DOCUMENT RESUME

ED 104 832 SP 009 066

AUTHOR	Bergquist, Constance C.; Graham, Darol L.
TITLE	Methodological Considerations for the Validation of Teacher Competencies.
PUB DATE	Apr 75
NOTE	15p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., April 1975)
EDRS PRICE	MF-$0.76 HC-$1.58 PLUS POSTAGE
DESCRIPTORS	Criterion Referenced Tests; *Educational Research; Higher Education; Measurement; Performance Based Teacher Education; Performance Criteria; *Research Methodology; Teacher Behavior; *Teaching Skills; Validity

ABSTRACT

Findings of educational research offer little indication of a relationship between teacher performance and pupil achievement. A review of reported studies suggest that the difficulty may be methodological. Most investigations are either correlational studies or comparative studies using control group designs. Most correlational research is characterized by a lack of theoretical guidance or explicit rationale, and results are often not replicable. Group comparison studies are impeded by difficulty in establishing preexperimental sampling equivalence of groups, inadequacy of traditional data analysis techniques, and within-group variance. Two alternative methodologies which may be useful for educational research are the time-series and multiple baseline designs. The former is an example of a design where a discontinuity in a series of measurements that occurs coincidentally with the introduction of the treatment variable suggests a possible relationship. The multiple time-series design introduces a nonequivalent control situation in order to rule out the possibility that other potential influences on the dependent variable were responsible for the observed changes. The multiple baseline design is an extension of the time-series model in which the researcher attempts to replicate the effect of a treatment variable across a number of behaviors. Baseline behaviors are established, treatments applied successively, and behavior changes recorded. (HMD)

METHODOLOGICAL CONSIDERATIONS FOR THE VALIDATION OF TEACHER COMPETENCIES

Constance C. Bergquist
Florida State University

and

Darol L. Graham
University of Texas Health Science Center at Dallas

Paper presented at the annual meeting of the American Educational Research Association, Washington, April, 1975

Methodological Considerations for the Validation of Teacher Competencies

Constance Bergquist
Florida State University

Darol Graham
University of Texas Health Science Center at Dallas

One of the basic assumptions guiding the operation of our educational system is that the teacher causes or at least facilitates pupil achievement. Accordingly, teacher education programs have assumed the role of training teachers to demonstrate those behaviors believed to enhance or produce classroom learning. Although considerable effort has been exerted to provide evidence of teacher effectiveness, the findings of educational research offer little indication of a relationship between teacher performance and pupil achievement (Rosenshine, 1971; Rosenshine and Furst, 1971; and Rand, 1971).

The failure of research efforts to provide conclusive and consistent evidence of instructional effectiveness can be attributed to a number of factors, however, a review of reported studies suggests that many of the major problems are methodological in nature. The present paper reviews some of the more prevalent methodological problems and identifies alternative approaches that should receive greater consideration, particularly for the validation of teacher competencies.

Methodological Problems

As indicated by Gall (1973) and Potter (1973), an adequate methodology must effectively handle three aspects of the research situation. First, the teacher behavior or independent variable must be carefully defined, controlled, and measured. Second, the dependent variable, student achievement, must also be

operationally defined and accurately measured. Finally, a design must be employed that is capable of showing any relationship that may exist between the independent and dependent variables. Studies that fail to give adequate consideration to any of these critical elements may draw unjustifiable conclusions, either positive or negative.

Teacher Performance Variables

In the past, many researchers have failed to operationally define the teacher behavior under consideration. In fact, many of the teacher behavior variables receive definition only as constructs represented by a sub-scale score on some personality inventory or a subjective rating on some global behavior scale. Rosenshine (1971) discussed a number of these high inference variables and the problems associated with their use in teacher behavior research. Even the low inference variables described by Rosenshine are seldom defined with sufficient clarity to permit replication of a study.

Studies which attempt to validate teacher competencies must begin with a precise definition of the variable of interest. The definition must specify, in behavioral terms, all subordinate skills that comprise the competency. In addition, behaviors that are not encompassed by the competency should receive comparable attention as an initial step in providing adequate experimental control. Once the characteristics of a teaching competency are carefully delineated, the development of reliable measures of the competency becomes a more manageable task.

Normally, the measurement of teacher behavior should be a two-step process. The process first involves determination that the teacher possesses the capability of interest and, secondly, that the teacher actually employs the competency in the experimental setting. Failure to determine in advance that the competency exists can waste many valuable hours of experimental time. Failure to substantiate that

an acquired competency is demonstrated properly during the study can render the results of the investigation meaningless. Both steps are essential in a well-conceptualized competency validation experiment.

Recent interest in performance-based teacher education has emphasized the need for improved procedures for measuring teacher behavior. This movement has focused attention upon behavioral specification of teacher performance. It is believed that performance-based teacher training programs may provide both the stimulus and the environment for meaningful research in the area of teacher effectiveness.

Pupil Performance Variables

Many studies have failed to demonstrate instructional effects due to problems associated with the measurement of the dependent variable. Frequently, insufficient attention is devoted to the task of ensuring valid and reliable assessment of pupil performance. Adequate measurement of pupil performance is dependent upon the same type of rigorous, operational delineation of the behavior of interest that is required for measures of teacher performance.

Many investigators have defined student behavior in terms of performance on standardized achievement tests. Gall (1973) suggested that a major problem in the use of such achievement tests for measuring student performance is the failure of the test to provide a measure of the content taught by the teacher. Tests with demonstrated validity for their designed purpose may not provide a valid measure of pupil achievement relative to the specific performance objectives identified for a given research study. In addition, standardized achievement measures may lack sufficient sensitivity to detect behavioral changes occurring in studies of short duration.

Other problems associated with pupil performance assessment include the operation

of statistical regression where groups have been selected on the basis of their extreme scores (Campbell and Stanley, 1963), and the artificial restriction of the range of gain scores when the test is not appropriate for the aptitude level of the class (Gall, 1973). A thorough treatment of the relationship between regression artifacts and the use of matching, gain scores, analysis of covariance, and partial correlation was presented by Campbell and Erlebacher (1970).

Gagne (1970) referred to standardized achievement tests as correlated measures of pupil behavior and discussed the difficulties involved in interpreting results of studies in which they are used. In particular, Gagne suggested that such tests possess many of the characteristics of intelligence tests and are not valid as direct output measures. Anastasi (1968) also indicated that if a standardized achievement battery is used, the analyzed relationship may be that between teacher behaviors and student aptitude instead of performance. In response to this problem, Gagne called for studies which employ more direct or proximal output measures, commonly referred to as criterion-referenced tests, to ascertain student performance instead of correlational measures.

While the use of direct or criterion-referenced achievement measures may circumvent many of the problems associated with norm-referenced instruments, a new set of concerns are introduced. The measurement practices which most frequently negate research efforts are: (1) assuming that content validity is either inherent in criterion-referenced tests or that it can be readily judged by a content expert and (2) constructing tests comprised of one or two items per objective and assuming their reliability. Although theories and methodologies for criterion-referenced measurement are presently inadequate, existing guidelines should not be ignored by the conscientious investigator. The use of more direct measures of student achievement may contribute substantially to the future success of educational research efforts.

Research Methodologies

The final factor that has hampered teaching effectiveness research is the lack of a suitable methodology for demonstrating a relationship between teacher behavior and pupil performance. Most investigations reported in the literature are either correlational studies or comparative studies employing control group designs. A few of the problems associated with these traditional methodologies are identified in the discussion that follows.

Most of the correlational research is of the type identified by Gall (1973) as "empirical-exploratory" and characterized by not being guided by theory or explicit rationale. The results of such studies often cannot be replicated because of the occurrence of chance relationships in large correlation matrices. Guralnick (1973) also pointed to the confusion that occurs in the interpretation of correlational data as a result of the excessive number of variables operating in educational programs. Even correlational research of the variety described by Gall as "hypothesis-testing" may be of little value because of the instruments used to measure the variables. In fact, Gagne (1970) indicated that he perceives the credibility of such investigations as very low and draws practically no conclusions from them.

Group comparison studies are also subject to a number of limitations. As indicated by Baer, Wolf, and Risley (1968), the relationship between teacher behavior and pupil performance must ultimately be demonstrated in the classroom. However, experimental control of relevant variables in the classroom situation is extremely difficult. A major limitation in comparative studies is presented by the difficulty of establishing pre-experimental sampling equivalence of the groups in naturally occurring educational settings. Randomization procedures for establishing equivalent

control groups rarely can be employed since subjects are often selected because of their identified need for treatment. Withholding treatment from certain subjects is often morally and socially unjustified, thus eliminating the availability of appropriate controls.

The typical solution to randomization problems is the use of nonequivalent control group designs. The adoption of such procedures, however, introduces another set of concerns. Campbell and Stanley (1963) discussed at length the possibility of interaction between selection and maturation affecting internal validity in such situations. Another source of internal validity problems in this design is the effect of regression. If either the control or experimental group has been chosen on the basis of its extreme scores, the pretest-posttest gains may well be a product of regression rather than the effect of the treatment.

Even if equivalent groups can be established, traditional data analysis techniques may be inadequate for many investigations. Due to the previously mentioned limitations of norm-referenced measurement procedures. The use of criterion-referenced measures, particularly in mastery learning situations, may produce data with little or no variance. Thus, the analysis of variance techniques typically employed with control group designs may be inappropriate.

Design problems are magnified in the individualized instructional setting that is being adopted with increasing frequency. In many instances, each individual pupil may receive an unique treatment. In addition, unique outcomes may be expected for each pupil. Thus, the sample size for each experimental group may become extremely small, even reaching unity. The problems associated with the use of traditional analysis of variance techniques in such situations should be obvious.

Finally, group comparison studies are also impeded by within-group variability

which may mask treatment effects. The failure of many ATI studies (Cronbach and Snow, 1969) to show significant differences may be influenced by the averaging of treatment gains in comparison to controls or other treatments. This phenomenon was discussed at length in the report published by the Rand Corporation (1971). Light and Smith (1970) argued that the variability within a treatment may be equally as important as the average because a given program may be effective for only a limited segment of the treatment population.

Alternative Methodologies

The problems associated with efforts to use laboratory research methods in applied settings are extensive. Many investigators are turning away from traditional methodologies and are seeking procedures that are more appropriate for the circumstances. The following discussion will attempt to present a number of alternatives that are available. Some of these alternatives are highly familiar to experienced researchers, while others may be less familiar to investigators in the field of education. Perhaps by focusing attention upon these methods, the quality of educational research can be enhanced.

In situations where control groups are not available, researchers often turn to techniques that demonstrate the ability to effect and observe change at prespecified points in time. The time-series experiment (Campbell and Stanley, 1963) is an example of a design where a discontinuity in a series of measurements that occurs coincidentally with the introduction of the treatment variable suggests a possible relationship. The strength of any causal inference, however, is dependent upon the ability of the researcher to eliminate competing hypotheses. In some manner, convincing evidence must be provided that other potential influences upon the dependent variables were not responsible for the observed changes.

In an effort to rule out alternative explanations for apparent treatment effects

found in time-series experiments, the multiple time-series design (Campbell and Stanley, 1963) introduces a nonequivalent control situation that is isolated from the experimental variable. Failure of a discontinuity to appear in the control data increases the believability of an experimental treatment effect. Another means of lending credibility to the influence of an experimental variable is by repeatedly demonstrating the treatment effect at the will of the experimenter. One of the strongest arguments available to the researcher is the demonstration of repeated success, or as Baer, et al. (1968) succinctly stated, "replication is the essence of believability [P. 95]."

A methodology that exemplifies the concept of replication is the multiple baseline design employed routinely in experimental psychology. This design is actually an extension of the multiple time-series experiment with the introduction of additional control conditions that also receive the experimental treatment but at different points in time. In the most common application of the multiple baseline technique, the experimenter attempts to replicate the effect of the treatment variable across a number of behaviors. After establishing a number of baselines by measuring each behavior over time, an experimental treatment is applied successively to each behavior with any changes being recorded against the baseline of that behavior. An effective treatment variable is indicated by producing a noticeable change at the time the variable is introduced and at no other time. That is, measures of responses prior to and after the treatment application must remain stable. Every replication of the treatment variable resulting in a change from the baseline increases its reliability as being an effective variable.

The replication of an experimental treatment across two or more behaviors of the same individual is just one example of the multiple baseline approach. Application of the same systematic procedures across individuals and across situations were

successfully demonstrated by Hall, Cristler, Cranston, and Tucker (1970). The major benefit of the design resides in the ability to establish a strong inference of a causal relationship through numerous replications under a variety of conditions.

Another advantage of the multiple baseline design is its adaptability to analysis of the behavior of individuals. Examination of within treatment variability may provide explanations of "why" the treatment was effective for certain individuals and, equally important, "why" it was ineffective or less effective for others. Increased emphasis on process rather than outcome variables may facilitate refinement of the treatment and establishment of the conditions in other environments.

The precision teaching movement has provided considerable evidence of the utility of multiple baseline designs in studies of teacher effectiveness. One example of the application of this approach in the validation of teacher competencies was reported by Pennypacker and Pennypacker (1973). The study demonstrated the effectiveness of the Standard Behavior Chart (Pennypacker, Koenig, and Lindsley; 1972) for recording changes in pupil behavior. It is interesting to note, however, that the investigators felt compelled to superimpose a nonequivalent control group design upon the multiple baseline design in order to enhance the credibility of their efforts in the eyes of traditional researchers.

Although most multiple baseline experiments have been concerned with individual behavior, Hall, et al. (1970) suggested that the procedures may apply equally well to the behavior of groups. The comprehensive achievement monitoring (CAM) model presented by Allen, Gorth, and Wightman (1970) for the evaluation of school achievement provides an example of the application of multiple baseline principles to group behavior. Another example is provided by Ellzey (1974) who employed precision teaching procedures in the evaluation of an ESEA Title III project.

The research-service model proposed by Guralnick (1973) is still another adaptation of multiple baseline procedures for the assessment of instructional programs. Designs such as the research-service model and precision teaching methods, that can examine effects on both groups and individuals, offer many advantages for the evaluation of innovative projects.

Graham (1973) developed a model for the validation of teacher competencies that combines the control qualities of the multiple time-series experiment with the replication characteristics provided by the multiple baseline design. In the study reported, the model provided both replication across behaviors defined by performance objectives and replication across individuals attaining common objectives. It was noted that by reassigning pupils during the experiment, the methodology would permit further replication across teachers or situations.

Graham identified a number of other advantages of the methodology for studies of instructional effects. These advantages are as follows:

1. The multiple replications of an experiment provide an extensive and convincing data base from which to draw conclusions concerning treatment effects.
2. The records maintained during administration of the treatment provide a means of examining within treatment variability. Examination of these records may provide more useful information about interaction effects than traditional ATI studies.
3. The design may be employed in investigations of either individuals or groups. In an individualized instructional situation where each pupil may receive an unique treatment, the ability to analyze the behavior of individuals is essential.
4. The design can be employed in naturally occurring educational settings with

students selected because of their need for treatment.

5. The model imposes no restrictions upon the nature of the data collected. The dependent variable may be measured by means of either criterion-referenced tests or norm-referenced tests or, for experiments concerned with instructional efficiency, data may take the form of instructional time rather than achievement.
6. The model is appropriate across a wide range of pupil and teacher behaviors of varying specificity.
7. The model may be applied to studies of varying duration, ranging from short-term experiments to large-scale evaluations.
8. The design provides information concerning retention and transfer effects as well as learning.
9. The design provides evidence concerning the reliability of tests used to measure pupil performance.

In conclusion, methodological problems have seriously hampered educational research in the search for contributors to instructional effectiveness. Although competency-based training programs and criterion-referenced testing techniques offer potential for solution of many of the measurement problems, adherence to traditional research methods tend to negate their impact. It is imperative that increased emphasis be placed upon innovative methodological developments. As a beginning, the time-series and multiple baseline designs should receive greater attention. Experience with these techniques may provide the insight and perspective necessary for the development of more creative and appropriate research methods.

References

Allen, D. W., Gorth, W. P., and Wightman, L. E. An introduction to longitudinal testing using item sampling techniques: Comprehensive achievement monitoring. Paper presented at the annual meeting of the American Educational Research Association, Minneapolis, March, 1970.

Anastasi, A. Psychological testing. (3rd ed.) New York: Macmillan, 1968.

Baer, D. M., Wolf, M. M., and Risley, T. R. Some current dimensions of applied behavior analysis. Journal of Applied Behavior Analysis, 1968, 1, 91-97.

Campbell, D. T. and Erlebacher, A. How regression artifacts in quasi-experimental evaluation can mistakenly make compensatory education look harmful. Disadvantaged Child, 1970, 3, 185-211.

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

Cronbach, L. J. and Snow, R. E. Final Report: individual differences in learning ability as a function of instructional variables. Palo Alto: Stanford University, 1969.

Ellzey, M. A. Developing models for special education. ESEA Title III Grant Application and Report of First Operational Year, May, 1974, Board of Education, Jefferson County, Florida.

Gagne, R. M. Policy implications and future research: a response. In Do teachers make a difference: a report on recent research on pupil achievement. Washington: U.S. Government Printing Office, 1970.

Gall, M. D. The problem of "student achievement" in research on teacher effects. Report A73-2, 1973, Far West Laboratory for Educational Research and Development.

Graham, D. L. Research to relate clinical teacher competencies to exceptional pupil performance: Summary. Paper presented at the Teacher Education Research Conference, Florida Department of Education, Tallahassee, November, 1973.

Guralnick, M. J. A research-service model for support of handicapped children. Exceptional Children, 1973, 40, 277-282.

Hall, R. V., Cristler, C., Cranston, S. S., and Tucker, B. Teachers and parents as researchers using multiple baseline designs. Journal of Applied Behavioral Analysis, 1970, 3, 247-255.

Light, R. J. and Smith, P. V. Choosing a future: strategies for designing and evaluating new programs. Harvard Educational Review, 1970, 40, 1-28.

References (Cont'd.)

Pennypacker, H. S., Koenig, C. H., and Lindsley, O. R. Handbook of the standard behavior chart. Kansas City: Precision Media, 1972.

Pennypacker, H. S. and Pennypacker, S. F. The relation of teacher competence in daily charting to changes in pupil performance in mathematics. Paper presented at the Teacher Education Research Conference, Florida Department of Education, Tallahassee, November, 1973.

Potter, D. A. A research strategy for performance-based education. Paper presented at the meeting of the American Educational Research Association, New Orleans, February, 1973.

Rand Corporation. How effective is schooling? a critical review and synthesis of research findings. Santa Monica, Cal.: 1971.

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

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.