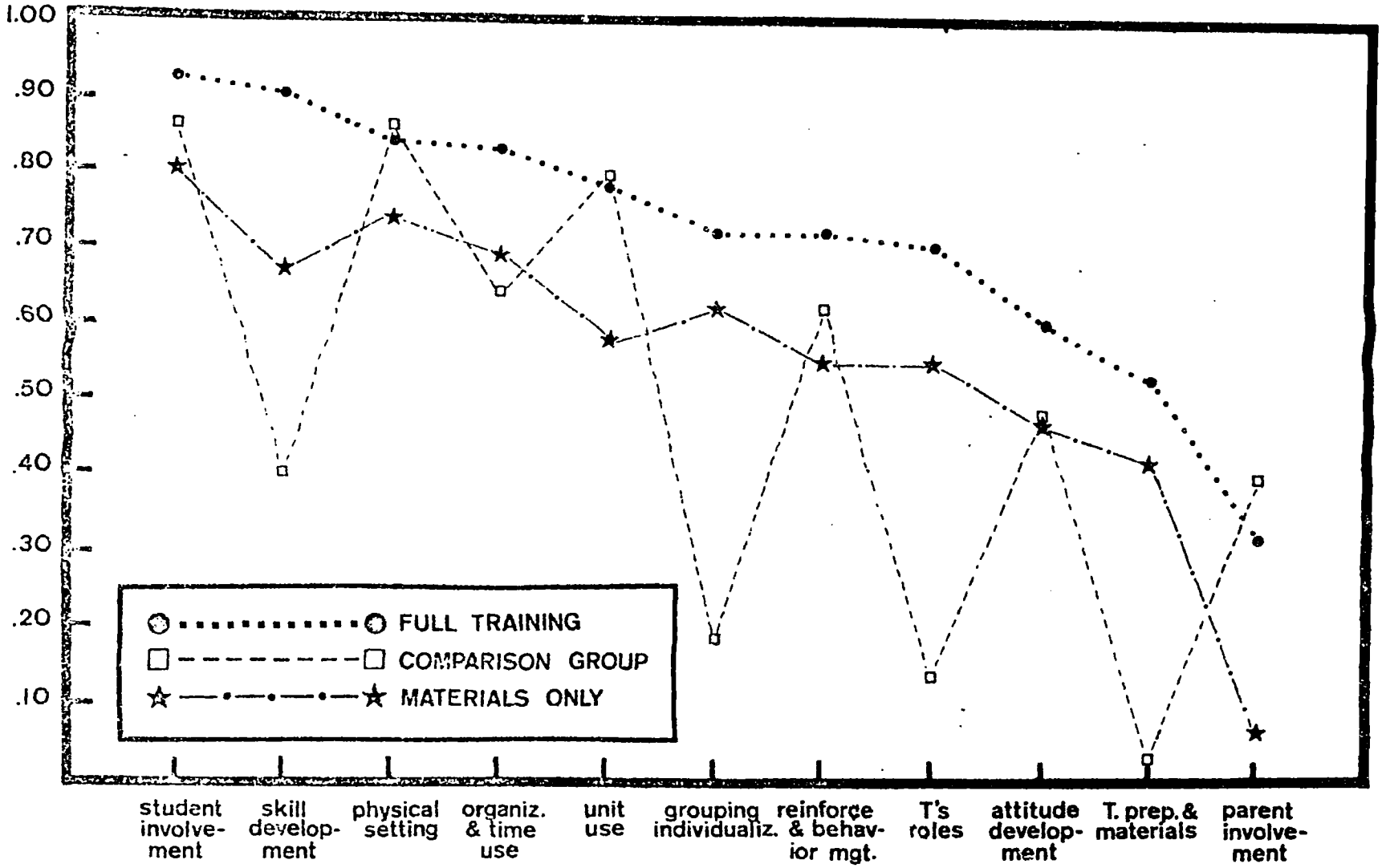
As Figure 5 shows, classrooms with maximum training scored on the average approximately 10 per cent higher on each of the essentials than did the classes with materials only. With the comparison classes, however, we don't find the same consistency. On the essentials of physical setting, unit use, and parent involvement the comparison classrooms actually scored higher than the DARCEE group with the maximum training, on two other essentials (reinforcement and behavior management and attitude development) and on student involvement they scored higher than the DARCEE classrooms with the minimum training, whereas on the essentials of skill development, organization and use of time, grouping, teacher roles and responsibilities, and teacher preparation these classes scored lower than both DARCEE classroom treatments.

Implications

The kind of instrument described in this paper certainly has some limitations. An observer who is busy scoring items on "The DARCEE Classroom Assessment Scale" is less likely to discover subtle differences that exist

FIGURE 5:

AVERAGE RATINGS OF TWO VARIATIONS OF DARCEE TRAINING AND A COMPARISON GROUP OF NON-DARCEE CLASSROOMS



PROGRAM ESSENTIALS

from teacher to teacher in how they structure activities, ask questions, and react to student behaviors than would be a non-participant observer who spends much time in classrooms focusing on such phenomena. Moreover, some of the items were less reliable than we had hoped.

Despite such problems, instruments such as this one may be used by people with a relatively short period of training to answer a variety of questions, such as:

1. Was the program being evaluated actually used?
2. Which of the program's components have shown themselves to be most difficult or easy to implement?
3. The answer to Question 2 may be used to evaluate the success of prior training efforts and modify future training plans.
4. Analysis of subscores in relation to child outcomes could help test the developer's hypothesized relationships between program elements and program outcomes.

A final and by no means minor value of developing such instruments is that developers and evaluators in the process will portray programs in concrete terms.[5] If program portrayal is a major function of evaluation as Stake (1972) has suggested, certainly an effort like this one to specify items that are designed to determine the degree of implementation is a constructive move in that direction.

---

[5]Of course, there is a danger that the effort to develop such an instrument could impose a rigidity in the thinking of program developers and trainers that could have undesirable consequences if the development of the instrument occurs before the developers have decided what alternative teaching behaviors they would regard as acceptable to their program.

References

Brown, C., Dokecki, P. R., O'Connor, M., & Stinson, J. The DARCEE approach to preschool education: Current status and implications for model installation. Demonstration and Research Center for Early Education, Project No. 3A0P02, Planning Document. Nashville, Tenn.: George Peabody College for Teachers, 1971.

Flanders, N. A. Teacher influence, pupil attitudes, and achievement. In R. T. Hyman (Ed.), Teaching: Vantage Points for Inquiry. Philadelphia: Lippincott, 1968.

Gallagher, J. J., & Aschner, M. J. A preliminary report on analysis of classroom interaction. In R. T. Hyman (Ed.), Teaching: Vantage Points for Inquiry. Philadelphia: Lippincott, 1968.

Gross, N., Giaquinta, J. B., & Bernstein, M. Implementing Organizational Innovations. New York: Basic Books, 1971.

Johnson, T. J. Program development and evaluation emphasizing full program characterization. Paper presented at the American Educational Research Association Annual Meeting, New Orleans, February 1973.

Oliver, D. W., & Shaver, J. P. Teaching Public Issues in the High School. Boston: Houghton-Mifflin, 1966.

Stake, R. E. Toward a technology for the evaluation of educational programs. In R. W. Tyler, R. M. Gagne, and M. Scriven (Eds.), Perspectives of Curriculum Evaluation. American Educational Research Association Monograph Series on Curriculum Evaluation, No. 1. Chicago: Rand McNally, 1967.

Stake, R. E. An approach to the evaluation of instructional programs (program portrayal vs. analysis). Paper presented at the American Educational Research Association Annual Meeting, Chicago, April 1972.