Supervision by Objectives: Pupil Achievement as a Measure of Teacher Performance[1]

*HAROLD H. SMITHMAN**
WILLIAM H. LUCIO

DURING the past five decades, numerous researchers have sought to identify indices of teacher performance, but the search for valid criteria has resulted in findings which, in general, have proved to be inconclusive. Morsh and Wilder (1954), in their extensive review of studies of teacher effectiveness, indicated that they found little evidence of particular teaching acts which were associated consistently with learner achievement. Barr (1961), in his summary of the Wisconsin Studies, concluded that the "good" teacher cannot be distinguished from the "poor" teacher on the basis of specific teacher behavior. Other investigators have come to similar conclusions (Howsam, 1960; Ryans, 1960; Openshaw and Cypert, 1966; Kleinman, 1966; Popham, 1971). In general, most of the studies which have attempted to discover criteria of teacher performance have focused on analyzing the instructional *means*—what the teacher does—rather than on the *outcomes* which occur in pupil behavior as a result of instruction.

Since circa 1960 there has been a trend toward studying teacher performance as a correlate of defined and predicted changes in pupil behavior—the outcomes of teaching acts—rather than studying teaching acts; or particular teacher characteristics assumed to relate to teacher effectiveness. Evidence that measures of pupil achievement can serve as practical and effective indices of teacher performance has been presented by a number of investigators (Morsh, Burgess, and Smith, 1958; Moffett, 1966; McNeil, 1966, 1967; Popham, 1968, 1971; Justiz, 1969).

The present investigation was modeled, in part, on the study conducted by Moffett (1966) in which he compared the performance of student teachers who were evaluated on the basis of attaining agreed-upon instructional objectives with the performance of teachers who were evaluated by means of a rating instrument. Data were obtained on the extent of pupils' attainment of instructional objectives (in geography skills) and their attitudes toward subject matter, and on teachers' attitudes toward supervisory help, satisfaction with midterm grades, and preferences for types of performance rating. Secondary school pupils in grades 7 through

[1] This article is based on an unpublished doctoral dissertation submitted to the University of California, Los Angeles, in 1970 by one of the authors, Harold H. Smithman.

* *Harold H. Smithman, Education Officer, Curriculum Development Services, Lakeshore Regional School Board, Beaconsfield, Quebec, Canada; and William H. Lucio, Professor of Education, University of California, Los Angeles*

12, selected by a randomized sampling technique, served as subjects. After pretesting pupils on geography skills, 36 student teachers were randomly assigned to either an experimental or a control group. Teachers in the experimental group executed preinstructional contracts based on the instructional objectives to be achieved as revealed by the pretests, and their teaching performance was evaluated in terms of pupil achievement. Control group teachers, while informed about pupils' pretest deficiencies and the need to correct them, did not enter into an instructional contract; their performance was evaluated by means of rating scale assessments.

Among the findings reported by Moffett were the following: (a) the pupils in the experimental group performed significantly better on a post-test of geography skills than did pupils of teachers in the control group; (b) teachers in the experimental group expressed more confidence in supervisory help and reported satisfaction with their midterm grades significantly more frequently than did teachers in the control group; and (c) 94 percent of all teachers, regardless of whether they were in the experimental or control group, expressed a preference for having their teaching performance based on pupil achievement as a result of instruction rather than on rating scale measures.

Purpose

The major purpose of this study, employing the instructional strategy of supervision by objectives, was to determine the extent to which a supervisor and teacher agreeing on instructional objectives stated in behavioral terms would increase pupil performance. A second purpose was to discover whether the process of supervision by objectives resulted in evaluations of teaching more germane to instructional performance. In conceptualizing the study two propositions were accepted: (a) learning is evidenced by a change in behavior, and (b) teaching is successful only when the instructor's predetermined and intentional changes sought in the learner actually occur (Lucio and McNeil, 1969). Certified teachers and elementary pupils in a Canadian school district served as subjects; mathematics was the curriculum vehicle.

Hypotheses

The following hypotheses, matching those used by Moffett, were tested:

1. Will the pupils of teachers who are evaluated using a system of teaching by objectives: (a) achieve higher scores on post-tests, and (b) show a more positive attitude (affect) toward mathematics than pupils whose teachers are evaluated by standard rating procedures?

2. Will the teachers who are evaluated using a system of teaching by objectives show a more positive attitude (affect) toward supervision and evaluation than those who are evaluated by standard rating procedures?

3. Will the teachers who are evaluated using a system of teaching by objectives indicate a greater preference for this process over evaluation by standard rating procedures?

Procedures

Selection of Subjects. Twenty classroom units, in nine schools, consisting of 20 teachers and 558 pupils in their sixth year of school were selected as the sample using a stratified randomization technique, as described by Van Dalen (1962). A table of random numbers was employed in choosing the experimental and control groups. Ten teachers and 272 pupils were assigned to the control group, and 10 teachers and 286 pupils to the experimental group.

Program and Evaluation Measures. Mathematics was chosen as the vehicle for teaching for the following reasons: (a) the Canadian controlling agency stresses the importance of teaching concepts and the application of knowledge in mathematics; (b) curriculum guides, texts, worksheets, and audio-visual equipment for mathematics instruction were provided for all teachers; and (c) the services of mathematics consultants were readily available. In addition, similar research investigations have focused

on one curriculum segment and the same procedure was followed in this study. For example, Moffett (1966) used geography skills, Popham (1969) an automotive and electronics curriculum, and McNeil (1967) language skills.

Prior to initiating the investigation, and in close collaboration with school district mathematics consultants, the skills in geometry, graphing, and long division that would be taught during the one-month period of the experiment were selected. Objectives were then defined, and a criterion-referenced test was constructed. An analysis of the selected individual items revealed that approximately 50 percent of the test was judged to be low level and 50 percent high level cognition (a category above recall of knowledge being considered high level). Construct validity as described by Davis (1965) was used to validate the test.

In order to maintain a real environment and to prevent biasing the results, the usual school district rating procedures for evaluating teachers and programs were continued. All pretests and post-tests were administered by the classroom teachers under the guidance of the principals and the mathematics consultants, and arrangements were made for the correction of the tests and for informing the teachers of the scores. The principals continued to carry out their usual assigned responsibility for evaluating teachers.

After pretesting the pupils, using the criterion-referenced test, the identified weaknesses in geometry, graphing, and long division skills, as exemplified in pupils' performances, were communicated to the teachers in both experimental and control groups. The instructional objectives, based on mathematical content to be taught in the experimental classes, were prepared by the mathematics consultants and one of the investigators. This procedure was followed because it was determined that teachers should not be expected to devote the time necessary to acquire skills in the preparation of specific and detailed instructional objectives.

For purposes of the study, the treatment for the experimental classes was as follows:

1. In a pre-instructional conference, the teacher and the principal agreed upon those objectives, selected from the prepared list, to be attained by the pupils in each of the teacher's mathematics classes; and the criteria acceptable as proof of the pupils' having reached the stated objectives were agreed upon mutually.

2. The supervisor (principal) visited the classroom for two observational periods to gather evidence regarding the extent to which the objectives had been reached. Observation data on pupil performance were recorded as facts observed without the injection of value judgments or inferences not relevant to instructional results.

3. Based on the observational record, a post-instructional conference was held between the teacher and the supervisor to determine whether or not the objectives had been reached. At this point, or at any point in the procedure, the parties could have renegotiated the contract in order to modify the teacher's original analysis of predicted pupil change. If the objectives were found to be too ambitious or too simple, or if a different procedure appeared advisable, changes could be made.

In supervising the control group, the principal pursued the usual practice in the school district—two observational periods followed by a conference to discuss the teacher's performance based upon the district evaluation guide which contained indices on: (a) teaching techniques, (b) administrative ability, and (c) personal qualities. In some instances the principal presented a written report to the teacher he observed, but in the majority of cases he expressed his judgment verbally.

The individual training sessions conducted for the principals and teachers in the experimental group included the methodology of teaching by objectives, and information on the formulation and the writing of instructional objectives. (Procedures for the conduct of the study were explained to the teachers in the control group in a separate interview.) Topics included in the training sessions were as follows:

1. The Nature of an Instructional Objective (as described by Okumu, Rupert, and Tyler, 1966)

2. A Model for Supervision by Objectives (as developed by Smithman, 1968)

3. Principles To Guide One in Making Classroom Decisions (as described by Lucio and McNeil, 1969).

By the end of the training period the principals were enabled: (a) to write instructional objectives in terms of learner behavioral change for each of the areas in mathematics selected for the investigation, (b) to construct an observational model for supervision by objectives, and (c) to list principles as a guide to making classroom decisions.

To minimize the Hawthorne Effect, it was requested that equal attention and assistance to the teachers in the experimental and control groups be given. There was no evidence to indicate that the consultants did not follow instructions. When the one month's experiment had been completed, a post-test, identical to the pretest, was administered to all pupils in both the experimental and control groups to measure changes if any in mathematical achievement. A total of 544 pupils completed the post-test, 286 from the experimental classes and 258 from the control classes. The decrease in the total completing the post-test was due to absence of 14 pupils from the control group when the test was given.

A pupil questionnaire was administered to the pupils in order: (a) to assess the pupils' attitudes toward mathematics, (b) to determine the pupils' perception of how much they had learned in mathematics, and (c) to assess how much individual instruction the pupils were given. Each teacher completed a questionnaire designed: (a) to assess the teacher's perception of the fairness of his evaluation, (b) to assess the teacher's perception of the objectivity of his evaluation, and (c) to assess the teacher's perception of his principal's help.

In addition the teacher was asked to indicate his preference for one or the other of the two methods of teacher evaluation. The questionnaires administered to pupils and teachers were essentially the same as those used by Moffett in his investigation. No names of teachers were entered on any of the tests or questionnaires, colors being used to distinguish the experimental from the control group.

Statistical Treatment. Although no direction was stated in the hypotheses, it was implied in the theoretical constructs and in the hypotheses that the experimental group would outperform the control group. Therefore, one-tailed "t" tests of significance, acceptable at the .05 level, were used to measure any significant differences between the means of the control and experimental groups on mathematics performance.

Data from the teacher questionnaires received two different treatments. Questions on fairness, objectivity, goal success, and principals' help were treated by one-tailed "t" tests of significance acceptable at the .05 level. Responses to the final question on the teacher attitude questionnaire were analyzed to determine the percentage of choice to the various alternatives.

In analyzing the data on the pupil attitude questionnaire, one-tailed "t" tests were employed. Preference for mathematics, the amount learned, and the percentage of individual instruction were analyzed to determine the statistical significance necessary for acceptance at the .05 level. In reference to the extent of individual instruction, a percentage was arrived at by dividing the total number of times the items were checked by the maximum number of times the items could have been checked.

Results

Pupil Performance. Mathematics pretests were administered to 558 pupils, from 20 classroom units. Although the mean score for the 10 experimental classes was slightly higher than the mean score for the 10 control classes, the difference was not statistically significant (Table 1).

At the conclusion of the treatment, pupils from the 20 sample classes were administered the post-test. A mean difference of 10.31 (.05 level of significance) was

	Experimental (N = 10)		Control (N = 10)			
	Mean	S.D.	Mean	S.D.	t	p
Pretest	22.18	7.38	18.55	8.02	1.055	n.s.
Post-test	48.50	10.41	38.19*	15.05	1.7593	.05
M-Gain Diff.	26.32	4.76	19.64*	11.83	1.5905	.10

* Teacher N = 9 when calculating the post-test mean for control group. Experimental pre- and post-test pupil N = 286. Control pretest pupil N = 272; Control post-test pupil N = 258.

Table 1. Pre- and Post-Test Mean Differences Between the Experimental and Control Classes on the Mathematics Tests

		Experimental (N = 10)		Control (N = 10)			
	Scale	Mean	S.D.	Mean	S.D.	t	p
Fairness	(3,2,1)	2.8	0.63	2.6	0.73	0.645	n.s.
Objectivity	(3,2,1)	2.5	0.71	2.4	0.72	0.312	n.s.
Goal success	(3,2,1)	2.6	0.51	2.7	0.50	0.434	n.s.
Principal's help	(4,3,2,1)	2.5	1.08	2.1	0.92	0.869	n.s.

Table 2. Mean Differences Between Teacher Responses on Four Measures of Teacher Attitude Toward Evaluation

		Experimental* (N = 10)		Control** (N = 10)			
	Scale	Mean	S.D.	Mean	S.D.	t	p
Preference	(3,2,1)	1.84	0.33	2.04	0.42	1.162	n.s.
Amount of Learning	(4,3,2,1)	3.69	0.13	3.74	0.14	0.8333	n.s.

* Pupil N = 286.
** Pupil N = 258.

Table 3. Mean Differences Between Pupils in the Experimental and Control Classes on Two Measures of Pupil Attitude Toward Mathematics

found in favor of the experimental classes. Since the variance of the sample groups, for both the pretests and the post-tests, was considered to be statistically homogeneous, the pooled variance "t" model was utilized in testing for significance. In addition, a mean gain difference of 6.68 between the two groups was tested for significance by the separate variance "t" model (Table 1) and found to be significant at the .10 level (one-tailed). This finding was judged as further evidence that the two sample groups were statistically equated initially.

As a result of monitoring the procedures followed by all the teachers in the study, it was found that one teacher in the control group made a copy of the criterion test and used it to plan the month's instructional activities. Since such a procedure might have influenced the results, the data on this teacher's group were excluded from the final analysis.

Teacher Attitudes Toward Evaluation. The teacher questionnaire attempted to measure the following five teacher attitudes: (a) fairness of the supervisor's evaluation, (b) objectivity of the supervisor while evaluating, (c) goal success as perceived by the teacher, (d) amount of assistance the supervisor gave the teacher, and (e) the evaluative technique most preferred by the teacher. In analyzing the responses it was found that no significant statistical differences were observed between the experimental and control group teachers on the first four characteristics (Table 2).

On the fifth measure of teacher attitude ("the evaluative technique most preferred by the teacher"), 90 percent of the teachers in the experimental group and 88.8 percent of the teachers in the control group favored an evaluation based on pupil performance.

Student Attitude Toward Mathematics. The data obtained from the pupil attitude questionnaire revealed that there were no significant differences between the experimental and control classes relative to the pupils' preference for mathematics and to the pupils' perception of the amount of mathematics learned (Table 3).

In addition it was found that the percentage of individual instruction provided for each group of pupils was similar, since 62.7 percent of the pupils in the experimental classes and 63.1 percent of the pupils in the control classes reported that they had received individual instruction from their teachers.

Tests of Hypotheses

Pupil Achievement. The results supported Hypothesis 1 (a): "Will the pupils of teachers who are evaluated using a system of teaching by objectives achieve higher scores on post-tests?" A difference of 10.3 in the mean scores favoring the experimental group was found to be statistically significant at the .05 level of confidence. The data supported Moffett's findings in which a signifi-

cant difference (.01 level of confidence) in pupil achievement favoring teachers supervised by objectives was reported.

Teacher Attitude. No statistical difference was found between the experimental and the control group in relation to the teachers' perception of the amount of help received from the supervising principal. These findings were not compatible with the data reported by Moffett, who discovered that student teachers in the experimental group were of the opinion that they had realized more help from the supervisor (significant at the .05 level of confidence) than did the student teachers in the control group.

No evidence was obtained to support Hypothesis 2: "Will the teachers who are evaluated using a system of teaching by objectives show a more positive attitude (affect) toward the supervisor and evaluation than those who are evaluated by standard rating procedures?" The data in this study did not support the implication that teachers supervised by objectives were more amenable to evaluation than teachers who were evaluated on a rating scale.

Hypothesis 3: "Will the teachers who are evaluated using a system of teaching by objectives indicate a greater preference for this process over evaluation by standard rating procedures?" was confirmed—a significantly large percentage of teachers in each group reporting that they preferred to be evaluated in terms of their pupils' performance. Apparently teachers perceived the evaluation of teacher effectiveness on the basis of the accomplishments of their pupils as a fair and reasonable procedure. Again, this finding was in concurrence with Moffett's reporting that 94 percent of the teachers preferred to be evaluated on the basis of pupil output.

The results of the study did not support Hypothesis 1 (b): "Will the pupils of teachers who are evaluated using a system of teaching by objectives show a more positive attitude (affect) toward mathematics than those pupils whose teachers are evaluated by standard rating procedures?" No statistically significant difference was found between the experimental group and the control group with regard to pupil attitude toward learning. Pupils in the experimental classes did not indicate any more preference for the subject matter than pupils in the control classes. Similarly, there was no difference in the pupils' perceptions of how much they had learned.

One major criticism of teaching by objectives has been that pupils might receive the impression that they were being "programmed," and thus acquire a dislike for the subject matter. The data in this study, nevertheless, supported Moffett's contention that the lack of a significant difference between the groups discounted the possibility of undesirable consequences.

There was no statistically significant difference in the amount of individual instruction given to each group of pupils by the teachers in either the experimental group or the control group. Once more the data supported the evidence gathered by Moffett. It may be concluded, therefore, that teachers in the experimental group (teaching by objectives) were attentive not only to subject matter and instructional objectives but also as aware of each pupil as the teachers in the control group in providing for individual instruction.

Summary

In an attempt to test a technique for teacher evaluation labeled supervision by objectives, 20 classroom units, in nine schools, consisting of 20 teachers and 558 pupils in their sixth year of school were selected as the sample using a stratified randomization technique. Ten teachers were evaluated using the procedure of supervision by objectives, and 10 by means of a school district rating scale. Pretests and post-tests in mathematics were administered to the pupils in the two sample classes, and the results from the two groups were compared by the pooled variance "t" model. The treatment was administered over a one-month period.

Results on the post-tests indicated that the performance of the experimental classes

was superior to that of the control classes in mathematics, and it was concluded that pupils whose teachers were evaluated by objectives outperformed those pupils whose teachers were evaluated on a rating scale.

In addition to measuring the performance level of pupils, the teachers and pupils participating in the research completed a set of attitude questionnaires. On the five measures of teacher attitude measured by the questionnaire, no significant differences were found between the experimental and the control group. It might be noted, however, that a majority of the 20 teachers expressed a preference for evaluation in terms of the performance of their pupils. The responses to the pupil questionnaire revealed that no differences existed between the experimental and the control group in reference to a partiality for the subject matter and the amount learned. Since the data also indicated that the amount of individual instruction provided for each group was approximately equal, it was concluded that there were no undesirable side effects for the pupils of the teachers teaching by objectives.

References

A. S. Barr *et al. Wisconsin Studies of the Measurement and Prediction of Teacher Effectiveness.* Madison, Wisconsin: Dembar Publications, Inc., 1961.

Frederick B. Davis. *Educational Measurements and Their Interpretation.* Belmont, California: Wadsworth Publishing Company, 1965.

Robert B. Howsam. "Who's a Good Teacher?" A special project prepared for and published by the Joint Committee on Personnel Procedures, California School Boards Association and the California Teachers Association, Burlingame, California, 1960.

Thomas B. Justiz. "A Method for Identifying the Effective Teacher." Unpublished Ed.D. dissertation, University of California, Los Angeles, 1968. (Reported in a paper presented at the American Educational Research Association annual meeting, Los Angeles, California, 1969.)

Gladys S. Kleinman. "Assessing Teacher Effectiveness: The State of the Art." *Science Education* 50 (3): 234-38; April 1966.

William H. Lucio and John D. McNeil. *Supervision: A Synthesis of Thought and Action.* Second Edition. New York: McGraw-Hill Book Company, 1969.

John D. McNeil. "Antidote to a School Scandal." *Educational Forum* 31 (1): 69-77; November 1966.

John D. McNeil. "Concomitants of Using Behavioral Objectives in the Assessment of Teacher Effectiveness." *Journal of Experimental Education* 36 (1): 67-74; Fall 1967.

George McHatton Moffett. "Use of Instructional Objectives in the Supervision of Student Teachers." Unpublished Ed.D. dissertation, University of California, Los Angeles, 1966.

Joseph Morsh and Eleanor Wilder. *Identifying the Effective Instructor: A Review of the Quantitative Studies, 1900-1952.* Research Bulletin AFPTRC-T-54-44. San Antonio, Texas: Lackland Air Force Base, 1954.

J. E. Morsh, G. G. Burgess, and P. N. Smith. *Student Achievement as a Measure of Instructor Effectiveness.* Research Bulletin AFPTC-TN-55-12. San Antonio, Texas: Lackland Air Force Base, 1958. p. 8.

L. Okumu, K. Rupert, and L. Tyler. *Using a Curriculum Rationale.* Los Angeles: The Students' Store, University of California, Los Angeles, 1966.

M. Karl Openshaw and Frederick R. Cypert. *The Development of a Taxonomy for the Classification of Teacher Classroom Behavior.* Cooperative Research Project No. 2288. Washington, D.C.: U.S. Office of Education, 1966.

W. James Popham. *Final Report: Performance Tests of Instructor Competence for Trade and Technical Education.* Office of Education Project No. 5-004, Contract No. OE-5-85-051, University of California, Los Angeles, June 1968.

W. James Popham. "Performance Tests of Teaching Proficiency: Rationale, Development, and Validation." *American Educational Research Journal* 8 (1): 105-17; January 1971.

W. James Popham. "Validation Results: Performance Tests of Teaching Proficiency in Vocational Education." A paper presented at the American Educational Research Association annual meeting, Los Angeles, California, 1969.

D. G. Ryans. *Characteristics of Teachers: Their Description, Comparison, and Appraisal.* Washington, D.C.: American Council on Education, 1960.

Harold Henry Smithman. "Student Achievement as a Measure of Teacher Performance." Unpublished Ed.D. dissertation, University of California, Los Angeles, 1970.

Harold Henry Smithman. "Supervision by Objectives." An instructional program for administrators, 1968. (Mimeographed.)

Deobold B. Van Dalen. *Understanding Educational Research.* New York: McGraw-Hill Book Company, 1962. □