DOCUMENT RESUME

ED 074 057 SP 006 284

| | |
|---|---|
| AUTHOR | Okey, James R.; Ciesla, Jerome L. |
| TITLE | Designs for the Evaluation of Teacher Training Materials. Report No. 2. |
| INSTITUTION | National Center for the Development of Training Materials in Teacher Education, Bloomington, Ind. |
| SPONS AGENCY | National Center for the Improvement of Educational Systems (DHEW/OE), Washington, D. C. |
| PUB DATE | Oct 72 |
| NOTE | 19p. |
| EDRS PRICE | MF-$0.65 HC-$3.29 |
| DESCRIPTORS | Curriculum Evaluation; *Evaluation Methods; *Instructional Materials; Measurement Instruments; *Research Design; Teacher Education; *Teacher Evaluation; *Teaching Skills |

ABSTRACT

This paper describes methods to assess the impact on students of a teacher using skills learned in a training program. Three designs for assessing the effects of teacher training materials are presented: time series design, equivalent time-samples design, and posttest-only control group design. Data obtained by classroom teachers while using the designs are included. Some of the considerations when selecting appropriate research and evaluation designs are discussed in addition to the problems of analyzing data from the designs. An eight-item bibliography is included. (Author/MJM)

# Acquiring Teaching Competencies:

## *Reports and Studies*

*National Center for the Development of Training Materials in Teacher Education*

SCHOOL OF EDUCATION

INDIANA UNIVERSTIY

BLOOMINGTON

This series is published and distributed under the auspices of the National Center for the Development of Training Materials in Teacher Education. The National Center has been initiated and supported by a grant from the National Center for the Improvement of Educational Systems, U.S. Office of Education.

The primary objective of this publication series is to provide an outlet for theoretical, procedural, technical and evaluational reports and studies in the development of protocol and training materials, and in their use in the acquisition of teaching competencies.

The editorial advisory board functions primarily to set policy regarding directions and purposes of the publication and areas of needed publication. Editors for each report will be selected from those listed below.



Manuscripts for consideration should be submitted to:

Laurence D. Brown
National Center for the Development of Training Materials in Teacher Education
School of Education
Indiana University
Bloomington, Indiana 47401

# Designs for the Evaluation of Teacher Training Materials

James R. Okey and Jerome L. Ciesla

Report #2, October, 1972

# *Foreword*

A mandated concern of all those individuals engaged in the production of protocol and training materials is that of evaluation. The specific objective of all the protocol and training materials projects is to produce materials which have been tested and revised until it can be demonstrated that the materials are effective. Unfortunately, not all of us are sophisticated in the concepts and strategies of evaluation although we may know, or are coming to know, much about the strategies by which materials are developed. The authors of this article have both kinds of skills and are exceptionally well suited therefore to speak to the specific problems of evaluation of the particular kinds of materials which concern the National Center for the Development of Training Materials in Teacher Education

The senior author, Dr. Okey, was one of the original project directors at Indiana University with whom the National Center contracted to produce a set of training materials. His project consisted of a self-instructional program entitled *TEACHING FOR MASTERY* and is based upon Bloom's well-known formulation. The materials focus on the acquisition of skills in the preparation and use of diagnostic examinations which provide information for student remedial work. After having completed a preliminary set of materials, Dr. Okey proceeded to evaluate them not only in terms of the immediate effect on the learner but also in terms of the effects on the students of the learners. The end product is a validated set of materials. In addition, however, there is a by-product in the form of an evaluation procedure which may provide a prototypical model for evaluation of other such projects. The subject of the article of course is the evaluation model. I believe it will prove to be helpful for many developers interested in evaluating their training materials.

L. D. Brown, *Editor*

# Designs for the Evaluation of Teacher Training Materials

James R. Okey and Jerome L. Ciesla

Indiana University

The intention in this paper will be to describe methods to assess the impact on students of a teacher using skills learned in a training program. To accomplish this, a program designed to train teachers in a particular set of classroom skills will be described. Then, designs used to assess the effect of teachers using these skills will be given. Thus, while the paper describes a particular set of training materials and methods for measuring their effectiveness, the intention is to illustrate evaluation designs that have wide application for assessing the effects of using teaching skills in terms of student outcomes.

## Evaluation Questions

There are three crucial questions a developer or evaluator of teacher training materials needs to ask:

1. Do teachers attain skills which the materials are designed to teach?

To answer this question requires measurement of whether a training program is effective in producing stated performance outcomes. This amounts to an internal or intrinsic evaluation (Scriven, 1967) of the training package. For example, if an objective of a training package is to learn to construct divergent questions, a posttest would be given to a teacher following study of the package to assess achievement of this skill. If an objective is to learn to construct evaluation items for given objectives, a test administered to anyone studying the package would indicate whether or not this skill was acquired. In either event, the important question for the developer is whether the training

program produces the outcomes specified for it. Other aspects of internal or intrinsic evaluation could be used by a developer or evaluator, but these shall not be considered here.

2. Do teachers use skills from the training materials in their classrooms?

This evaluation is commonly performed with observation schedules or rating forms (e.g., see Amidon and Hough, 1967). Observers enter the classroom directly or vicariously to record what a teacher does. Amount of teacher talk, frequency of verbal praise, or the type and number of questions asked may be recorded, depending on whatever skills were included in the training program being evaluated.

3. Does the use of skills by teachers have any effect on student learning?

This question concerns not the training package itself, but the "payoff" for using the skills in it (Scriven, 1967). For example, if teachers learn to construct diagnostic tests by studying training materials, a payoff evaluation might determine whether use of this skill increased student achievement. If teachers learned to use praise to reward classroom participation, the effect of use of praise by teachers on student attitude could be measured. The emphasis in each case is not on acquisition of a skill, but on the effects of using it.

Each of the three questions posed above is important. A thorough evaluation of a training program will attend to each one. The intention in this paper, however, is to focus on designs to aid in answering the third question, whether the use of certain skills by teachers has any payoff in altered student achievement. The reason for focusing on the

latter question is that little attention has been given to the relationships between teaching skills and student achievement (cf. Rosenshine and Furst, 1971) and to the means of obtaining evidence of these relationships.

Selecting Evaluation Designs

Campbell and Stanley (1963) describe an extensive set of designs for research and evaluation studies. For each design included in their work, they discuss threats to validity, procedures for organizing groups, methods of scheduling treatments and measurements, and suggestions for analyzing data. Among the sixteen designs they describe, three are identified as true experimental designs (Pretest-Posttest Control Group Design, Solomon Four-Group Design, Posttest-Only Control Group Design) and are recommended for use when possible.

Despite their acknowledged superiority for gathering data to answer questions, the three recommended designs of Campbell and Stanley are frequently difficult to use because each of the designs specifies one or more control groups. Use of control groups, however beneficial for obtaining reliable answers to questions, is often not practical because:

a. few subjects (teachers) may be available and dividing a small population reduces the number of subjects for measuring treatment effects.

b. subjects (teachers) resent placebo treatments or serving as members of untreated control groups.

c. ethical questions arise regarding the use of control groups or placebo treatments.

Payoff evaluation studies, by definition, must be done with teach-

ers who have students. To find teachers with students, a developer or evaluator may go directly to schools to locate volunteers or work through in-service classes. These employed teachers may be enticed into trying new materials or techniques when they see an advantage to themselves in doing so. However, it is difficult to convince teachers with a heavy work load and numerous problems for which they desire help that they should participate in a study as a member of a control group.

When it is impossible to use the recommended designs the next best procedure can be tried--in this case, using what Campbell and Stanley call "quasi-experimental" designs. The difference between these and true experimental designs lies in the degree to which the experimenter has control over arranging treatments, selecting subjects, scheduling observations, and other events which occur during an experiment. Several of Campbell and Stanley's quasi-experimental designs are one-group designs in which the same teachers act as both experimental and control teachers; yet the designs allow a comparison of the effects of using and not using selected teaching skills.

In the remainder of this paper three designs taken from Campbell and Stanley (1963) will be used to demonstrate how data can be gathered for payoff evaluation studies while avoiding the problem of setting up separate groups of teachers for comparison purposes. The three designs are singled out to illustrate alternative procedures for evaluating the effects of training in a classroom setting. The training package used in the studies will be described briefly and will be followed by a description of each design and the sample data collected when using it.

## The Training Materials

A self-instructional program called Teaching for Mastery (Okey and

Ciesla, 1972) designed to train teachers to implement Bloom's mastery learning strategy (1968) was developed. The materials, which require about five hours to complete, consist of tape-slide and paper and pencil exercises. Frequent opportunities for practice and feedback are included and self-tests with answers are available for each of the six sections into which the program is divided. A total of 22 outcomes are stated in the program that range from sequencing objectives, to constructing diagnostic tests, to selecting alternative instruction for unsuccessful students.

The overall goal of the training program is to teach teachers to implement a five step plan for increasing the achievement of their students. The major skills required to do this are learning to prepare and administer diagnostic examinations on course objectives at frequent intervals, and then to direct students to remedial work as needed.

The Teaching for Mastery program was studied by all members of an in-service class of 21 elementary school teachers about mid-way through a 15 week term. Portions of two class periods were devoted to independent study of the program with the remainder done outside of class.

Time Series Design

Campbell and Stanley (1963) diagram the Time Series Design as follows:

$$O_1 \quad O_2 \quad O_3 \quad O_4 \quad X \quad O_5 \quad O_6 \quad O_7 \quad O_8$$

The diagram shows a time sequence of events from $O_1$ on the left to $O_8$ on the right. Measurements or observations ($O_1$, $O_2$, etc.) are made at intervals and then a treatment (X) is introduced. Following the treatment, measurements ($O_5$, $O_6$, etc.) are continued. This design has been

used tc measure such things as attitude changes both preceeding and following an event such as showing a motion picture on race relations. Another use might be to examine the number of students that leave school before and after setting up a dropout-prevention program.

The Time Series Design is well suited to evaluating the effects of teachers studying and using skills from a training package when a two group design is impossible. Multiple measurements before studying the package allow pre-treatment or baseline behavior to be established. Repeated measures after studying the package allow both immediate and long term effects to be measured. Using several observations before and after a treatment allows an evaluator to interpret results more confidently because transient or spurious effects are more apparent.

Figure 1 shows data gathered by a first grade teacher using a Time Series Design with a class of 24 students. The plotted points represent the percentage of children in the class scoring 90% and higher on summative tests in mathematics given at approximately two week intervals. The first three observations were made prior to studying the Teaching for Mastery materials and the last three after doing so. Thus, the graph shows achievement results for about 12 weeks of instruction.

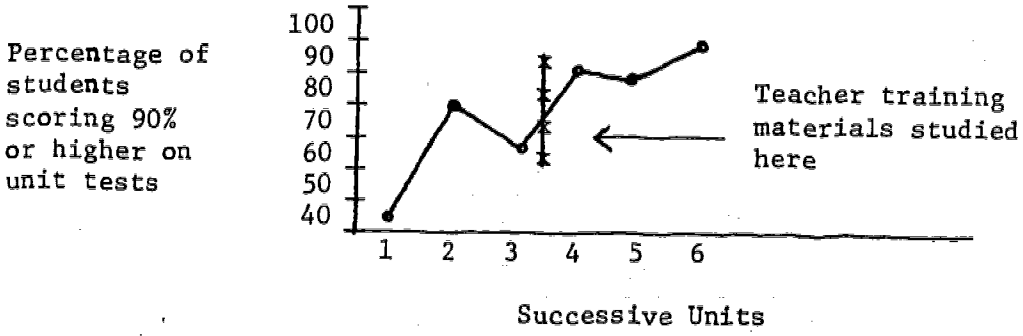


Figure 1. Student performance before and after teacher training materials are studied.

The reason for studying the Teaching for Mastery materials was to have teachers learn to use the skills taught in the package and thereby increase student achievement. One measure of this achievement is the number of students scoring at a selected level on tests over the objectives for a unit. In this study the teacher used a 90% criterion level; if students scored 90 or above on a unit test they were said to have mastered the material. Other criteria could, of course, be used such as an 80% criterion level or the mean test score for all students.

The problem of analyzing data from a Time Series Design is considerable. If the several observations before and after the treatment are the same (e.g., repeated administration of the same attitude measure), problems of comparison are simplified. In this case, however, the observations are different; six unit tests are given, each covering different objectives. To compare the scores is hazardous because objectives from one unit may be more difficult than those from another.

In this study the procedure for analyzing data from the Time Series Design was to compare the mean percentage of students achieving the 90% criterion before and after the treatment. These data are given in Table 1. Correlated proportions should be used for this comparison since the same class took a series of six tests. This was not done because only a set of scores for the entire class was available for each unit, not individual scores for individual pupils on each unit. Because three observations were made both before and after the treatment, the number of subjects used in calculating the z value is three times the number of students in the class to account for the three observations contained in the mean score. That is, a pre-treatment n of 60 (3 X 20)

and a post-treatment n of 69 (3 X 23) was used.[1] The difference between the proportions is also significant ($z = 1.99$, $p < .05$) when 20 and 23 subjects are used in the calculations.

Table 1

Comparison of Pre- and Post-Treatment Achievement in a Time Series Design

| Measure | n | Percentage of students scoring 90% or higher on successive tests | | | Mean | z |
|---|---|---|---|---|---|---|
| Pre-treatment | 60 | 40 | 75 | 57 | 57.3 | |
| Post-treatment | 69 | 83 | 78 | 87 | 82.7 | 4.6* |

* $p < .001$

More sophisticated data analyses than shown here are possible when using Time Series Designs. In this study the proportion of students scoring above a certain level on unit tests was selected because this was the criterion teachers were encouraged to use in the training program. Campbell and Stanley (1963) treat the problem of comparison of observations from Time Series Designs at greater length.

[1] The number of students in the class fluctuated during the study. The average number before the treatment was 20 and after the treatment was 23.

Equivalent Time-Samples Design

The Equivalent Time-Samples Design is diagrammed by Campbell and Stanley (1963) as follows:

$$X_1O \quad X_0O \quad X_1O \quad X_0O$$

A time sequence of events is shown starting with treatment $X_1$ on the left and proceeding to the final observation on the right. This design can be thought of as an "on and off" design. A treatment is introduced ($X_1$) and then withheld ($X_0$), then reintroduced and then withheld again, and so on. In other words, the treatment or experimental variable is turned on and off. After each use or non-use an observation (O) is made of the behavior being examined.

The Equivalent Time-Samples Design can be readily used for assessing the power of skills learned in a teacher training package. Suppose a teacher learns to use certain questioning skills. These skills can then alternately be used and not used in successive encounters with students. Students' attitudes or intellectual achievements under each treatment can serve as dependent variables to assess the effectiveness of the skills.

If teaching skills have an effect on student learning and are alternately turned on and off in successive units, a saw-tooth type of achievement record should result. When the skills are in effect student achievement should be up, when not used, student achievement should be down. Of course, a reverse situation would be expected if the teaching skills were designed to alter a behavior such as frequency of classroom fights.

Figure 2 shows the results obtained by a sixth grade teacher using the Equivalent Time-Samples Design with 29 students during four successive units in a mathematics class. During the four units, each approximately two weeks in length, the teacher alternately used and did not use the skills studied in the Teaching for Mastery training materials.

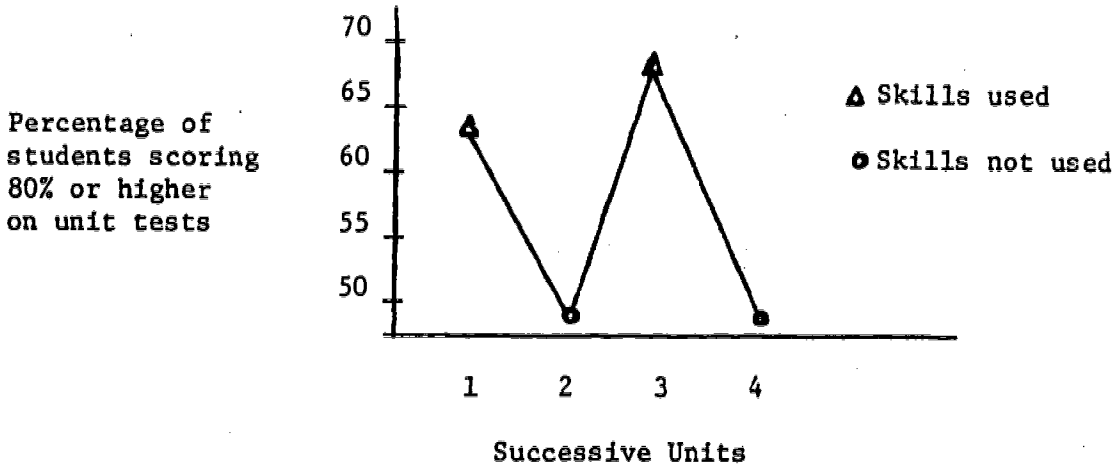


Figure 2. Performance on units during which the teacher turns skills on and off.

Results obtained with the on and off treatment confirmed expectations. When the teacher used skills learned in the training program, student achievement was up; when skills were not used, student achievement fell. Table 2 shows the results of a test for the significance of difference in achievement for the two units when the skills were used and the two for when they were not. A total of 29 students studied each of the four units. Values of $n = 58$ were used in the calculation to reflect the two observations under each treatment condition. When an n of 29 is used, a z value of 1.23 ($p < .09$) is obtained.

Table 2

Comparison of Achievement in an Equivalent Time-Samples Design

| Treatment Condition | n | Percentage of students scoring 80% or higher on successive tests | | Mean Percentage | z |
|---|---|---|---|---|---|
| Skills used | 58 | 62 | 66 | 64 | |
| Skills not used | 58 | 48 | 48 | 48 | 3.4* |

* $p < .001$

Posttest-Only Control Group Design

The Posttest-Only Control Group Design is diagrammed by Campbell and Stanley (1963) as follows:

$$\begin{array}{ccc} R & X & O_1 \\ R & & O_2 \end{array}$$

This design is an excellent one to use when testing the effectiveness of teaching skills except for the difficulty of withholding an experimental treatment from a group of teachers. A way around the problem of withholding treatments from teachers, however, is to have teachers withhold certain treatments from portions of their students for limited periods of time. For example, a teacher studies a set of training materials and learns certain skills that are intended to alter student behavior. To test the effectiveness of these skills, the teacher divides a class, using the randomization procedure, into two groups. For a short period of time, perhaps for two or three weeks, he teaches one of these

groups using the learned skills and teaches the other group without using them. Both groups of students pursue the same objectives and are judged using the same criteria whether a unit test, an observation instrument, or some other evaluation instrument. Appropriate procedures for isolating the groups during study (e.g., sending one group to the library while the other group is taught) and avoiding other sources of contamination (e.g., alternating the order in which the two groups are taught on successive days) are necessary.

Table 3 shows data obtained by a third grade teacher using the Post-test-Only Control Group Design with 26 students during a two week unit on fractions. The students were divided at random into two groups and taught by using and not using the skills from the Teaching for Mastery program. Both groups took the same test over the same set of 20 objectives at the end of the unit.

Table 3

Scores for students taught while the teacher used and did not use Mastery Teaching Skills

| Group | n | $\bar{X}$ | SD | t |
|---|---|---|---|---|
| Mastery skills not used | 13 | 10.2 | 3.20 | 2.0* |
| Mastery skills used | 13 | 12.8 | 3.26 | |

* $p < .05$

Discussion

The first point to be made is that the designs illustrated in this paper for measuring the effects of teacher training materials are not new designs. They have been described at length by a variety of people and have been used extensively. They have not, however, been used often for measuring teacher training effects. As Rosenshine and Furst (1971) point out, there have not been many studies (they report approximately 50) in which the relationship between teacher behavior and student achievement is examined. Even among the studies reported, most have been correlational. The studies reported in this paper are experimental and illustrate the use of designs for examining cause and effect relationships between teaching skills and student achievement.

Another point to be made is that these designs are not better than others that might be used. An investigator or developer of training materials should select the design that is possible to use under the circumstances that exist. For example, a time series design, because it is a single group design, is probably less ideal than several of the "two group" designs, but it is not always possible to constitute several treatment groups in a study. If the most that the developer has available to him is a single group, he has to select some design that will collect the maximum data in that situation.

Whenever investigations are carried out it is important to keep in mind the audience to whom one wishes to speak. Different information may be necessary to demonstrate to different groups the effectiveness of a treatment. Classroom teachers and principals (who are probably less sophisticated in statistical analysis) are likely to be more interested

in descriptive data (of the type shown in the graphs in the previous pages) than in an analysis of variance table or the results of a z or t test. Persons who have a background in statistics will be likely to require different results to be convinced of the power of a treatment. One can see that both descriptive and inferential statistics play a role in communicating the results of an investigation to potential users.

Perhaps too often we have decided that inferential statistics are needed in order to assess the effects of treatments. If you look, however, at the results that the teacher obtained in the Equivalent Time Samples study in this paper, you will see a fairly pronounced treatment effect between the times the teacher was using the skills and the times she was not. Although this is fairly dramatic when presented graphically, the results are not significantly different when a .05 level of significance is used with an n of 29 students. Thus, inferential statistics may lead one to the conclusion that there was no significant treatment effect, while a descriptive display of the data leads one to conclude the opposite.

Additional analytic power could have been achieved in these studies by selecting appropriate classification variables and blocking on these for precision of analysis. For example, one could have obtained IQ scores, motivation scores, or creativity scores from students, and then blocked accordingly. Not only would this have given more power to the analysis, but it would have allowed identification of interactions among different sub-groups on classification variables with certain treatments. One might find, for example, that high achievement-oriented students do best under mastery conditions or that certain IQ groups are differentially

affected by use of certain teaching skills. In other words, certain aptitude-treatment interactions could be identified by selecting appropriate classification variables and determining which sub-groups on these variables interact favorably or unfavorably with certain treatment conditions.

Throughout this study the investigators had minimal contact with the teachers when they were in their classrooms. In fact, no visits were made to any classrooms. The only intervention by the investigators was to tell the teachers what data to gather, what time intervals to use, and what design to follow. The teachers collected all data and instituted all treatments. It should also be pointed out that because of this there was no check on the teachers' fidelity regarding use of the skills that they learned in the training package. For future studies, observation instruments or rating scales should be developed in the manner of Worthen (1968) to establish the degree to which the teachers incorporate the strategies or use the skills that they learned in the training materials in their actual classroom work.

Data from only three teachers from the in-service class of 21 are reported in this paper. Quite obviously some of the more successful ones are reported. Statistically significant results were obtained by teachers using each of the three designs although outcomes varied; some teachers were able to cause highly significant changes in student performance and others were not. Most of the teachers in the study, however, (more than 80%) were able to effect some degree of improved performance with their students. What the three studies describe, therefore, is what certain teachers were able to do after receiving a limited amount of instruction from a short piece of training material that was in a preliminary

phase of development. Data obtained from the 21 teachers are being used to revise the Teaching for Mastery training package.

A final comment should be made about the rigor of the studies reported here, the analysis of data obtained from them, and the confidence one can place in the results. Certainly the results obtained in any of the studies fall short of a full scale validation of the training program. Little control was maintained over the teachers and no measures were made of their ability to institute the treatments. Some students of statistics may quarrel with the data analysis for each of the designs. In particular, the number of degrees of freedom to use when calculating the z values is arguable. We have analyzed the data using one set of assumptions and made a case for doing so. An alternative and more conservative analysis is also reported. As Guba (1969) has noted, evaluation studies in a field setting almost invariably fail to meet some of the criteria for traditional research studies. Because this is true for the three studies described, the confidence in the results falls somewhat short of that obtained from a laboratory controlled study, but is a good deal greater than the confidence one has in an untried instructional program.

## Conclusion

Three designs for assessing the effects of studying teacher training materials are given along with data obtained by classroom teachers when they were used. The teachers had studied a self-instructional training package designed to teach them to use Bloom's mastery learning strategy. Some of the considerations when selecting appropriate research and evaluation designs are discussed. Problems of analyzing data from the designs are also considered.

## References


Amidon, E. and Hough, J. (Eds.) Interaction analysis: Theory, research, and application. Reading, Massachusetts: Addison-Wesley, 1967.

Bloom, B. Mastery learning. In J. Block (Ed.), Mastery learning: Theory and practice. New York: Holt, Rinehart, and Winston, Inc., 1971, 47-63.

Campbell, D. and Stanley, J. Experimental and quasi-experimental designs for research. Chicago: Rand McNally and Co., 1963.

Guba, E. Significant differences. Educational Researcher, 1969, 20, 4-5.

Okey, J. and Ciesla, J. Teaching for mastery. Bloomington, Indiana: National Center for the Development of Training Materials in Teacher Education, Indiana University, 1972.

Rosenshine, B. and Furst, N. Research in teacher performance criteria. In B.O. Smith (Ed.) Research in teacher education. Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1971, 37-72.

Scriven, M. The methodology of evaluation. In B.O. Smith (Ed.) Perspectives of curriculum evaluation. Chicago: Rand McNally and Co., 1967, 39-83.

Worthen, B. A study of discovery and expository presentation: Implications for teaching. Journal of Teacher Education, 1968, 19, 223-242.