DOCUMENT RESUME

ED 088 857 SP 007 828

AUTHOR Flanders, Ned A.
TITLE Knowledge About Teacher Effectiveness. Teacher Education Publication Series. Report A73-17.
INSTITUTION Far West Lab. for Educational Research and Development, San Francisco, Calif.
PUB DATE 73
NOTE 47p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, Louisiana, February 1973)

EDRS PRICE MF-$0.75 HC-$1.85
DESCRIPTORS Behavioral Science Research; *Educational Research; Effective Teaching; Research Design; *Research Methodology; *Research Problems; *Research Reviews (Publications); Research Utilization

ABSTRACT

This paper was presented at a symposium concerned with the results of research on teacher effectiveness and the procedures which might be used to review such research. The first part of the paper describes Rosenshine's research review procedure and analyzes the logic used to group research reports. Three possible standards which might be applicable to the writing of reviews on teaching are identified. The second section consists of a discussion of five technical problems which occur in the conduct of research on teaching and which affect research reviews in this area. Two of these problems, the unit of sampling and analysis of replicated studies are highly controversial and may have important consequences for research on teaching. The third section presents and discusses the kinds of knowledge that researchers should seek to gain in the future. (HMD)

ED 088857

TEACHER EDUCATION DIVISION PUBLICATION SERIES

KNOWLEDGE ABOUT TEACHER EFFECTIVENESS

Ned A. Flanders

Paper presented at the meeting of the American Educational Research Association, New Orleans, February 1973

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

REPORT A73-17

FAR WEST LABORATORY FOR EDUCATIONAL RESEARCH AND DEVELOPMENT
1855 Folsom Street, San Francisco, California, 94103, (415) 565-3000

SP 007 828

This book review is enclosed with the Flanders article at the request of Barak Rosenshine.

ROSENSHINE, BARAK, *Teaching Behaviours and Student Achievement.* Slough: National Foundation for Educational Research in England and Wales, 1971. pp. 229. £ 3.25.

This book reviews over one hundred and twenty researches into teacher behaviour and student achievement and does so with a thoroughness which would be difficult to better. Under six heads the effects of teacher classroom behaviour ranging from "teacher approval and disapproval" through "enthusiasm" to "overall competence" are reviewed in careful and considerable detail. An eighth heading deals with "time" as a variable and a final chapter with "antecedent and demographic variables".

A major difficulty which the author faces and copes with well is caused by the wide variations of measures, meanings and styles of analysis used by researchers in similar general areas of teacher behaviour. It does, however, become increasingly clear that the lack of agreement on terms, modes of analysis and measures of student achievement in the studies reviewed leaves the author with a larger interpretive role than he would wish though he shows commendable restraint, and evolves useful means for classifying the studies. In particular his 'high' and 'low' inference dichotomy is a valuable evaluative tool. But of even greater value are the discussions which follow each sub-area of teacher behaviour research.

These discussions are models of brevity, clarity and insight. Each not only picks out the salient findings of groups of studies, but also suggests possible improvements for future researches. It is because these brief discussions are so good that one is even more unhappy that no overall general discussion should appear as a final chapter. One can only speculate at why this was so – an act of self-denial on the author's part, a fear it would be clear that when summarized a large research effort had produced rather meagre results or a behaviourist's unwillingness to raise questions about the nature of the general model which informs most of the studies and has possibly reduced the level of their insight generating capability?

It is to this last issue that the book in the end draws dramatic attention. How ill served is research into teaching by theories of teaching and why? These are the questions which, though the author does not himself raise them, are raised so forcefully as one follows the careful, ordered account of study after study. If we cannot see in what ways theory ill serves the researcher into teaching, it is hard to understand why researchers continue to expend scarce resources on refinements of modes of research which are clearly of small productivity. To be able to render a rational account of the effectiveness of the teaching process may call for something other, or in addition to, the kinds of researches which this book so successfully recounts.

P. H. TAYLOR, University of Birmingham

International Review of Education, XVIII, 1972, #4, (Special Number), p. 582. UNESCO Institute for Education.

KNOWLEDGE ABOUT TEACHER EFFECTIVENESS

The Problem

This Symposium[1] is an attempt by Meredith Gall, Robert Heath, Mark Nielson, Barak Rosenshine, and me to sharpen contrasting points of view regarding the results of research on teacher effectiveness and the procedures which might be used to review such research. Our search is for areas of agreement and disagreement as we make inferences from research studies and discuss different ways to summarize our present state of knowlege. Our method is to use the published reviews of Rosenshine (1970a, 1971a, and with Furst, 1971, 1973) as material to be analyzed in order to sort out various inferences and alternative procedures for summarizing research results.

Rosenshine is probably the world's leading critic of research on teaching effectiveness and teacher education when judged by total pages of material published in 1970, 1971, and 1972. He was commissioned to write a review on teaching effectiveness for the International Association for the Evaluation of Educational Achievement which became the basis of his

[1] See 1973 AERA Annaul Meeting Program, p. 135.

book, Teaching Behaviors and Student Achievement, recently published in England. According to publicity announcements for English consumption,[2] "This book . . . is likely to arouse wide interest as a possible basis for a real science of teaching." A similar announcement in the Phi Delta Kappan[3] includes such statements as, "The book is cautiously and conservatively conceived . . . (will reduce) the ratio of polemic to fact . . . most up-to-date and thorough (regarding) relationships between specific behaviors and student achievement . . . may be used to study current knowledge about items in assessment instruments . . . focuses squarely on the work of teachers with real students in real schools . . . provides the basis of future research . . ." Granted that an author usually doesn't write advertising copy for his book, these aspirations may be questioned. Should a science of teaching start its development from knowledge consisting of many correlations between one predictor variable and one outcome variable? What is the rationale for grouping the predictor variables? Why wasn't student achievement defined and variation in its measurement discussed? Would further research on a promising predictor variable like clarity be a wise investment of scarce research funds? More than one answer to each of these questions is likely to occur in this Symposium.

In fairness to all of us, but especially to Rosenshine, there have been few if any research reviews which synthesize knowledge about teacher effectiveness in ways that are noncontroversial and widely accepted. It is not easy to decide, in a paraphrase of Herbert Spencer, "what knowledge of pedagogy is of the most worth." These difficulties can be illustrated

---

[2]From a flyer printed by Book Publishing Division, National Foundation for Educational Research, Windsor. Copy supplied by Rosenshine in a personal communication.

[3]Phi Delta Kappan, July, 1972, p. 21

by Rosenshine's (1970, p. 445) criticisms of a review on teaching effectiveness written by Campbell and Barnes (1969), namely, "The flaws include: (1) inappropriate statistical analyses, (2) limits in the external validity or generalizability, (3) data omitted from the summary reports, and (4) misinterpretations in reading." The perception of incorrect results is apparently becoming epidemical to both researchers and reviewers alike because Heath and Nielson (1973) now find essentially these same four flaws as criticisms of Rosenshine's reviews. Gall (1973) argues that Rosenshine failed to define student achievement and to discuss the consequences of different ways of measuring it. One consequence of this omission is that grouping according to the same predictor variable might inadvertently combine studies with different operational definitions of student achievement.

This third article of criticism is divided into three sections. In the first, Rosenshine's procedure is described and then the logic which he uses to group research reports is analyzed with special attention to the different assumptions which are made for each type of conclusion. A distinction is made between what might be called an "ideal logic" and the "practical logic;" the latter exists because the assumptions of the former are always less than perfectly satisfied. It is suggested that when there is a discrepancy between an ideal and a practical logic, the reader of a review will need conceptual definitions as well as operational definitions in order to appraise any grouping procedure. Rosenshine's reviews seldom provide adequate conceptual definitions. The first section concludes by identifying three possible standards which might apply to the writing of reviews on teaching.

In the second section, there is a discussion of five technical problems which occur in the conduct of research on teaching and therefore are of interest to those who review this kind of research. Of these, the first and fourth problems on "the unit of sampling" and on "replication" are the most controversial and may have important consequences for this kind of research.

In the third section, the discussion focuses on the kind of knowledge about teaching that the research community should seek. Skeptical questions are discussed about the utility of simple, linear correlations between two variables. Some alternatives are suggested.

The Logic of Grouping Studies

Reviewers, just like researchers, follow a procedure in the conduct of their work. In both research and reviews of research, the procedure chosen can be the basis of either commendation or criticism, especially in terms of the basic logic on which the procedure rests. Is this logic appropriately chosen? Does it fit the purpose at hand and is it consistently used? In this section the logic of Rosenshine's procedure is discussed.

Rosenshine's Procedure

The procedure that Rosenshine or Rosenshine and Furst followed is perfectly clear, which itself is commendable. It can be described as follows.

Setting limits. The first sentence of Rosenshine's book (1971a) reads "The major purpose of this book is to report the results of all

available studies in which teacher behavior was studied in relation to student achievement" (p. 11)[4]. Studies which report educational outcomes other than achievement are not included.

Preferred statistic. Rosenshine prefers to report correlation coefficients between a predictor variable and student achievement which indicate one-to-one associations. While I would quarrel[5] with the reasons Rosenshine gives for preferring correlation, it is clear that Rosenshine's procedure of grouping and counting does require a consistent statistic across different studies in order to facilitate direct comparisons of one group of studies with another. Estimating the chance occurrance of a linear correlation, of course, involves a "two-tailed" test of significance which is appropriate to purely descriptive studies. Notice that the statistic is not appropriate to research designs based on theoretical hypotheses when such hypotheses permit a "one-tailed" test for accepting or rejecting the null hypothesis.

---

[4]It is unfortunate to begin a book with an inaccurate statement. Rosenshine did not include all the studies that relate teacher behavior to student achievement which were known to him and available to him. The sentence should have the following phrase added after the period: " . . . except those which the author purposely excluded." However, the exclusion of a few studies is not the central issue since their inclusion would not alter Rosenshine's conclusions.

[5]Rosenshine's discussion of how analysis of variance was devised for experimental studies (his book, p. 26) begs the question of when is a study an experiment. It also contravenes the assertion that analysis of variance is to be preferred to correlation for analyzing data with the small "Ns" typically found in this kind of research.

Grouping and counting results. In his book, Rosenshine sorts out more than 150 one-to-one associations[6] from 70 separate research studies, each of which included a measure of student achievement. Each association is assigned to a group or cluster according to Rosenshine's interpretation of the predictor variable. Differences in the measure of student achievement are ignored in the clustering, but a table in the first chapter lists the particular test of achievement that is used in each study. Once these associations are clustered and given a label (like clarity), Rosenshine performs a number of logical operations within and between clusters. Within a cluster "hits and misses" are counted by noting how many correlation coefficients are statistically significant and nonsignificant. Between clusters, the signs, median, and range of the coefficients in one cluster are subjectively compared with those in another cluster in order to determine that one predictor variable is more promising than another, promising in the sense of predicting student achievement more effectively.

Generalizations and conclusions. The most frequent conclusion to be found in Rosenshine's reviews is of the type "out of X studies reviewed, Y showed significant results (for this association)" and these statements usually occur following each of the 44 tables of research results. A second type of generalization is of the type "the results for variable Q

---

[6]Rosenshine does not use the phrase one-to-one associations which, in this article, refers to the correlation of one predictor variable with some measure of student achievement. He uses the word study to mean (1) a one-to-one association, (2) the study of that association, and (3) the study in which the association is cited. He uses the word variable to refer to a particular cluster of similar associations. Rosenshine's readers may reach the misconception that each association reported is from a study in which that association was the primary target of the research. Rosenshine fails to help his readers distinguish between associations which the researcher intended to investigate and those for which data were merely reported serendipitously.

are stronger than those for variable R" which is the basis of one chapter by Rosenshine and Furst (1971), is cited by them again, (1973), and occurs at least once in his book (p. 74). For generalizations of the second type, the comparison is between one cluster of associations and another. The third type of generalization has to do with recommendations for future research such as the one on page 91 of Rosenshine's book " . . . mere frequency counts of 10 or 20 teacher behaviors are not sufficient, and recommendations have been made for the collection of data on the intensity, content, and context of the specific behaviors."

The assumptions on which this procedure rests are discussed in the section that now follows.

Assumptions and Generalizations

Given the three types of generalizations mentioned above, there are a number of assumptions to keep in mind.

Hits and misses within a cluster. The first type of generalization is of the type "Out of Y studies, X results were significant." This generalization speaks to the proportion of hits and misses or the number of significant and nonsignificant correlations. The assumptions are (1) that the opportunity to test an association does exist in Y different studies; (2) that these opportunities are reasonably comparable in the sense of being equally fair; and (3) that a standardized procedure for identifying a hit, which in this case is a test of significance, is being applied consistently in each case.

Variable Q is more promising than R. The second type of generalization compares one group of associations with another, for example, "Clarity is a more promising variable than teacher criticism." The intent of this

generalizations is to compare the evidence for two different predictor variables. The first requirement for generalizations which compare groups is that the three assumptions just mentioned above apply consistently within each of the two groups that are being compared. A fourth assumption is that all of the various measures of pupil achievement in each one-to-one association of both groups are reasonably equivalent. The fifth assumption is that the tests of significance, their power, the magnitude of the correlations, or the sign of correlations, taken together or in some combination, constitute sufficient evidence for stating the preference for one group of associations compared with another.

When a generalization of this second type predicts that a high inference variable like clarity is more promising for future research compared with a low inference variable like teacher criticism, then the fifth assumption (above) runs into real trouble. In this case we must assume that a high inference variable is just as clearly defined as a low inference variable so that they both lend themselves equally to developing clear operational definitions in research or to identifying skills in teacher education. Such an assumption is ridiculous since by definition (see Rosenshine, p. 19, 1971a) high inference variables are ratings which lack the behavioral specificity of identifying low inference acts and thus they cannot lend themselves equally to future research nor to teacher education. For these reasons it is very difficult to understand why Rosenshine, with his co-author Furst, concluded that the variable clarity had more promise than any other predictor variable in two separate

publications (Rosenshine and Furst, 1971, 1973). The reader is left with the problem, which Rosenshine fails to discuss, of why a high inference variable is more promising than a low inference variable simply because of higher correlations.

Recommendations for future research. The third type of generalization is a recommendation about what researchers ought to be doing rather than what they are doing. The assumptions which lie behind such statements are not easily identified because of the complex judgments which enter into each recommendation. Yet it does seem reasonable to believe that such recommendations do rest on how well the assumptions of the type one and type two generalizations are met and, in large measure, on the general quality of the entire review.

Logic-in-use

Kaplan (1964, Chap. 1) distinguishes between reconstructed logic or "what ought to be," and logic-in-use or "what actually exists." I would assign the asusmptions discussed in the previous section to be "ideal," reconstructed logic and, in this section, turn to Rosenshine's logic-in-use.

Imperfect equivalents. No one, least of all Rosenshine, assumes that each one-to-one association cited from a study is exactly equivalent to another one-to-one association no matter whether one group or several groups are involved. Human behavior is never exactly duplicated and the possibility of exactly replicating an entire study exists only in our imagination. Knowing that behavior is never exactly duplicated and, in spite of this, grouping "similar" one-to-one associations into clusters, Rosenshine's reader must decide how imperfect an equivalent can be

(in the sense of somewhat similar associations) and still support the logical comparisons which Rosenshine chooses to make. Consider his comments about 17 studies involving 60 one-to-one associations on teacher criticism (Rosenshine, pp. 52-62, 1971a).

> A single table describing the results of 17 studies is too gross a summary because a variety of behaviors ranging from giving simple directions to extreme teacher hostility are contained in these variables. The specific categories which one investigator developed overlap those another developed, and so this table cannot be divided easily into smaller tables. However, an attempt is made to describe clusters of behaviors within the larger variable 'criticism and control,' but the reader should be aware that the definitions investigators gave may not be comparable, and these definitions may not be identical to the operational definitions which the observers developed in the course of coding. (p. 52 & 59)

This is but one example of many[7] in which Rosenshine is faced with a variety of operational definitions to which the original researchers gave similar labels. Perhaps Rosenshine recognizes that both he and his reader live in the real world in which the requirements of ideal assumptions are seldom met or cannot be met so that the very least he can do, as a responsible reviewer, is to apprise his readers of this state of affairs. Nevertheless, when we read the above quotation literally, it is clear that as far as Rosenshine is concerned, there is no requirement that a cluster of one-to-one associations be formed with a degree of homogeneity that is implied by the ideal assumptions mentioned earlier. Apparently Rosenshine is grouping the associations for convenient description and discussion. He is not grouping them because they represent equally fair tests for the same predictor variable which can then be compared in quantitative terms. He deals only with imperfect equivalents in each group.

---

[7]For other examples, see p. 64--teacher praise; pp. 70-71--use of student ideas; pp. 74 & 77--combined measures of teacher approval; p. 78--ratios of approval and disapproval; p. 93--business like behavior; p. 99--organization; p. 100--clarity, and so on.

The central issue. At this point in the discussion there are three questions which beg for answers.

(1) Are the five assumptions just mentioned really relevant to the generalizations Rosenshine makes; are they a fair or unfair model on which to base the logic of a review? My answer is that these five assumptions are both fair and relevant because of the quantitative comparisons, the type one and type two generalizations, which appear throughout Rosenshine's reviews. Consider the following from his book.

For clarity Rosenshine states:

> "Eight studies which used high-inference ratings . . . (are listed) (p. 100) . . . Significant results were obtained in all eight studies (p. 103). . . The results on clarity are most consistent and significant, particularly in contrast to the results on other variables. Therefore, variables such as 'clarity' are highly recommended for future study (p. 197)."

For teacher criticism Rosenshine states:

> "Seventeen studies were found which included variables that might be labelled 'teacher criticism of pupils' (p. 52) . . . significant negative correlations between teacher use of criticism and pupil achievement on at least one criterion measure were obtained in half of the 17 studies (p. 59) . . . There is no question that variables such as criticism, teacher directness, or giving directions should be included in future research . . .
>
> However, the existing research on teacher disapproval or teacher criticism appears inadequate because insufficient attention has been given to the context in which these behaviors occur." (p. 61--the text that follows makes suggestions on ways that this research might be improved.)

Any reader will conclude that research on clarity is more likely to be productive than research on teacher criticism and this is exactly what Rosenshine and Furst (1971 and 1973) intend to convey. He also intends to convey the same preference in his book from which the above quotations are taken. Presumably such a conclusion is justified because a higher

proportion of the studies on clarity showed significant results and, in general, the coefficients were larger. Yet the validity of this comparison rests on the consistency with which the predictor variable is operationalized in each group of associations, the consistency of the operations for quantifying pupil achievement in both groups, equal fairness in all of the tests of significance, and the assumption that a higher inference variable has as much promise in future research as a low inference variable.

(2) Since the ideal assumptions of a review procedure cannot be met, how does one decide how much deviation from the ideal assumptions can be tolerated? More specifically, are the groups identified by Rosenshine sufficiently homogeneous to justify discussing them as a group in order to reach the type one and type two generalizations? This question articulates one of the most important issues between Rosenshine and his critics in this Symposium and needs to be explored if we are to identify adequate procedures for conducting reviews of research results. As a result, the topic will be discussed in the next section of this paper. At this point, however, there is a third question which I prefer to discuss.

(3) If there is less homogeneity in Rosenshine's groups than is desirable, why not alter the generalizations so they are less quantitative? It might be argued that the phrase "Out of Y studies, X results were significant" is not really a quantitative comparison, instead it is merely descriptive. Such a phrase could be in a context in which a reviewer discusses the studies of Brown, then Smith, then Jones, and so on until the supply of studies on the topic is exhausted and the reviewer then summarizes with the statement "Y studies have been discussed, the results were significant in X studies" with much less emphasis on counting hits

and misses. My reaction to this alternative is that choosing to write less quantitative generalizations fails to eliminate the central problem of homogeneous groups. Less precise quantitative comparisons are clearly possible. Gage (1973, pp. 5-22) proposes to ignore the magnitude of a correlation and, therefore, whether it is significant or not, and count instead the number of positive, doubtful, and negative results for a particular association. This clearly eliminates any assumptions about equally fair tests of significance within and between groups. Gage also shows, in this same reference, that there are subjective similarities between some of Rosenshine's groups and the earlier research of Ryans (1960) and Bush (1954) and discusses what might be called "the emerging consensus on teacher behavior dimensions." While these developments do contribute to a summarization of current knowledge, they do not eliminate the need for homogeneity within a group when that group is chosen to represent the research on a single predictor variable. Even though one-to-one associations are grouped together merely for the purpose of discussion, it is still the responsibility of the reviewer to show <u>why</u> each association is assigned to its particular group.

Possible Standards for Grouping Variables in a Review

Deciding whether there is too much variation and not enough homogeneity within a group of one-to-one associations may include the following steps: (a) checking the conceptual definition of the entire group with the operational definitons in the one-to-one associations, (b) comparing the teacher and pupil samples, and (c) analyzing situational features of the learning opportunity. These aspects of writing a review will be discussed briefly making use of examples from Rosenshine's writing.

Conceptual and operational definitions. An evaluation of the groups in Rosenshine's review begins by understanding the label assigned to a group of associations. For example, what is the meaning of such labels as clarity or teacher criticism? This meaning is communicated by what I choose to call a conceptual definition. A conceptual definition usually consists of sentences which identify what is to be included and what is to be excluded when we think about a particular label. Like all definitions it is an arbitrary convention which aids communication. Here is an example--by teacher criticism is meant acts of a teacher which communicate disapproval to a student or which correct misconceptions which the student has expressed. Such acts have the intent of changing unacceptable ideas or behaviors so that they are more acceptable. Thus, teacher criticism occurs when a teacher makes a judgment about right/wrong, good/bad, correct/incorrect, etc. and communicates his perceptions to the student with the expectation that the student will modify his behavior or change his ideas. In general such acts create a social context in which the teacher initiates and the student complies.

The foregoing sentences provide an example of a conceptual definition, but in this case for the predictor variable only. A complete conceptual definition of a group label should include both the predictor and the outcome variable. Rosenshine presumably avoids the latter by asserting that all studies reviewed include the same outcome variable student achievement. However, according to Gall (1973), Rosenshine does not provide an adequate conceptual definition of student achievement anywhere in his reviews or his book. For example, Rosenshine does not specify whether student achievement is a measure of the subject matter that a teacher intends to teach or whether it is knowledge that a student of a particular grade level is expected to know.

An operational definition consists of a series of sentences which describe the procedure to be used in quantifying a concept. It is like a recipe which, if carried out, will identify a number to represent the variable being measured. Operational definitions have been called conditional definitions by Ennis (1969, p. 236) because they make explicit a set of conditions within which operations are carried out. For example, if an observer is trained to record reliably teacher statements which criticize student behavior or ideas, the incidence of such statements per unit of time can be a measure of teacher criticism. Obviously there are other operations which might be used such as ratings by the students according to their perceptions of how critical the teacher is, or similar ratings by a trained observer. Each one-to-one association involves an operational definition for the predictor variable and the outcome variable.

Given Rosenshine's preference for organizing his review around groups of associations, the crucial question is how can a reader determine how much variation exists in each group? Ideally a reader could complete the following two steps.

(1) Read the reviewer's conceptual definition of a group label. This should include a definition of the predictor and outcome variables since it is the combination that constitutes the association.

(2) Next compare the operational definitions of the predictor and outcome variables for each one-to-one association with the conceptual definitions of the label and try to judge whether each association falls reasonably within the limits set by the label definitions.

Being able to carry out the above steps leads me to recommend the following standards for reviews in which research evidence is discussed in groups. The conceptual definitions of both the predictor and outcome variables for each group lavel should be clearly and fully developed and both operational definitions should be documented for each association so that the match between the former and the latter can be discussed by the reviewer. This recommendation involves increasing the size and bulk of a review which flies in the face of page limitations not to mention reader fatigue. Perhaps the answer is a supplemental appendix to be made available by the author of a review to each reader who is curious enough to want to check how associations were grouped together. No review of research on teaching effectiveness that I have read or written meets this relatively high standard. Let me plead guilty to lack of attention to these matters (see Flanders and Simon, 1969), but the standard is long overdue.

Although Rosenshine does as well or better than many reviewers, he fails to meet these standards in the following way. First, he fails to develop adequate conceptual definitions of the group labels in nearly every case, instead he lets operational definitions speak for themselves. His description of use of student's ideas (1971a, p. 70) is an exception rather than the rule. Second, his lack of attention to the outcome variable at both the conceptual and operational levels is criticized in depth by Gall (1973) and indicates that the outcome variable was ignored in forming groups. Rosenshine asserts that each study included in the review has a measure of student achievement, but this is inadequate if a reader wishes to judge the homogeneity of groups. And third, he fails to discuss the consequences of the matches and mismatches between conceptual definitions and operational definitions of associations although he warns the reader of considerable heterogeneity, among operational definitions. As a result, it is impossible for a reader--just as it was for the four critics of this Sumposium--to check on the heterogeneity of Rosenshine's groups except by rereading the original reports cited in the review.

What is most disturbing about at least one of Rosenshine's groups is that by using only the information that he provides on the operational definitions of the predictor variables, the four critics of this Symposium believe that he was much too optimistic about what fits together. For example, in discussing variability (with Furst, 1971, p. 45) the authors interpret one or two items from a test which are part of a larger

factor[8] (Solomon, Bezdek, and Rosenberg); several studies which involved adult ratings (like Fortune); listing instructional aids to be found in the room (Anthony); a biographical inventory about the teacher (Walberg); isolating three items from a pupil response scale (Torrence and Parent); and the range of i/d ratios (Flanders). The four critics simply cannot imagine any reasonable, conceptual definition of variability which would admit such a wide range of operations. If Rosenshine had developed a clear conceptual definition of variability and discussed the match between it and the pair of operations for each association, I am certain that he would be dissatisfied with this group.

We might note in passing that diversity of operational definitions in the quantification of a variable within the same study is a very desirable and a powerful form of assessment. Thus if a teacher were rated by an observer on criticism, rated by students on criticism, and then systematically observed using a category system that separated criticism from other teacher statements, all in the same study, a more powerful analysis would come from different configurations of data.

Samples of teachers and pupils. One-to-one associations are products of a research study and each study involves at least one sample of students and their teachers. On one group Rosenshine places college

---

[8]Heath and Nielson (1973, p. 10) assert that separate tests for items must be made in order to determine whether or not they are associated with the outcome variable. They insist that high factor leadings for the items involved and a significant association for the entire factor with the outcome variable, taken altogether, constitute insufficient evidence. One might also note the interpretation of paper-and-pencil items one or two at a time involves great risk compared with a well developed monotonic scale consisting of more than ten items.

student-teachers and experienced teachers, or in another group he will combine teachers of pre-school, elementary, high school and evening school. Similarly student samples may include pre-school students up through adults. In response to this kind of heterogeneity, Heath and Nielson (1973) criticized Rosenshine by writing "... the idea that effective teacher behavior might be different for different age groups is ignored when conclusions are drawn from such a collection (p.11)." Patterns of teacher-pupil interaction are quite different for very young children who do not have independent study skills like reading and writing compared with children who have these skills. It is misleading to combine the results or to expect the same results from widely different samples of teachers and students. Results are often further confounded because variation of this kind often coexists with variation of other features of a learning situation in ways which are discussed in the next section. It is difficult to specify a standard about sampling for reviews, but one suggestion might be phrased as follows. Results from widely different samples of students should not be combined into the same group for interpretation; at least one can think of pre-school, primary, intermediate, and high school as possible boundaries between the universes to be sampled.

Situational factors in learning. The length of time for instruction, microteaching versus regular classrooms versus 15 minute lectures, socio-economic level of the home, inner versus outer city, and other features of a learning situation can have very pervasive effects on teacher behavior. Each setting may have much to contribute in the search for knowledge about teaching, but to assign one-to-one associations from widely diverse settings to the same group for the purpose of interpretation may confuse rather than

clarify data trends. Just as the preceding paragraph on samples of teachers and students contained no hard and fast rules, one can only observe that good judgment should be used in combining the results from different learning situations. Perhaps the only standard one can suggest is results from widely different educational settings should not be combined for purposes of interpretation except when such heterogeneity is called for in a carefully designed plan of inquiry.

Summary of grouping in a review of research. In this section, the assumptions of an ideal logic and a practical logic in grouping studies have been discussed. Given questionable homogeneity within groups, the possibility of making less quantitative generalizations has been discussed and discarded as an alternative in writing a review. The steps of (a) providing careful conceptual definitions of group labels, (b) completely citing operational definitions, (c) discussing the match between the former and the latter, and (d) exercising judgment when combining samples and situations have been proposed and illustrated. I would risk the opinion, at this point, that if Rosenshine had followed the above steps he would have identified different groups and coined different labels. As a further aid to this summary, the group of associations from eight studies which Rosenshine listed under clarity is discussed. This is the predictor variable which Rosenshine (with Furst) decided had the strongest supportive evidence and the most promise for future research.

One can begin by noting clarity lacks clarity because there is no conceptual definition given; only operational definitions can be found. For the predictor variable there is a strong theme of "explaining things clearly

while lecturing" in Studies One, Three, and Four[9]; in Study Two the theme is the teacher's skill "in matching instructional materials to the interests of Puerto Rican and Negro first grade students, inner city"; in the Fifth Study, clarity means the "instructor's understanding of his students and not his knowledge of subject matter" for a military instructor teaching future airplane mechanics; and in Studies Six, Seven, and Eight the main theme seems to be overall teaching competence, knowledge of subject matter, and teaching methods. The outcome variable, in this one-to-one association, has several meanings: first, what youngsters ought to know in the first and third grade levels (Two, Seven, and Eight); second, immediate recall of a lecture topic or demonstration lesson (One, Three, and Four); and third, the course objectives for repairing airplane engines (Five) and a college history class (Six).

The operational definitions are more or less reasonable, depending on the standards one would like to see in this kind of research. One of the weaker procedures for quantifying the predictor variable is a single item on "clarity of presentation" filled out by students from the 8th to 11th grades, who are employed to play the role of students in a summer session microteaching clinic, and then averaged for the three to six students (Four). Student achievement tests included nationally standardized tests in the first and third grades, special recall tests for lectures, and course finals for adult classes.

---

[9]To conserve space, reference to the eight studies cited by Rosenshine in his book, Table 3.4, p. 104-106, is made by the numerical order within the table.

The range of samples and situations is quite extreme. In Studies Two, Seven, and Eight, regular classes, taught for two semesters, and the first and third grade levels were involved. Study Two made use of inner city Puerto Rican and Negro children. Studies One and Three involved experienced teachers giving short lectures or demonstration lessons 10 to 15 minutes long. Study Four involved college interns with no teaching experience, attending a microteaching clinic during the summer. Study Five was concerned with teaching military airplane mechanics and Study Six, the college course in history, was given as an extension class in the evening.

Taken altogether, the studies supporting clarity represent a mixed bag. In spite of this, or because of this, Rosenshine sees considerable promise for future research in the evidence reported and the studies reviewed. Just what direction future research might take remains unclear since there are so many alternatives. The recommendation of Rosenshine seems to be a poor choice when one considers the consistency, of concepts, samples, and situations in some other groups, especially in Chapter Two of Rosenshine's book in which teacher reactions such as criticism and use of student ideas can be found.

Technical Problems in Writing Reviews of Research on Teaching Effectiveness

In this second section, brief observations are made about some technical problems in reviewing research on teaching effectiveness. Each problem discussed could be a criticism directed at a researcher by a reviewer ot it might be directed toward a reviewer.

The Unit of Sampling

Even though there are hundreds of universes which might be sampled in the study of teaching and its effects, it is practical to divide opinion about what to sample into just two positions: those who agree with Rosenshine that the unit of sampling mist be the teacher versus those who are willing to admit other alternatives depending on the research design. In his book Rosenshine writes --

> *"The class rather than the number of students appears to be the appropriate statistical unit for research of this type because the investigator wishes to generalize to the behaviors of teachers. Studies in which the student was used as the statistical unit were not excluded from this review, but some note is usually made ..." (p. 17).*

in another paragraph he writes --

> *"One study (Hunter, 1968) was completely re-analyzed by this reviewer in order to provide data on correlational procedures using the class as the statistical unit (p. 17)."*

in making recommendations about future research, he writes --

> *"There is no orthodoxy for statistical analysis, other than the need to use the classroom, or subgroups within the classroom, as the unit of analysis instead of each student as the statistical unit." (1971b, p. 84)*

(with Furst, 1971) the authors describe experimental studies --

> *"In order to furnish conclusions which can be applied to teacher education programs, we need studies in which (1) the teacher is the statistical unit of analysis, (2) ..." (p. 41)*

One might choose the teacher as the unit to be sampled because "statisticians" give such advice, but this is a superficial response to a complex problem that offers many different alternatives.

There are few decisions a researcher makes that have greater consequences than deciding on the unit of sampling. Rosenshine emphasizes teachers rather than teaching throughout his writing and this choice becomes a pervasive feature of his thinking. The researcher who decides to evaluate teachers will posit a universe of teachers, he will follow well known procedures to choose a sample from this universe, and the "N" which sets the degrees of freedom in his analysis will be the number of teachers involved. As the science (and art) of research on the effects of teaching progresses, it is very unlikely that researchers will remain satisfied with this rigid prescription. Defensible designs in which the unit of sampling is a student, or an encounter, or a single act, or a pattern of acts can and no doubt will be developed.

The student as the unit of sampling. It is possible to design research on teaching in which the student is the unit of analysis. One rationale is that patterns of teaching can create treatments and the effects

on pupils of different treatments can then be analyzed. The requirement here is that the burden of proof rests with the researcher to show that variables of interest are reasonably homogeneous within treatments (however these conditions may have been created), that outcome variables reveal significantly different variances when within treatment versus between treatment comparisons are made, and that differences between means occur in a direction that supports the theoretical hypotheses. One possible approach (Flanders, 1965, p. 50) is to test the differences between the means and variances of just two classes on a pupil attitude inventory and demonstrate that there are no significant differences. Next, compare these two combined classes with a third and if there is no significant difference combine all three. Then compare these three, in a similar fashion, with a fourth, and so on. By starting at one end of a distribution of means, a significant difference will result, sooner or later, and when it does, that class is not added to the group. In one study using this procedure, three classes were grouped as high, positive attitude and six classes were identified as low, negative attitude. Each group formed a homogeneous cluster and the two groups were significantly different from each other with regard to pupil attitude. In this case it is possible to assume that each group represents a "treatment" insofar as pupil attitude is concerned.

It is also possible to create treatment differences by using one or two trained teachers (Flanders and Amidon, 1961, or Schantz, 1963) to which students are exposed. Teachers are trained to create particular patterns of interaction and systematic observation can be used to establish that within treatment homogeneity and between treatment

differences exist. Classroom sessions can then be created with these trained teachers and combined with the random assignment of students to each treatment. Under these circumstances using the teacher as the unit of sampling makes no sense at all.

Other sampling units. It is quite probable that when research on teaching behavior reaches the more advanced stage of becoming concerned with teaching strategies that a pattern of behavior which characterizes a strategy will become the sampling unit. An early attempt at this kind of analysis can be found in Flanders (1960, pp. 95-109) who hypothesized the strategy of shifting from indirect to more direct patterns of teaching during a two week unit. Others who have studied strategies include Spaulding (1965), and Freitag (1970). As encoding systems become more sophisticated and experimental conditions more precisely controlled, we should expect that researchers will be able to satisfy the requirement already mentioned and can create the necessary conditions within and between treatments.

Problems with correlation

The most consistent feature of research on teaching effectiveness that requires observation is the relatively high cost of adding one additional teacher to the study. Often the teacher is the unit of sampling and the "N" is small. Under these circumstances linear product-moment correlation is not likely to be a wise choise. There is insufficient space here to discuss the various issues that are involved. Such a discussion should include considering the lack of symmetrical distributions (J shaped curves) for between class scores on many typical variables,

the "inverted U hypothesis" of Soar (1968), and modern safeguards in the use of regression analysis which make it much more applicable to field studies.

Rosenshine encourages the use of correlation and questions the use of analysis of variance (Rosenshine, 1971a, p. 26). It is possible that he is describing the analytical habits of the researchers whose work he reviews and not really giving advice for future research. When he does give advice[10] it may read like the following excerpt.

> " . . . it does not seem appropriate for investigators to limit themselves to any given level of statistical significance or to any one set of statistical procedures. Rather, a variety of procedures should be used. . ." (with Furst, 1971, p. 63).

New Knowledge of Representative Samples

It should be perfectly clear that representative samples of teacher and students will focus attention on what is going on in today's schools, especially average, ordinary teaching practices. Non-normal samples of gifted teachers or special experiments are necessary to investigate creative, unusual teaching practices which may have the most to offer if we are to discover more effective patterns of teaching or contribute new knowledge to teacher education.

Rosenshine is especially critical of research on teaching when real teachers teaching real kids in practical school settings are not involved. He criticizes "laboratory" studies (with Furst, 1971, p. 40), objects to experiments in which one person acts like two different teachers (in order

---

[10] This advice is curiously inconsistent with his own interpretation of research.

to provide different patterns of teaching), and goes on to criticize laboratory studies by saying --

> "... periods of instruction are seldom longer than three hours... each subject (in an experiment) studies individually, without group interaction; the treatments are usually highly structured; and the 'teacher' is the experimenter or his assistant." (with Furst, 1971, p. 40)

Rosenshine is remarkably flexible about following his own advice and chooses to review a number of studies in which the time for teaching is less than three hours. For example, among the studies he reviews are exceptions which came mostly from Stanford Unversity: Belgard, Rosenshine, and Gage is a study in which two 17-minute learning periods are used; one study by Fortune and another by Fortune, Gage, and Shutes involving three 10-15-minute microteaching episodes; Penny has two 45-minute lessons; Rosenshine has two 15-minute lectures; and Shutes, two 45-minute lessons. Another example is the three out of the eight studies listed under the variable clarity, which involve lessons or lectures which are 10 to 15 minutes long.

The Problem of Replication

Research on teaching effectiveness has probably progressed far enough to propose that all research conclusions should be replicated (i.e., verified with an independent sample) before they are taken very seriously[11] in the summarization of pedagogical knowledge. In the

[11] Similarly, professorial promotions in education might better be based on the number of journal articles, books, etc. that are reprinted one or more times, rather than printed only once.

research reviewed by Rosenshine there are several examples of replication such as studies by Beiderman, Harris, Flanders, and Waller. In this section, the problem is to decide how replication can best be handled in a review of research. One alternative is to follow the lead of Rosenshine and merely list replications as no more and no less than one more study investigating a particular one-to-one association. However, this procedure seriously underestimates the basis for having greater confidence in replicated findings. My own research on responsiveness may serve as an illustration.

Responsiveness[12] is a variable which can be quantified by combining category totals from the FIAC system (Flanders, 1970, p. 102). One measure is the i/i + d ratio which is calculated by adding the total tallies of categories 1 + 2 + 3 and dividing by the total tallies in categories 1 + 2 + 3 + 6 + 7. The more responsive the teacher is, the higher this ratio becomes. Associations between this ratio and two outcome variables, adjusted content achievement and positive pupil attitudes, have been investigated in a number of different studies. Replication enters in because these associations have been investigated at the 2nd, 4th, 6th, 7th, and 8th grade levels.

The results in the second grade sample failed to support the associations mentioned above, but in the other four samples the associations

---

[12]Response and initiation are to be preferred to indirect and direct, although these words refer to the same variables. Like Rosenshine, Flanders makes changes when he rewrites (see 1970, p.102).

were greater than would be expected by chance, using "N" as the number of teachers, a "t" test, and assuming a one-tailed test of significance. It is my purpose to discuss the 4th, 7th, and 8th grade samples because they would be the studies most likely to qualify as replications.

It is impractical to think of perfect replications in research on teaching. Instead, replication is a matter of degree. Some kind of chart, such as the one in Table 1, is helpful in deciding when a replication may have occurred. The chart shows features of a study which may

Table 1

Features of the 4th, 7th, and 8th Grade Studies

	4th	7th	8th
School setting	self-contained classroom	2 hour block, junior high	one hour class, junior high
Subject matter	two-week unit on New Zealand	two-week unit on New Zealand	special two-week unit in math
Sample	16 observed, selected from 72 by pupil attitude score	same as 4th except 15 selected from 63	same as 4th except 16 selected from 85
Test of Achievement	designed to fit unit of study	designed to fit unit of study	designed to fit unit of study
Observation system	trained to use FIAC	trained to use FIAC	trained to use FIAC
Days of observation	all ten teaching days	1st, middle, and last two days of ten teaching days (6 days)	1st, middle, and last two days of ten teaching days (6 days)
Period of instruc-	one hour, 15 minutes	about two hours	about one hour

be taken into consideration. These three studies come fairly close to being replications in analyzing the association between teacher responsiveness and both pupil achievement and positive pupil attitude. Statistically significant results supporting both associations are claimed for each of the three studies based on "t" tests. (See Flanders, 1969, p. 42 for the 4th grade; Flanders, 1960, for the 7th and 8th grades--see also Rosenshine's discussion in his book, p. 40 for these latter two grades). Yet in his book (p. 77ff), Rosenshine has chosen to report the results of all three studies as not significant.

Special interpretation of replicated findings. In many different books on elementary statistics there are discussions about combining several independent tests of the same hypothesis. Two formulas for estimating a test of the same hypothesis when several different studies are combined can be found in Winer (1962, p. 44). They are --

$$(1)\ X^2 = 2\left(\sum -\ln P_i\right) \quad \text{and} \quad (2)\ z = \frac{\sum t_i}{\sqrt{k}}$$

The main thrust of these two formulas can be summarized by saying -- if a study is repeated several times with independent samples in order to test the same hypothesis it is possible to combine the results in ways that take advantage of the increased information which becomes available, the word increased referring to data from several studies rather than just one.

In the case of the three studies at the 4th, 7th, and 8th grade levels, mentioned above, the results reported separately included the following product-moment correlations: for the 4th grade, N = 16, r = .31;

for the 7th grade, N = 15, r = .48; and for the 8th grade, N = 16, r = .43. Only the 7th grade correlation would occur by chance at about the 0.05 level. However, a t-test of the association in each of the three studies was significant at the 0.05 level or better. When using the Chi-square test of formula (1) above, at six degrees of freedom, the support for the hypothesis could be expected to occur by chance at less than the 0.01 level.

The reviewer's responsibility. It is not the responsibility of a reviewer to calculate the combined results from replications. Nevertheless, it is his responsibility to point out replicated findings and interpret them in ways that are responsive to the increased information. This latter responsibility is not met adequately by reporting three non-significant correlations or by choosing a policy which simply counts each study as equal to any other study no matter to what degree a replication may exist. In my own research on teacher responsiveness, for example, replication exists to the following extent: in six out of seven studies significant results occurred if either measures of positive pupil attitude OR content achievement is a pted as a desirable learning outcome. In studies conducted by independent researchers, one can state that in 11 out of 15 studies, some desirable educational outcome has been found to be significantly associated with some measure of teacher responsiveness using the FIAC system of observation. In no case was there a significant negative association.[13]

---

[13]Flanders, 1970, p. 410.

The Main Purpose of the Researcher

There is an interesting dilemma, for those who choose to review research on teaching, in that the reviewer can keep track only of the main purposes of research OR he can dig out, extract, and then regroup odd bits of pieces of data which were not part of the main thrust of a research project. One way to illustrate this is to point out that Rosenshine cited many more than 150 separate one-to-one associations from no more than 70 separate studies. In one case he cited 54 separate one-to-one associations from just five studies conducted by the same researcher. Presumably, the more a reviewer digs out, extracts, and regroups the data, the more enterprising he will appear to be. Rosenshine can be commended for this aspect of his work, but there is one important qualification which can be illustrated most easily by referring to my own research.

In his book (p. 164) Rosenshine cites a nonsignificant one-to-one association between percent teacher talk and student achievement in each of several research projects which I conducted. My question is how will the reader of the review learn that such a relationship was not expected and the research was not designed to provide a fair test of that particular association. The possibility that this is not an isolated example can be inferred by noticing how often single items from larger scales are used to quantify the predictor variable, for example, this occurs in three out of eight variables under the label clarity. However, the incidence of such citations is not at issue here, instead the question is what is a proper procedure for summarizing such a data?

I would strongly suggest the following guideline: A reviewer should decide whether the association being cited was or was not a primary topic of the research and communicate this information to his readers. When the association is not a primary target of research, an appropriate phrase can be used such as -- "In a study not designed to test the association, Smith found that . . ."

The Kind of Knowledge That We Seek

Deciding whether the progress in analyzing teaching effectiveness during the last two decades is remarkably productive or dismally inadequate depends on one's point of view. For oldtimers, like myself, the progress may appear more remarkable than dismal because we can remember the state of affairs two decades ago compared with today's scene. During the last twenty years schemes have been developed for the systematic observation of interactive events, computers have been invented and programs designed to handle huge quantities of data, initial attempts have been completed to design tests of subject matter which are more sensitive to variation in classroom learning activities than are traditional tests; techniques for assisting a person who wishes to analyze his own teaching behavior appear to be more and more promising, and even a hesitant step or two has been taken in the search for mathematical models which will help us understand and cope with chains of events. All of this progress is cause for optimism.

Yet Rosenshine's opening sentence in his article with Furst (1971, p. 37) is "This review is an admission that we know very little about the relationship between classroom behavior and student gains. It is a plea for more research on teaching... for educational researchers and teacher educators to devote more time and money to the study of classroom teaching." When these statements are taken at their face value, they present a rather dismal picture. Nevertheless, the details of this picture, as painted by Rosenshine, were much too cheerful and optimistic

for Heath and Nielson (1973) who judged the data base too weak for the conclusions that Rosenshine reached. How are these conflicting perspectives to be resolved? How does one answer a Dutch student who asked me in Nijmegen, the Netherlands, on May 10, 1972, "My doctorate Committee wants to know why I propose to investigate classroom interaction when Rosenshine has shown that teaching behavior does not affect student learning?"[14] Is the viewpoint of Rosenshine a natural consequence of his review procedure? What kind of knowledge do we seek? These are some of the questions which guide what is written in this third section of this paper.

The Limitations of Simple, Linear, One-to-One Associations

As we seek the pedagogical knowledge which has the most worth, we should be skeptical of each type of knowledge nominated. What are the strengths and weaknesses of simple, linear, one-to-one associations?

Curvilinear associations. Soar's inverted "U" hypothesis (1968) proposes that a measure of teaching behavior will have an optimum incidence at which its association with an outcome variable is at its maximum, that a higher or lower incidence of this teaching behavior will create conditions which are less effective, and that the optimum level will vary with different types of student learning. In a secondary analysis of my data

[14] The right answer, of course, is that Rosenshine did not show and was not trying to show that teaching does not affect student learning. Yet how would reasonably qualified persons obtain this misconception?

from the 6th, 7th, and 8th grade levels, Coats (1966) independently found similar curvilinear relationships, but placed them in the Appendix of his thesis without interpretation. In a secondary analysis of my 2nd and 4th grade data, Nuthall[15] is currently finding additional evidence supporting curvilinearity.

One important consequence of curvilinearity is that one study may report a positive linear association and another study may report a negative linear association, depending on whether the sample was mostly above or below the optimum level. Under these conditions, the interpretation of positive and negative linear associations is impractical and misleading, as is the magnitude of the correlation itself, unless the procedures for assessing the predictor variable permit comparisons between the samples of two or more studies.

Sequences. Interactive events occur one after another in what may be thought of as a chain of events. It has been proposed (Flanders, 1970, Chaps. 1 and 9) that a sequence of events forms patterns, and a sequence of patterns forms a teaching strategy. Strategies, in turn, probably occur in cycles which occupy even longer segments of time. In my judgment, there is a very high probability that the meaning of almost every event, if not all, depends of what occurred before and after it. The existence of chains is now well established and some evidence about the variation of events within strategies has been reported (see Flanders, 1965, pp. 102-108; Spaulding, 1965; Bellack, 1966, to name a few).

---

[15]A personal letter from Graham Nuthall, dated July 13, 1973, indicates that the work is underway.

One consequence of chain phenomena is that two classrooms in which measures of a particular teaching behavior are equal in incidence but different in context may then be quite different in terms of the effects of such behavior. This kind of reasoning may lie behind Rosenshine's suggestion (his book, p. 61) that the context of teacher criticism should be taken into account. In any case, if the same events can have different meanings according to context, then a misleading interpretation is possible for a group of one-to-one linear associations. One-to-one associations from research designs in which situational cues are taken into account would be more valid.

Multiple correlations. Nearly all multidimensional studies, like those of Soar for example, show that a single type of event can be associated with several outcomes and a single outcome can be associated with several kinds of events. Under these circumstances it is difficult to interpret a single one-to-one association which has been so to speak, pulled out of its own context.

The existence of multiple correlations between predictor and outcome variables and the consequences of this for interpreting single, one-to-one associations are likely to be different in each study. Nevertheless, most one-to-one associations are incomplete bits of information which have been plucked from a larger mosaic, but it is the mosaic which helps to make their interpretation more valid. Thus, the isolation of such a bit, no matter whether this is done by the researcher or the reviewer--or both--does involve a risk. When the reviewer then places this bit of information into a new mosaic of his

own design, another risk is taken. Interpreting a one-to-one association may involve double jeopardy, first, when it is taken out of context, and second, when it is placed into a new context. For those who believe that one-to-one associations can be used as a guide to future research, like Rosenshine, the assumption is that the entire process provides an opportunity for more valid interpretation rather than less valid.

Summary. It is precisely because one-to-one associations, standing alone, do not lend themselves easily to interpretation that we expect reviewers to cluster them into some kind of organization which gives a better perspective. As has been discussed on earlier pages of this article, the essence of such an organization is the logic used to group the associations. Given the above limitations of one-to-one associations, reviewers should be reminded that interpreting one-to-one associations involves considerable risk and this risk can only be reduced by using special care in the grouping of the associations and in the interpretation of the results.

The Kind of Knowledge About Teaching and Learning That We Need

The decade of the seventies may be a turning point in that researchers will recognize and respond to the conceptual and procedural requirements of high quality research on teaching. It is in describing this response that the views of the Symposium participants, Rosenshine, Heath and Nielson, Gall and myself, are more likely to be in general agreement.

1. Knowledge that fits together. If the meaning of behavior is "situation specific," that is, the same behavior can have a different meaning in different situations, then we need information about events that covers a wide range of phenomena. The instructional materials, the availability and use of space, the expected student outcomes compared with their present performance, and the role of the teacher as coordinator, are all aspects of teaching and learning to be described by parameters. These parameters should be carefully conceptualized in the planning phase of research after, and only after, direct contact with[16] the situations and behaviors has occurred. For each parameter, the number of variables must necessarily be limited, but most important of all, the variables, parameters, and sets of parameters must fit together.

Information is more likely to fit together when it is conceptualized, quantified, and synchronized according to the rules of one or more theoretical models. A student at his desk, completing a worksheet by himself, initiates a question to the teacher, and the teacher responds by giving directions, all this may be included in the description of a short time-frame. One variable from each parameter is scored in each time-frame and that includes the time-frames that precede and follow. In this case the model coordinates instructional material, space and its use, the student act, and the teacher act. The model may have the

---

[16] "Contact with" may be too weak, perhaps "immersion in" would be a better phrase.

capability of assigning such a time-frame to the set called "individualized seatwork, student compliance, and teacher direction." In short, the information collected fits together.

2. Short term and long term outcome variables. No variable, like long term content achievement for example, is any more important or less important than the immediate response of a student, or his perception of recently completed events, or his long term attitude toward learning and toward the teacher, or his distractability scores averaged for the year, etc. If any general admonition is appropriate, it would be that a research design with a single outcome variable is inadequate, not cost/effective for the researcher, and contains the risks that have just been discussed.

Just as a model helps to ensure that information will fit together, it also helps to identify which variables will be scored as short term and long term educational outcomes. We might note in passing that long term outcomes which cannot be explained are not very useful in understanding teaching and learning. What are sometimes called "intervening variables" may be short term educational outcomes that are essential to explaining long term results. Thus, this student learned more arithmetic because he felt good about it and his positive feelings, in turn, seemed to be associated with his exposure to positive feedback, and so on.

3. The problem of lateral context. As long as multiple scored time frames are used, there is the possibility that a single type of event can

be interpreted in terms of its context, at least to some degree. When the total quantity of raw behavioral events is quite large, even events which occur less frequently can be classified into different contexts for more accurate interpretation. It would be in this way that we might be able to find out in what situations mild criticism has a positive effect on student effort and when it has a negative effect; when harsh criticism is more likely to occur compared with mild criticism; when praising an alternative behavior can help to extinguish an unwanted behavior and when praise can help to increase the incidence of desirable student behaviors; and so on. When we focus only on one-to-one associations and fail to attend to contextual cues, our generalizations are so far removed from the classroom that teachers in the chalk pits cannot use the results. How and when to use praise, for example, may be questions which are more important to a teacher than knowing that the overall incidence of praise is associated with higher student achievement. As researchers, we must be responsive to such needs.

4. The problem of longitudinal context. There is plenty of evidence from research on different kinds of reinforcement schedules to suggest that when somethings occurs is as important as how often it occurs. The longitudinal context is concerned with what occurred before and after an event of interest. However, we need not restrict our interest to one event. A moving window (after Semmel, 1972) can scan a chain of events such that it focusses on two, three, four,.... up to "n" events thereby creating a context of any desired segment length. It is in this manner that single events, pairs, etc. become patterns and these in turn

can become strategies. We will need theoretical models to help us decide the number of different parameters to be scored within one time frame and we will need additional models which will help us analyze chains using variable segment length. It is this latter type of model which will provide clues about longitudinal contexts.

Summary

This paper has been revised several times in an effort to emphasize the constructive aspects of criticism. All five members of this AERA Symposium are interested in improving the quality of research on teaching effectiveness and the quality of reviews which attempt to summarize this research. The four critics are indebted to Rosenshine for his tireless work in abstracting research articles and for providing the critics with manuscripts that would otherwise not be available or at least difficult to obtain.

The first section of this paper described logical considerations which pertain to procedures for grouping research studies when the purpose is to interpret the results from a large number of studies. The second section dealt with some technical issues with regard to sampling, linear correlation, new knowledge versus current practice, and the interpretation of replicated findings. The third section briefly mentioned four kinds of information which we will seek in our quest for more useful knowledge about teaching and learning.

REFERENCES

Anderson, H. H., and Brewer, H. M. "Studies of Teachers' Classroom Personalities, I: Dominative and Socially Integrative Behavior of Kindergarten Teachers," Applied Psychology Monographs, 1945, No. 6.

Anderson, H. H., and Brewer, J. E. "Studies of Teachers' Classroom Personalities, II: Effects of Teacher's Dominative and Integrative Contacts on Children's Classroom Behavior," Applied Psychology Monographs, 1946, No. 8.

Anderson, H. H., and Reed, M. F. "Studies of Teachers' Classroom Personalities, III: Follow-up Studies of the Effects of Dominative and Integrative Contacts on Children's Behavior," Applied Psychology Monographs of the American Psychological Association, December, 1946, No. 11, Stanford University Press.

Bellack, A. A., Kliebard, H. M., Hyman, R. T., and Smith, F. L., Jr., The language of the classroom. New York: Teachers College Press, Columbia University, 1966.

Bush, R. N. The teacher-pupil relationship. Englewood Cliffs, N.J.: Prentice-Hall, 1954.

Campbell, J. R., and Barnes, C. W. Interaction analysis--a breakthrough. Phi Delta Kappan, 1969, 7, 587-590.

Coats, W. D., "Investigations and simulation of the relationships among selected classroom variables." Unpublished doctoral dissertation, The University of Michigan, 1966.

Ennis, R. H. Logic in teaching. Englewood Cliffs, N.J.: Prentice-Hall, 1969.

Flanders, N. A. Analyzing teaching behavior. Reading, Mass.: Addison-Wesley, 1970, pp. 448 + xvi.

Flanders, et al. Teacher influence patterns and pupil achievement in the second, fourth, and sixth grade levels, Volume I and Volume II. Terminal contract report (OE-4-10-243) Project #5-1055, USOE, HEW. Ann Arbor: School of Education, University of Michigan, 1969. Volume I: pp. 100 + xii. Volume II: pp. 374 + vi.

Flanders, N. A. Teacher influence, pupil attitudes and achievement. Cooperative Research Monogaph No. 12 (OE-25040). Washington, D.C.: U. S. Government Printing Office, 1965, pp. 126 + ix.

Flanders, N. A. Teacher influence, pupil attitudes and achievement. Cooperative Research Project #397 (USOE). Minneapolis: School of Education, University of Minnesota, 1960, pp. 121 + appendices.

Flanders, N. A., and Amidon, E. J. The effects of direct and indirect teacher influence on dependent-prone students learning geometry. Journal of Educational Psychology, 1961, 52, 286-291.

Flanders, N. A., and Simon, A. Teacher effectiveness. In R. L. Ebel (Eds.), Encyclopedia of Educational Research, New York: Macmillan, 1969, 1423-1437.

Freitag, N. F. Social issues classroom discourse: a study of expository inquiry-nonprobing, inquiry-probing classes. In B. G. Massialas (Director) Structure and process of inquiry into social issues in secondary schools. Cooperative Research Project #1352, USOE, Hofstra University, 1965.

Gage, N. L. Program on teaching effectiveness, current organization and plans. In house paper, Stanford Center for Research and Development in Teaching, May, 1973.

Gall, M. D. The problem of "student achievement" in research on teacher effects. AERA paper, 1973, available from the author.

Heath, R. W., and Nielson, M. A. The myth of performance based teacher education. AERA paper, 1973, available from the authors.

Kaplan, A. The conduct of inquiry. San Francisco: Chandler, 1964.

Rosenshine, B. Interaction analysis: a tardy comment. Phi Delta Kappan, 1970b, 8, 445-446.

Rosenshine, B. New directions for research on teaching. In How teachers make a difference. Washington: U.S. Government Printing Office, 1971b. (Catalog No. HE 5.258.58044).

Rosenshine, B. Teaching behaviours and student achievement. London: National Foundation for Educational Research in England and Wales, 1971a.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970a, 40, 647-662.

Rosenshine, B., and Furst, N. Research in teacher performance criteria. In B. O. Smith (Ed.) Research in teacher education: A symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

Rosenshine, B., and Furst, N. The use of direct observation to study teaching. In R.M.W. Travers (Ed.) Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Ryans, D. G. Characteristics of teachers. Washington, D.C.: American Council on Education, 1960.

Schantz, B.M.B. "An experimental study comparing the effects of verbal recall by children in direct and indirect teaching methods as a tool of measurement." Unpublished doctoral dissertation, Pennsylvania State University, 1963.

Semmel, Melvyn I. "Toward the development of a computer assisted teacher training system." In N. Flanders and G. Nuthall (Eds.) The classroom behavior of teachers, Vol. XVIII, UNESCO Institute for Education, Hamburg, Germany, 1972.

Smith, B. O., and Meux, M. O., "A study of the logic of teaching." In Anita Simon and E. G. Boyer (Eds.), Mirrors for behavior: An anthology of classroom observation instruments. Volume IV, 121 S. Broad Street, Philadelphia, Pennsylvania: Research for Better Schools, Inc., 1967.

Soar, R. S. "Optimum teacher-pupil interaction for pupil growth." Educational Leadership. 1968, 26, 275-280.

Spaulding, R. L. Achievement, creativity, and self-concept correlates of teacher-pupil transactions in elementary school classrooms. Cooperative Research Project #1352, USOE, Hofstra University, 1965.

Taba, H., Levine, S., and Elzey, F. Thinking in elementary school children. Cooperative Research Project No. 1574, USOE, 1964.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.