DOCUMENT RESUME

ED 120 262 TM 005 226

AUTHOR Borich, Gary D.
TITLE Sources of Invalidity in Measuring Classroom Behavior.
INSTITUTION Texas Univ., Austin. Research and Development Center for Teacher Education.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [76]
CONTRACT NIE-C-74-0088
NOTE 55p.

EDRS PRICE MF-$0.83 HC-$3.50 Plus Postage
DESCRIPTORS *Classroom Observation Techniques; Effective Teaching; Elementary Secondary Education; Guidelines; Teacher Behavior; *Teacher Evaluation; *Testing Problems

ABSTRACT

This paper is a review of the methodological problems recently uncovered in studying the nature of teacher effectiveness and evaluating the performance of individual teachers. Four problems encountered in the literature are range of measurements, inconsistent instrumentation across similar studies, lack of a generic framework from which to select behaviors to be measured, and use of instruments with inadequate psychometric characteristics. These problems are discussed. From a review of the literature, three general dimensions were selected from the purpose of categorizing classroom behavior and the instruments used to measure it. These dimensions were: (1) stage of behavior on a process-product continuum; (2) level of inference required in measuring behavior, and (3) objectives of the instruction. If the measurement of behavior is viewed as a longitudinal process, four distinct and consecutive measurement stages are apparent: (1) Preoperational (personality, attitude, experience, and aptitude/achievement); (2) Immediate (sign, counting, and rating systems); (3) Intermediate (Likert and Guttman Scales, semantic differentials and check lists); (4) Product (influences other than the teacher, unreliability of the raw gain score, and the teacher's desire to teach to the test). Last, some guidelines are offered for improving the measurement process. (RC)

Sources of Invalidity in Measuring Classroom Behavior

Gary D. Borich

The University of Texas at Austin

This work has been supported in part by the National Institute of Education Contract NIE-C-74-0088, The Evaluation of Teaching Project. The opinions expressed herein do not necessarily reflect the position or policies of the National Institute of Education and no official endorsement by that office should be inferred.

Sources of Invalidity in Measuring Classroom Behavior

Gary D. Borich

The University of Texas at Austin

This paper is a review of the methodological problems uncovered by relatively recent efforts in the U.S. to study the nature of teacher effectiveness and to evaluate the performance of individual teachers. The former concept--teacher effectiveness--derives from almost two decades of research, conducted in this country and elsewhere, to identify the behavioral correlates of "more effective" and "less effective" teachers. The latter concept--teacher evaluation--stems from comparatively recent efforts to design and implement schemes for appraising individual teachers, a practice stimulated primarily by an increasing number of state and local mandates requiring yearly, systematic evaluation of elementary and secondary school teachers.

The organizational framework of this review is sketched below to identify for the reader the bounding points of the discussion.

A. Four Generic Methodological Problems

1. Range of Measurements
2. Instrumentation
3. Frameworks
4. Psychometrics

B. Characteristics of a Measurement Framework

1. Process-Product Stages
2. Levels of Inference
3. Objectives of the Instruction

C. Stages of Measurement

1. Preoperational Stage Measurements

a. Personality
b. Attitude
c. Experience
d. Aptitude/Achievement

2. Immediate Process Stage Measurements

a. Sign Systems
b. Counting Systems
c. Rating Systems

3. Intermediate Process Stage Measurements

    a. Likert Scales
    b. Semantic Differentials
    c. Guttman Scales
    d. Checklists

4. Product Stage Measurements

    a. Influences Other Than the Teacher
    b. Unreliability of the Raw Gain Score
    c. The Teacher's Desire to Teach to the Test

D. Some Guidelines for Improving the Measurement Process

1. Criterion-Referenced Testing vs. Norm-Referenced Testing
2. Relationship between Process and Product
3. Relationship between Performance Measured and Objectives Planned
4. Relationship between Objectives Planned and Objectives Taught
5. Time between Product Measurements
6. Raw vs. Adjusted Gain

In reviewing empirical studies of teacher effectiveness for the Evaluation of Teaching Project, funded by the National Institute of Education and conducted at the Research and Development Center for Teacher Education, The University of Texas, I found the research generally characterized by four problems, which I believe are largely responsible for both the dearth of convincing relationships identified between teacher behavior and pupil achievement[1] and the failure of researchers to build credible quantitative systems by which individual teacher performance can be reliably and validly measured. These four problems are listed below:

1. a narrow range of measurements frequently employed in individual studies of teacher behavior;

2. inconsistent use of specific instruments across studies measuring the same or similar hypotheses;

3. lack of a generic framework or guide from which to select behaviors to be measured in the classroom; and

4. use of instruments with inadequate psychometric characteristics to measure these classroom behaviors.

---

[1]Footnotes appear at the conclusion of this paper.

Problem 1: Range of Measurement

This first problem came to my attention from an examination of literally hundreds of empirical studies investigating relationships between teacher behaviors and pupil outcomes (Borich, 1977; Borich & Madden, 1977; Kash, Borich, & Fenton, 1977). During this review an inordinate number of research studies were noted which measured only a single criterion behavior. While it was apparent that a single criterion provided investigators with a parsimonious research design and a "clean" interpretation of results, the large number of nonsignificant findings produced by studies of this kind suggested that such simplistic methodology represented too narrow and theoretically vacuous an approach to measuring classroom behavior. A number of studies defined teacher behavior, treatments, or instructional programs so broadly that the reader was led to expect wide ranging effects upon pupils. Yet in many of these studies only one treatment or teacher effect was actually measured. Even when multiple criteria were used, they were often applied to only one area of behavior (e.g., classroom interaction variables) or to closely related areas (e.g., self-, pupil, and supervisor evaluations of the teacher). Rarely did researchers employ instruments that captured a range of both pupil and teacher affective and cognitive behaviors. Surprisingly, concurrent measurement of teacher process and pupil product variables was infrequently incorporated into research designs, and few investigators focused on causal sequences of behavior which might have accounted for the effects of classroom instruction. This limited scope might have been avoided had researchers utilized a multivariate approach to the study of classroom behavior. Most of the research encountered in my review dealt with only a single "slice" of the classroom behavior shown in Figure 1. Relatively few studies investigated the sequence of classroom behaviors, taking into account the interactive effects of context, classroom, school, pupil and teacher variables.

Insert Figure 1 about here

Problem 2: Inconsistent Instrument Use

A second problem which emerged from the literature was the inconsistent use of any one instrument across studies purported to be conceptually similar. In a number of areas, there were as many instruments as there were studies, since no two investigations employed instruments which possessed the same reliability or validity and measured variables operationally defined in the same way, regardless of the supposed similarity between the constructs assessed. It was apparent that researchers preferred to develop their own instruments, rather than to use or adapt those already constructed for the same or similar purposes. This emphasis on new instrument development seems to have reduced the opportunity for researchers to improve upon existing measures, to replicate an instrument's reliability and validity, and to use the same operations to measure the same constructs.

Replication of research findings is more likely to occur under some conditions than others. Valid replicated results are most likely to be obtained when the same instrumentation is used across studies which test the same hypotheses. Less congruence between findings is expected when different instruments purporting to measure the same variables are used within the same study; and even less agreement is anticipated when such instruments are used in different studies. Although findings replicated across studies using different instruments are encouraging, the most systematic approach to replication involves use of the same instrument (and, therefore, the same operationally defined constructs) in different studies. Given that "no significant differences" are generally the norm in classroom research,[2] we can place most confidence in those "no difference" findings which are

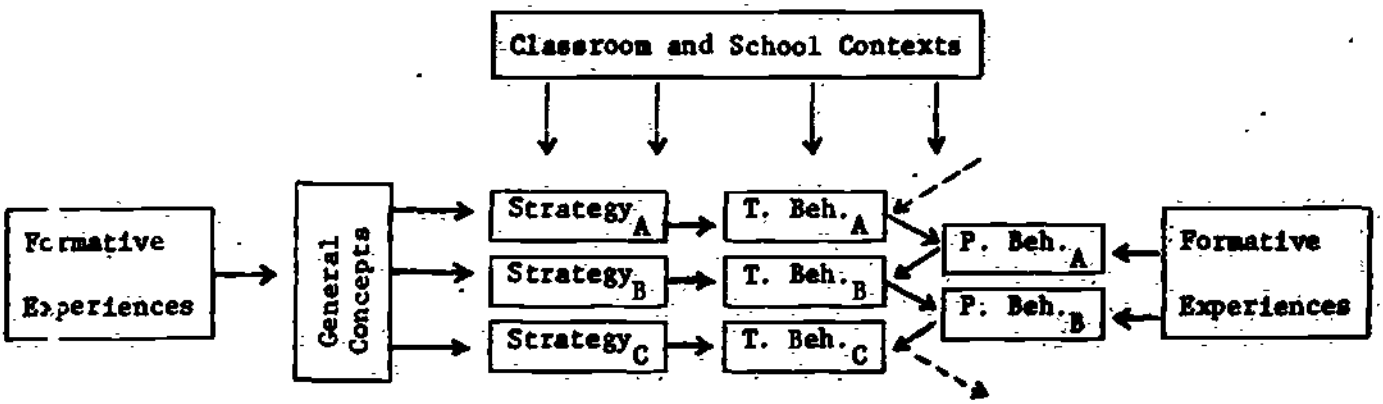

Figure 1. A multivariate model of classroom behavior.

replicated across studies using the same instrument, somewhat less confidence in those obtained with different instruments in the same study, and least confidence in those produced by different studies using different instruments.

Problem 3: Lack of a Measurement Framework

A third problem apparent in the literature was the absence of an overall framework or model to guide researchers in selecting teacher and pupil behaviors to be measured. Few researchers provided rationale for the kinds of teacher behavior they assessed, and even fewer showed interest in (or knowledge of) the causal sequences of behavior possibly prerequisite to the single variable they did measure. Although promising process-product relationships (Rosenshine, 1971) sometimes encouraged researchers to collect both teacher and pupil data within the same study, other variables (e.g., context, ethnographic, presage, and affective) were frequently ignored. Researchers seemed averse to studying those behaviors which were prerequisite to teacher process variables, (e.g., formative experiences of the teacher), or which were likely to confound the measurement of student achievement, (e.g., formative experiences of the pupil). Figure 2 presents an example of the general type of eclectic framework from which investigators must work in order to assure that classroom, school, context, teacher, and pupil variables, as well as the relationships among these variables, are included in the research design. This model will be fully described later in this discussion.

Insert Figure 2 about here

'AGE	Preoperational Measures		Immediate Process Measures		Intermediate Process Measures		Product Measures	
MAIN	Affective	Cognitive	Affective	Cognitive	Affective	Cognitive	Affective	Cognitive
IGH I N F E R E N C E	Personality, attitude, cognitive style	Achievement, experience, and aptitude	Observations of general affective characteristics, e.g., teacher's warmth	Observations of general content-related characteristics, e.g., teacher's business-like or systematic behavior	Ratings of attitudes toward teaching and learning	Ratings of teacher's knowledge of unit or grade level content	Pupil attitudes toward learning, the school, and teachers	Pupil achievement of unit or grade level content
OW I N F E R E N C E	Attitudes related to teaching and learning	Knowledge of teaching methods and content	Observations of specific affective characteristics, e.g., times teacher praises student	Observations of specific content-related characteristics, e.g., teacher's lecture to discussion ratio	Ratings of attitudes toward classroom tasks and lessons	Ratings of knowledge of classroom tasks and lessons	Pupil attitudes toward lesson content	Pupil achievement of lesson content

igure 2. A measurement framework for evaluating classroom instruction.

Problem 4: Psychometric Standards

My review of the literature also revealed the gross psychometric inadequacy of most tests and measures used in these research studies. The literally hundreds of instruments employed in the field of teacher effectiveness research had been subject to little critical assessment, and in the rare cases when such assessment had been conducted and reported in a journal article or test manual, researchers seemed to be unaware of or indifferent to it. In addition, when psychometric evaluations were available, the instruments were generally judged according to absolute, rather than relative standards. Few sources currently exist whereby one can judge the reliability and validity of an instrument in relation to the reliability and validity of other instruments with similar objectives. It was not uncommon, therefore, for researchers to choose an instrument with relatively low validity and reliability without realizing that a more reliable and valid instrument was available and suitable for measuring the same construct. Revision and further development of an existing instrument would in many cases have been more appropriate and productive than construction of still another instrument of equal, or more questionable, reliability and validity. The inadequate psychometric properties characterizing many of the instruments in the studies reviewed, and the apparent availability of other more appropriate measures, suggests that no systematic approach was used in selecting these instruments.

It is important to note that researchers, for whatever reasons, rarely reported adequate psychometric data concerning the structural integrity of their instruments. They provided little information about replicability, constancy, uniformity, and stability of the measures they used, even when an instrument had not previously appeared in the research literature. Replicability refers to the extent to which a pattern or configuration of the behavior being

measured recurs in essentially the same form in random samples on different occasions. Constancy is the degree to which a pattern or configuration appears in essentially the same form at different score ranges (e.g., Do pupils scoring low on an anxiety instrument demonstrate the same configuration of items as pupils scoring high on the instrument?). Uniformity refers to invariance in or similarity of the pattern across selected groups with varying characteristics (e.g., configurational similarity across race, sex, SES, age, etc.). And, stability involves consistency of the pattern across two or more administrations of the instrument to the same subjects. The structural integrity of an instrument is determined by all four of the above characteristics, and the extent to which an instrument has structural integrity in turn determines the reliability of the construct created to explain observed regularities in the behavior of teachers and pupils.

An even greater number of studies failed to provide information about the validity--particularly the convergent and discriminant validity--of the instruments used. Convergent validity is a confirmation of traits (or variables or categories) by independent measuring methods that requires significant correlation between two methods measuring the same trait. Discriminant validity is a requirement that the correlation between different measures measuring the same trait exceed (a) the correlations obtained between that trait and any other trait not having method in common and (b) the correlations between different traits which happen to employ the same method. By determining intercorrelations among constructs in a multitrait-multimethod matrix, one can identify categories which pass specified tests of convergent and discriminant validity. I will return to this concept later, hopefully to persuade the reader that any instrument which can not display both convergent and discriminant validity does not deserve our attention.

Characteristics of a Measurement Framework

In order to develop a framework to guide the selection and construction of instruments to be used in conducting research on classroom behavior, Borich and Madden (1977) reviewed all relevant teacher behavior research and evaluations published during the period from 1954 through 1974. Using Ryans' 1950's study of teacher characteristics (Ryans, 1960) and the publication of the Second Handbook of Research on Teaching (Travers, 1973) as approximate bounding points, they consulted significant journal articles, books, and monographs in order to define parameters for measuring classroom behavior. From this review of the literature, three general dimensions were selected for the purpose of categorizing classroom behavior and the instruments used to measure it. These dimensions were: (1) stage of behavior on a process-product continuum; (2) level of inference required in measuring behavior, and (3) objectives of the instruction.

Process-Product Stages

The inclusion of instruments covering a variety of process-product information stemmed from (1) the author's conviction that researchers should explicitly state in advance the specific behaviors they are attempting to predict or observe; and (2) the fact that research findings produced over the past two decades have repeatedly revealed greater variance between than within process-product stages for both teacher and pupil behaviors.

The process-product continuum suggested for studying classroom behavior contained four stages: (1) a preoperational stage characterized by the collection of antecedent, covariable, and predictor information such as personality, attitude, cognitive style, achievement, experience and aptitude data on pupils and teachers; (2) an immediate process stage during which the ongoing, interactive behaviors of both the teacher and the learner are assessed, usually through systematic observation coding schemes and classroom climate inventories using either a sign, rating or counting metric; (3) an intermediate process stage

involving self-, peer, supervisor, and pupil ratings of the teacher; and (4) a product stage characterized by the measurement of affective and cognitive pupil behaviors across content, subject, and grade level.

Level of Inference

The second dimension used to guide the study of classroom behavior--level of inference--has received increasing attention as a result of evidence which has shown that relationships between teacher behaviors and pupil outcomes differ depending upon whether teacher behavior is measured in a high- or low-inference fashion. High-inference behaviors are general characteristics of the teacher's performance, such as "enthusiasm," "organization" and "clarity," which are measured by overall ratings or observations. Low-inference behaviors, on the other hand, are discrete and easily observable units of activity, such as "teacher asks question," "student gives correct answer" or "teacher reinforces student." These behaviors are measured in terms of their presence or absence and are so clearly defined prior to coding that little inference is required of the observer in deciding whether or not the particular behavior occurred.

Contrary to initial expectation, high-inference measures of teacher behavior have often led to the strongest and most stable findings in the research literature. Such findings and related conclusions have recently stimulated interest in and examination of high-inference measures, while at the same time encouraging revision of low-inference techniques, which offer the advantage of greater definition and specification of the teacher's performance vis-a-vis a specific pupil outcome.

Objectives

The third dimension--objectives of the instruction--seemed particularly important to a measurement framework since outcome can be evaluated only in light of intent. While most previous studies and reviews (Rosenshine, 1971) have dealt almost exclusively with the cognitive achievement of pupils,

affective outcomes are also important and are sufficiently distinct from cognitive outcomes--even for the same teacher variable--to warrant separate treatment. Reviews by Kahn and Weiss (1973) and Loree (1971) remind us that relationships between teacher behaviors and pupil outcomes cannot be expected to be the same for cognitive and affective criteria. Teaching behaviors that have a consistent positive effect in the cognitive domain may have no effect, or even a negative effect in the affective domain (or vice versa). Most researchers and evaluators recognize this, but many fail to adapt their instrumentation to the full range of outcome behavior that could be measured.

The reader will note that the framework which emerges from the juxtaposition of these three dimensions is that which was illustrated in Figure 2. This framework will now serve as the basis for a discussion of sources of invalidity in the measurement of classroom behavior.

Stages of Measurement in the Study of Classroom Behavior

If the measurement of classroom behavior is viewed as a longitudinal process, with data collected at various points in time, four distinct and consecutive measurement stages are apparent: preoperational; immediate; intermediate; and product. These are the four process-product stages previously mentioned. In the following discussion, I have focused on problems which reduce the validity of measurement within each of these stages, including specific assessment and scoring techniques, which are, themselves, sources of invalidity.

The Preoperational Stage of Measurement

During the first stage of measurement, personality, attitude, experience, achievement and aptitude variables are measured to provide a composite picture of the teacher at the beginning of the research or appraisal period.

Though preoperational measurements do not involve the assessment of actual teaching behavior, the information they provide often aids in understanding and interpreting performance data collected at subsequent stages of the measurement process. Moreover, if teacher experience, aptitudes, attitudes, and personality characteristics consistently relate to pupil performance, then teachers may be differentially assigned to teaching positions in accordance with these data. Or, these data may be used in grouping teachers for experimental purposes. Therefore, depending on the nature of the inquiry (i.e., evaluative or experimental), preoperational information can serve as either the foundation for an empirical process of teacher selection and placement or a method of controlling sources of systematic variance that may (covariable) or may not (independent variable) be outside the purview of the study being conducted.

Table 1 lists some of the variables most commonly researched in the preoperational stage. The discussion which follows briefly summarizes some of the more ostensible problems pertaining to the measurements of each type of preoperational variable.

Insert Table 1 about here

Personality variables. Unfortunately, only a few personality constructs have been developed to describe characteristics specifically related to teaching and learning. Consequently, the application of most personality measures to the assessment of teacher performance may be inappropriate. Since personality measures are often designed for and validated in clinical settings, some of the constructs they measure may be irrelevant to the classroom. The more useful "personality" variables may actually represent teachers' concerns about or preferences for specific teaching tasks rather than what are commonly thought of as personality characteristics (See Fuller, 1969; Christensen, 1960;

Table 1. Summary of variables commonly researched in the preoperational stage.

Personality	Attitude	Experience	Aptitude/Achievement
permissiveness	motivation to teach	years of teaching experience	National Teacher exam
dogmatism	attitude toward children	experience in aubject taught	Graduate Record Exam
authoritarianism	attitude toward teaching	experience in grade level taught	Scholastic Aptitude Test
achievement motivation	attitude toward authority	workshops attended	1. verbal
introversion-extroversion	vocational interest	graduate courses taken	2. quantitative
abstractness-concreteness	attitude toward self	degrees held	special ability tests, e.g.,
directness-indirectness	(self concept)	professional papers written	reasoning ability, logical
locus of control	attitude toward subject		ability, verbal fluency
anxiety -- 1. general	taught		Grade Point Average
2. teaching			1. overall
			2. in major aubject
			profesaional recommendations
			atudent evaluationa of teachin
			effectivenesa
			atudent teaching evaluations

Doyal & Forsyth, 1973; DeBlassie, 1971; Duffey & Martin, 1973; Marjoribanks, 1970; Soar, Soar, & Ragosta, 1973; Weiss, Sales, & Bode, 1970; Yonge & Sassenrath, 1968.)

Attitude variables. Attitude assessments may be global (e.g., attitude toward the school and the educational system), or specific (e.g., attitude toward a particular task, text or curriculum). In either case, attitude instruments often suffer from inadequate predictive validity. Relationships between teacher attitude and performance in the classroom are commonly low and nonsignificant. Therefore, in the absence of clear-cut validity data, attitude measurement in the preoperational stage usually rests on the assumption that the attitudes assessed are intervening or enabling constructs, i.e., are prerequisite to certain affective and cognitive behaviors. Thus, as causative agents, responsible for engendering pupil change, these constructs are more remote and less credible than performance variables, which offer more immediate links to pupil achievement. Motivation to teach and attitude toward children are among the most important attitude variables measured in the preoperational stage. (See Horn & Morrison, 1965; Krasno, 1972; Neale, Gill & Tismer, 1970; McCallon, 1966.)

Experience variables. Although two decades of research have shown experience variables to be almost worthless in predicting teaching performance, it is possible that these variables have in the past been measured too grossly to yield significant findings. The standard biographical data form, on which years of teaching and extent and type of training are recorded, defines the teacher's experience so broadly that it cannot be used to identify teachers who will be more or less effective in relation to specific performance criteria. For example, a teacher's experience with the type of curriculum or the kind of pupils he will be expected to teach may be far more relevant to his performance than the number of years he has taught or the graduate credits he

has earned. Yet, the latter rather than the former typically appear on the standard biographical information form, often leaving specific data related to the teaching context untapped. (See Dumas, 1969; McCallon, 1966; Rutherford & Weaver, 1974.)

Achievement and aptitude variables. Like experience variables, most achievement and aptitude data have been of little value in predicting teacher performance. The prior achievement of the teacher, in terms of, say, college grades, has rarely shown any direct relationship to teaching performance. This may be accounted for by the relatively low variability which characterizes the prior achievement (e.g., course grades, GPA) of teachers. Standards set by training institutions generally insure that all teachers meet a minimum level of knowledge-related achievement, which is usually high enough to skew the distribution of this variable. Considerably greater success has been found in relating specific cognitive styles (e.g., verbal fluency, reasoning and logical ability) to teaching effectiveness. (See Alschuler, 1969; Dacey & Madaus, 1971; Knoell, 1953; Treffinger, Feldhusen, & Thomas, 1970; McDonald et al., 1975.)

The Immediate Process Stage of Measurement

The second phase of the measurement process is the immediate, or observation, stage. In this stage, the teacher's actual classroom behavior is recorded. He is observed as he applies procedures, strategies, and techniques in the course of teaching, and these observations are recorded on presumably reliable instruments, containing explicitly stated behavioral categories. These categories focus the observer's attention on either low-inference (i.e., discrete and specific) or high-inference (i.e., general and cumulative) behaviors. There are three characteristics which distinguish various observation instruments: (a) the recording procedure; (b) the item content; and (c) the coding format. Each of these is discussed briefly below.

Tools for observing ongoing classroom events may employ either of two recording procedures--sign or category. A sign system records an event only once regardless of how often it occurs within a specified time period. The behavior is given a code which indicates only its presence or absence within a particular block of time. A category syatem, on the other hand, records a given teacher behavior each time it appears and hence providea a frequency count for the occurrence of specific behaviors, rather than a mere indication of their presence or absence. A frequency count may also be obtained using a modified sign system, called a rating inatrument, which eatimates the degree to which a particular behavior occurs. For example, instead of simply noting the presence or absence of a behavior, a rating instrument may auggeat the frequency at which the behavior occurs on, say, a 1-5 scale, with "5" indicating a high frequency of occurrence and "1" a low frequency of occurrence.

Observation systems can be further differentiated on the baaia of item content. Generally, observation instruments, whether they be of the category, sign, or rating variety, focus on either high- or low-inference behaviors. Those which ask an observer to judge, for example, the presence, absence or degree of a teacher's warmth, effectiveness, clarity, or enthusiasm require high inference because the item content does not specify discrete behaviors which must occur in order for a teacher to be conaidered warm, effective, clear, or enthusiastic. Item content which is cumulative in nature, like that on many rating scales, forces the observer to make high-inference judgments about the behavior being observed. Obaervation instruments which name specific behaviors to be recorded, such as "teacher asks question" or "teacher uses example," require little inference on the observer's part. It should be noted that not all observation systems are either high- or low-inference. Some combine the two types of item content, while othera may require an intermediate level of inference from the observer.

Periods						Teacher Practices
I	II	III	IV	V	VI	
✓			✓			1. T occupies center of attention.
		✓				2. T makes p center of attention.
						3. T makes some thing as a thing center of p's attention.
						4. T makes doing something center of p's attention.
	✓					5. T has p spend time waiting, watching, listening.
						6. T has p participate actively.
					✓	7. T remains aloof or detached from p's activities.
						8. T joins or participates in p's activities.
						9. T discourages or prevents p from expressing self freely.
				✓		10. T encourages p to express self freely.

Figure 3. Sign system. (From the Teacher Practices Observation Record, in The Experimental Mind in Education, by B. B. Brown, New York: Harper & Row, 1968.

			Teacher							Pupil			
	Category		1	2	3	4	5	6	7	8	9	10	Total
Teacher	accepts feelings	1											0
	praises	2											0
	accepts ideas	3			1		1						2
	asks questions	4				2	1			12		1	16
	lectures	5				5	22	3					30
	gives directions	6					1	5		3		4	13
	criticizes	7											0
Pupil	responds	8			1	7	4	4		14		1	31
	initiates	9											0
	silence	10				2	1	1		2		3	9
		Total	0	0	2	16	30	13	0	31	0	9	101

Figure 4. Category system for recording sequential pairs of events. (From Flanders' Interaction Analysis System, in Teacher Influence, Pupil Attitudes, and Achievement by N. A. Flanders, Final Report of Cooperative Research Project, No. 397, U. S. Office of Education, University of Minnesota, 1960.)

1. Amount of Criticism: High-Low

High		Moderate		Low
1	2	3	4	5

2. Criticism: Personal-General

Personal		Mixed		General
1	2	3	4	5

3. Criticism: Kind-Harsh

Kind		Neutral		Harsh
1	2	3	4	5

4. Warmth: Warm-Cold

Warm		Neutral		Cold
1	2	3	4	5

5. Enthusiasm: Enthusiastic-Apathetic

Enthusiastic		Neutral		Apathetic
1	2	3	4	5

.
.
.

6.

Figure 5. Rating system.

Categories	Makes Statement		Asks Question	
	On task	Off task	On task	Off task
Teacher	3	1	5	2
Pupil	4	6	5	2

Figure 6. Multiple coding system.

A third distinction among observation instruments concerns differences in coding format. Two coding formats are available: single and multiple. A single coding format records a behavior on one dimension. Multiple coding, on the other hand, divides a general behavior into two or more discrete subcategories which further define it. Each subcode deals with a different aspect of the initial behavior observed. For example, a single comment might be coded in three ways, according to (a) the identity of the speaker (i.e., teacher or pupil), (b) whether the speaker is on or off task, and (c) whether the speaker is making a statement or asking a question. Other multiple coding formats might include observation and recording of the teacher's pedagogical behaviors and the pupils' responses as they occur sequentially. These sequential records show patterns of classroom interaction, which on a single coding format would appear as a number of separate, unrelated behaviors. Figures 3-6 illustrate differences in recording procedures, item content, and coding format among observation systems.

Insert Figures 3, 4, 5, and 6 about here

Observation coding systems provide a method for recording the teaching behaviors (i.e., strategies, procedures, and techniques) that are ostensibly used to produce pupil growth. The observation of teaching leads to the identification of two general types of behaviors: those which are considered desirable as an end in themselves and those which are considered desirable <u>because</u> they promote pupil growth. Those considered desirable in and of themselves are generally high-inference behaviors, and their inclusion in an observation system is easily justified since they reflect inherently "good" practices, such as "teacher shows warmth toward children," or "teacher uses student ideas." Because these behaviors are clearly desirable, they need not

always relate to pupil achievement to be employed in an observation instrument. The case for including low-inference item content in an observation instrument, however, is less obvious. Since it is not immediately apparent that low-inference items such as "teacher uses blackboard" or "teacher probes pupil for correct response" represent desirable behaviors, these items must be empirically linked to pupil performance.

The justification of item content, however, is only one of several methodological problems involved in the use of observation coding systems. Others concern the reliability and validity of these "systems."

It is important to note four distinct threats to the accuracy of any observation system. These are: (1) consistency of observations among those judging the behavior; (2) stability of the behavior measured across pupils, content, and time; (3) convergence of the behavior being observed with similar measures of teaching behavior; and (4) divergence of the behavior being observed from dissimilar measures of teacher behavior. Since a reliable index of teacher behavior is not necessarily a valid index, but a valid index must always be reliable, I will discuss the contribution of the concept of reliability to the observation of classroom behavior before turning to the more encompassing topic of validity.

In this context reliability refers to the consistency or agreement between two independently derived observations, recorded on the same coding instrument. It can be measured in several ways. For example, the reliability of a coding system can be determined by correlating observations recorded by different raters using the same instrument and observing a teacher for the same period of time. This procedure yields an estimate of interrater reliability, which is an index of consistency among raters. The interrater reliability of most observation systems is adequate, or can be made adequate, given sufficient resources and time in which to train observers in using the instrument. Of

greater concern, however, is test-retest reliability, which is a measure of the stability of teacher behavior as recorded by a given observation instrument across changes in time, content, or pupils. This type of reliability is determined by correlating the results of two observations of the same teacher, recorded at different times by the same observer. Reliability across time refers to the stability of teacher behavior or the capacity of an observation instrument to record the stable components of teacher behavior at different times, whether these times are separated by a week, a month, or a year. Similarly, reliability across content concerns the stability of teacher behavior or the capacity of an observation instrument to record this stability, regardless of the subject matter being taught to a particular group of pupils. And, reliability across pupils refers to the stability of teacher behavior from one class of pupils to another, with content held constant. Teacher behaviorists have been relatively unsuccessful in establishing the stability of teacher effects on pupils over long periods of time and across different content, though they have achieved some consistency over brief instructional units and across different pupils (Rosenshine, 1970; Shavelson & Dempsey, 1975). The results of these studies suggest that teacher behavior may not be stable across long periods of time and content, or that our assessment systems fail to record the kind of teacher behavior which remains constant across these dimensions.[3]

This instability may be explained in two ways. The most pessimistic stance assumes that teacher behavior of almost any type is basically unstable. That is, teachers do not perform consistently from day to day or from class to class. While this pessimistic explanation may eventually prove correct, at the moment it lacks convincing support for reasons I shall describe below.

An alternative explanation, which appears somewhat more tenable on the face of research evidence, is that our measures of teacher behavior are inadequate and, therefore, do not allow us to record the consistency which

may, in fact, characterize teacher behavior. This explanation contains two corollary assumptions: (1) at least some of our instruments for measuring teacher behavior are not tapping those specific behaviors which are relatively stable across subject matter and time; and (2) the constructs currently used as indices of teacher process are measured so poorly by existing instruments that stable teacher behaviors are almost impossible to record. These two assumptions are related to the concepts of validity and reliability, respectively

Validity may be defined as the extent to which an instrument measures the teacher or pupil behaviors it purports to measure. While the validity of an index of teacher behavior can only be improved through a reconceptualization of the construct being measured (a considerable investment in time and effort), reliability can be improved either by increasing the number of occasions on which the behavior is rated or observed or by increasing the number of individuals doing the rating or observation--or both.[4] The reliability estimates obtained for a particular behavior, of course, may not apply when the instrument is used in other contexts or when different content and different pupils are involved.

A lack of validity, as noted above, is more complex than a lack of reliability. The former leaves little alternative but to reconceptualize the operational definition of the behavior of interest and thus to create a new instrument to measure it.

Let us first review the well documented but often overlooked relationship between reliability and validity. I present the reader with the following exercise fully realizing that if teacher behaviorists seriously considered this relationship in selecting and constructing process stage instruments, a good portion of these instruments would be deemed unsound.

Reliability can be defined as:

$$r_{tt} = 1 - \frac{s_e^2}{s_t^2},$$

or, the proportion of error (s_e^2) variance to total test variance (s_t^2), subtracted from unity. Analogously, validity can be defined as:

$$val = \frac{s_{co}^2}{s_t^2}$$

or, the proportion of common factor variance (s_{co}^2) to total test variance (s_t^2). The total variance of any test can be divided into three compenents: common factor variance, specific variance, and error variance, as shown in the equation below:

$$s_t^2 = s_{co}^2 + s_{sp}^2 + s_e^2$$

In order to speak in terms of proportions of total variance, we can divide each member of the equation by the total variance:

$$\frac{s_t^2}{s_t^2} = \frac{s_{co}^2}{s_t^2} + \frac{s_{sp}^2}{s_t^2} + \frac{s_e^2}{s_t^2}$$

And, to move our definition of validity ($\frac{s_{co}^2}{s_t^2}$) to the left-hand side of the equation, we can transpose terms:

$$\frac{s_{co}^2}{s_t^2} = \frac{s_t^2}{s_t^2} - \frac{s_{sp}^2}{s_t^2} - \frac{s_e^2}{s_t^2}$$

so that now validity can be defined as that part of the total variance of a measure that is not specific variance and not error variance. Note the portion of the formula for validity that is the same as the formula for reliability:

$$\frac{s^2_{co}}{s^2_t} = \frac{s^2_t}{s^2_t} - \frac{s^2_e}{s^2_t} - \frac{s^2_{sp}}{s^2_t} .$$

Reliability is equal to the two right-hand terms of the formula and, thus, we arrive at the basis for the well-known principle that <u>any validity coefficient for a measure must always be equal to or less than its reliability</u>.

Suppose, for example, that the proportion of a test's error variance to total variance was .35, i.e., that the test was only moderately reliable.

If $\frac{s^2_e}{s^2_t} = .35$, its reliability will equal 1 - .35 or .65

then val = 1 - .35 or .65 - $\frac{s^2_{sp}}{s^2_t}$

or val $\leq$ .65.

Thus, validity is that proportion of the total variance which is left over after the test's error and specific variance have been subtracted from the total variance.

What practical implications do these formuli have for the validity of instruments which purport to measure classroom behavior? They imply, simply, that an instrument's validity will be less than its reliability.[5] In most cases for which adequate data exist, validity coefficients have been found to be as much as 25 to 50% less than reliability estimates (Borich & Madden, 1977). For example, subscale reliabilities for one-third of the instruments studied by Borich and Madden (1977) were in the moderate range (.50 - .70), while validity coefficients for a random selection of these instruments fell between .25 and 52.5 It should be noted that these instruments are popular assessment

tools, familiar to many researchers and commonly used in large-scale research projects.

The use of instruments with moderate to poor validity may account, in part, for some of the null findings which have occurred in teacher effectiveness studies. The following reasoning can be brought to bear on the problem: (a) if the validity of an instrument is low, the instrument fails to measure the construct intended; (b) if the instrument fails to measure the construct for which a research hypothesis is posed, the power to detect a significant finding related to that hypothesis must necessarily be weak; (c) null findings can then be attributed to constructs other than that which was defined at the beginning of the study. The effect is not unlike entering into a research study knowing that the chance of missing a significant finding (if one, in fact, exists) is equal to or greater than, say, .5. Who among us would gamble his precious resources so foolishly? It may not be coincidental that teacher effectiveness studies, no matter how well executed, commonly find "no significant differences." This is why, at least for the moment, I prefer to reject the pessimistic explanation that teacher behavior <u>is</u> basically unstable and to focus upon the means by which the validity of our instruments can be improved.

The premises underlying convergent and discriminant validity are: (1) the correlation between the same behavior measured by the same method (reliability) should be higher than (2) the correlation between the same behavior measured by two different methods--which, in turn, should be higher than (3) the correlation between two different behaviors measured by the same method--which, in turn, should be higher than (4) the correlation between two different behaviors measured by two different methods. A simple method-by-behavior design for determining the convergent and discriminant validity of two separate teacher behaviors, each measured by different instruments, is as follows.

		Methods A		Methods B	
		T. behaviors		T. behaviors	
		Accepts	Questions	Values	Delves
		1	2	1	2
A	1	(.86)			
	2	.23	(.70)		
B	1	.63	.22	(.58)	
	2	-.12	-.01	.27	(.84)

For illustrative purposes, let us assume that (1) A and B are two different classroom observation systems purporting to measure the same teacher behaviors and that (2) the operational definition of teacher accepts, on instrument A, is similar to that of teacher values on instrument B. Likewise, the behaviors questions and delves are similarly defined across the two instruments. By referring to the premises which underly convergent and discriminant validity, we can determine that relatively good convergent and discriminant validity is indicated for the behavior accepts, but poor convergent and discriminant validity is indicated for the behavior questions. Whether the behavior questions or the behavior delves is invalid or whether, in fact, both fail to measure the construct they purport to measure cannot be known. However, given the evidence above, it would be foolhardy to use either instrument for measuring the desired behavior.

While the above paradigm is rarely employed by teacher behaviorists, it provides an example of the type of reconceptualization which should be undertaken when the instability of teacher behavior is due to the invalidity of the instrument, rather than the unreliability of the measure. If lack of validity stems solely from a failure to consistently measure the behavior, we need only find the optimal number of occasions and observers needed to increase reliability to an acceptable level and thereby increase our validity.

If, however, reliability is not at issue, then we must redefine and remeasure the behavior.

The Intermediate Process Stage of Measurement

The next stage of measurement is the intermediate stage, in which the teacher's cumulative behavior is rated on predetermined scales. These ratings differ in two ways from the coding of classroom behavior which occurs during the previous stage. First, intermediate measurements are made after, not in conjunction with, classroom observation. Second, these ratings are cumulative in nature, summarizing the frequency and quality of many behaviors in a single judgment. At the intermediate stage, for example, the evaluator may rate a teacher's attitude toward teaching, his knowledge of unit or grade-level content, his attitude toward particular tasks and lessons, or his use of classroom management techniques. Such ratings are used primarily to fill the gap between observations of specific classroom events and various indices of pupil growth recorded on norm-referenced or criterion-referenced tests. Intermediate measures are thus, on the one hand, an attempt to summarize the numerous, discrete events in the classroom and, on the other hand, an attempt to provide a global description of the teacher behaviors responsible for pupil growth. These summative ratings can be recorded on a variety of scales, using a number of techniques. While all of the methods available for rating teacher performance are too numerous to mention here, several of the more popular varieties and the measurement problems they pose are noted below.

Summated ratings (Likert scales). The Likert scaling technique requires a large number of items which describe teacher behaviors, each yielding a high score for a favorable rating on a behavior and a lower score for a less favorable rating. The rater reacts to items on a 5-point response continuum, which reflects either the quality of a behavior or the frequency at which it was perceived to occur. The Likert procedure customarily yields scales with

moderate to high reliability. Validity, however, can vary, depending upon the following considerations. No attempt is made in the construction of a Likert scale to insure equal distances between units (e.g., between "very often" and "fairly often" or between "always relevant" and "mostly relevant"). Therefore, increments of change may have different meaning on different portions of the scale. This may encourage raters to make judgments more frequently at one end of the scale than the other. For example, raters often view judgments recorded on the bottom half of a scale as so detrimental to the teacher that they are reluctant to use that end regardless of their "true" observations. Furthermore, the unidimensionality of the scale, i.e., the extent to which it measures a single, distinct behavior, must be inferred from high correlations between item and total scores. The lack of such correlations makes the construct multifaceted and factorially complex, precluding any simple and direct interpretation of the behavior. Likert scores are interpreted according to a distribution of sample scores, and an individual teacher's score has meaning only in relation to the scores of other teachers. This may complicate interpretation since, ultimately, teachers should be judged according to their achievement of specific, well defined competencies, and not on the basis of their standing relative to others who also may have failed to achieve the desired behavior.

Semantic Differential scales. The Semantic Differential is another method used to cumulatively record the quality or frequency of teacher behaviors. It requires the rater to judge the teacher's performance on a series of 7-point bipolar scales. The rater checks the appropriate space, indicating both the direction and intensity of his judgment. Since the Semantic Differential and Likert techniques are similar, the cautions noted above also apply here: the Semantic Differential does not necessarily exhibit equal intervals between scale points; the unidimensionality of the concept being measured may vary

from one scale to another (particularly when bipolar responses are not exact opposites); and scores are interpreted relative to the rated performance of others. In practice, differences between Likert and Semantic Differential scales are minor and are generally related to the use of 5- or 7-point response formats. The similarity of these procedures is often reflected by high or moderate correlations between the two when they are used to measure the same behavior.

Scalogram analysis (the Guttman Scale). Another method of recording summative judgments of teacher performance is the Guttman Scale. This method is based upon the idea that behaviors can be arranged hierarchically so that a teacher who possesses a particular behavior may be assumed to poaseas all other behaviors having a lower rank. When such an arrangement is found to be valid, the behaviors are said to be scalable. In developing a Guttman Scale, items are formulated and arranged in a hierarchical order. These items are then administered to a group of teachers, whose response patterns are analyzed to determine whether or not the items are scalable. If items require only agreement or disagreement, i.e., an indication of the presence or absence of a behavior, there are 2^n response patterns that might occur. If items are scalable, however, only $n + 1$ of these patterns can be obtained. The relative nonoccurrence of deviant patterns allows the computation of what is called a coefficient of reproducibility (R). R is equal to the proportion of responses that can be correctly reproduced from the knowledge of a teacher's score. The extent to which such inferences can be made depends upon the level of the coefficient of reproducibility. This value represents a measure of the unidimensionality of the scale and is an index of the scale's validity. Like the Likert and Semantic Differential scales, the Guttman Scale makes no attempt to insure equal units between items. However, unlike the Likert and Semantic Differential, the Guttman Scale need

not be interpreted relative to the ratings of other teachers, since its items represent specific behaviors, the presence or absence of which can form the basis of an absolute as well as a relative judgment. This should be a desired characteristic of any instrument used to evaluate the performance of individual teachers.

Checklists. When a behavior cannot be easily rated on a continuum of values, a simple indication of its presence or absence is used. If the researcher is unable to make fine gradations in judging the quality or frequency of behavior, a simple yes-no, observed-unobserved, or present-absent format is used. Since checklists record only the presence or absence of behaviors, they assume that the rater has had ample opportunity to observe these behaviors. However, many times this assumption is unwarranted. When checklist data indicate the absence of a particular behavior, it should be determined whether this reflects a true absence or simply a lack of opportunity to observe the behavior. The latter situation may occur when the teacher's objectives are unrelated to or incompatible with the particular behavior in question or when the rater has visited the classroom too infrequently to have had an opportunity to observe the behavior. In order for the rater to distinguish the absence of an event from inopportunity to observe it, checklists should provide three response alternatives: (a) no opportunity to observe the event; (b) presence of the event; and (c) absence of the event. The rater would choose the first alternative whenever a behavior on the checklist was both unobserved and unlikely to have been observed, considering classroom conditions which existed at the time. The "true" absence of a behavior would then be recorded using the third alternative.

The Product Stage of Measurement

Although each stage of the measurement process involves problems, a system that assesses behavior at four stages--preoperational, immediate, intermediate, and product--provides a composite picture of teacher performance in which the errors of measurement may be counter-balanced and limited. The product stage, considered by some researchers the most important stage of measurement, is, therefore, best viewed as a component within a larger network of teacher and pupil behavior. Product-stage assessments confirm observations and ratings made at earlier stages while at the same time provide their own unique contribution to the measurement process.

The product stage of measurement involves the recording of changes in pupil achievement, both affective and cognitive, over a prespecified period of instruction. This period may be as brief as the span of a single lesson or as long as a semester or a school year. The teacher's pupils are assessed at Time 1, the beginning of a unit of instruction, and at Time 2, the end of the unit. The difference between pre- and posttest pupil achievement is attributed to the performance of the teacher. Pupil tests which are employed to measure teacher proficiency in this manner may be either standardized (i.e., norm-referenced) or criterion-referenced.

The major problems in the product stage of measurement are:

(1) Determining and controlling the extent to which pupil performance is affected by influences other than the teacher. Some studies have indicated that parental expectations, the pupil's prior achievement, the socioeconomic status of the family, and the general intellectual quality of the pupil's home life may have greater influence on the pupil's measured achievement than does the teacher. If this is true, to what extent can we infer teacher effectiveness from pupil performance?

(2) The unreliability of the "raw gain score," which is the difference between the pupils' pre- and posttest achievement. This score is unreliable for two reasons. First, in calculating the raw gain score, the unreliability inherent in the pretest is added to that in the posttest, making the resulting raw gain or difference score less reliable than either the pre- or posttest score alone. Second, research has indicated that teacher effects upon pupil achievement may not be consistent over long periods of time and across subject-matter. Thus, if a teacher's influence on pupil performance is inconsistent from one subject, or one time, to another, one can legitimately question the use of pupil gain (of any kind) as a measure of teacher effectiveness.

(3) The teacher's understandable desire to teach to the test when he knows that pupil achievement is to be an index of his effectiveness. Teachers may consciously or subconsciously plan classroom instruction which focuses upon content which they suspect will be measured by specific test items. For example, teachers may guess that pupil achievement tests will contain material which is easily measured, rather than higher-order learning which requires more complex pupil performance and testing procedures. Hence, they may proceed to teach the more straightforward, easily measured content. This is unfortunate since higher-order learning, reflecting more complex instructional objectives, may be more important than other criteria in distinguishing more effective from less effective teachers. Pupil growth in these areas, however, may be imperceptible during any given period of instruction.

Some Guidelines for Improving the Measurement Process

While many guidelines can be offered for improving the measurement of classroom behavior, six which address particularly distressing problems are presented here. These guidelines apply primarily to the measurement of pupil change, an area plagued by the most critical problems.

Guideline 1: Idiographic Rather Than Nomothetic Tests of Pupil Performance Should Be Used

An idiographic test produces a score which describes the individual's performance in relation to the test, while a nomothetic measure yields a score which describes the subject's performance in relation to that of other examinees who serve as a norm group. Idiographic tests are commonly referred to as "criterion-referenced" measures since they relate test performance to a predetermined standard or criterion rather than to the performance of others. The term "norm-referenced" applies to tests which compare an individual's performance to that of others who have taken the same test.

The suitability of criterion-referenced and norm-referenced measures as indices of teacher effectiveness becomes apparent when the objectives of each are compared. Idiographic, or criterion-referenced, tests attempt to determine whether or not the examinee has attained a particular skill, or mastered a given content area. The items on such tests deal with situations, problems, or tasks, mastery of which is essential to proficiency in the skill being measured. If the pupil can correctly answer a sufficient number of these items, he has achieved proficiency in the particular skill--regardless of how his classmates have performed on the test.

The purpose of norm-referenced tests, however, is to discriminate among pupils, to reveal differences in performance, rather than mastery of a skill or subject area. This objective demands the inclusion of a variety of items, some of which must be relatively obscure or difficult, in order to differentiate

among pupils. Accordingly, norm referenced tests must contain items which cover not only the main ideas or skills taught, but also the finer points, knowledge of which may not be essential to proficiency in the subject area or skill under consideration.

A good criterion-referenced test should produce less variability in pupil performance within each administration than between pre- and post-administrations. An ideal criterion-referenced measure would actually register zero variability among pupils on both the pre- and posttests but a maximum amount of variability between the two administrations. In other words, all the pupils would answer all the items incorrectly on the pretest and correctly on the poattest. The truly effective teacher should be able to reduce variability in achievement among pupils by obtaining approximately 100% mastery of the specified objectives for each pupil taught. A criterion-referenced test can show how well the teacher has achieved this goal by indicating the number of pupils who have mastered the material taught. A norm-referenced test, on the other hand, by presenting content more global and extensive than that which can be taught during a brief measurement period, intentionally prevents all pupils from obtaining 100% mastery and thereby increases rather than reduces variability in pupil performance.

It should be obvious that criterion-referenced tests measure the outcomes of specific teaching processes better than norm-referenced tests. The latter, in fact, may measure behaviors beyond what is taught by or even of interest to a particular teacher.

Guideline 2: Both Process and Product Behaviors Should Be Measured

Process behaviors refer to teacher performance while product behaviors involve pupil change. Classroom assessment should include both process and product measures as indices of teacher effectiveness. While questions raised

about the stability of teacher effects on pupil achievement have caused some researchers to advocate use of one or the other, both process and product measures are essential to valid measurement, since we cannot assume that stable teacher behavior always produces stable pupil outcomes or that stable pupil outcomes are always attributable to stable teacher behaviors.

The strongest evidence supporting the use of product measures comes from Rosenshine (1970), Brophy (1973), and Shavelson and Dempsey (1975). Rosenshine, examining nine studies of classroom behavior (both long- and short-term), found that teacher effects upon pupil achievement were moderately consistent when instructors were teaching the same material to different students--a circumstance which approximates the real, day-to-day teaching environment. However, when instructors were teaching different material to the same students, or different material to different students, pupil outcome was less consistent, suggesting that these latter situations are the least desirable contexts in which to measure teacher effects.

These findings are confirmed by Shavelson and Dempsey (1975) who, examining all available long- and short-term studies of teacher stability published since the Rosenshine review, also conclude that teachers teaching the same material to different students are moderately consistent.

The strongest support for this view, however, is provided by Brophy (1973), who reports stability data from the Texas Teacher Effectiveness Study. After collecting pupils' scores over a 4-year period on a given subtest of the Metropolitan Achievement Test, Brophy judged the mean scores of each teacher's pupils for a linear pattern. Overall, 28% of these judgments indicated linear constancy, 13% linear improvement, 11% linear decline, and 49% non-linearity; thus, about half of the assigned judgments indicated some form of consistency in the ways various teachers engender pupil achievement. When considered in conjunction with the findings of Rosenshine and Shavelson

and Dempsey, these results suggest a fairly high degree of consistency among teachers in generating pupil gains, at least in teaching the same content to different groups of similar students--circumstances which, as pointed out, resemble the natural teaching situation.

Why, then, in the face of such evidence, do some researchers discourage the use of product indices, in favor of process measures of teaching effectiveness? Glass (1974), the major proponent of process measures, bases his opposition to the use of pupil outcome on statistical grounds. Examining 21 short-term studies, including those reviewed by Rosenshine, Glass points out that confidence intervals for only 4 out of 21 stability coefficients failed to span zero. Glass effectively demonstrates that even when teachers are shown to be relatively consistent over content and pupils, the stability coefficients themselves may not be accurate estimates of the consistency of teacher performance--though product measures of teaching effectiveness often assume such consistency.

These findings lead Glass to argue against the use of either standardized achievement tests or performance tests of teaching effectiveness (which compare different teachers by requiring them to teach a specified topic to a randomly formed group of pupils for one class period, after which the pupils are tested for mastery of the material taught). He suggests, instead, that process evaluations of the teacher, made by trained observers or students, are the most stable indices of teacher effectiveness. Such evaluations, he proposes, should perhaps focus on the 11 teacher variables identified by Rosenshine (1971) as "promising." Glass qualifies his endorsement of these variables, however, by specifying that no characteristic of teaching should be incorporated into rating scales until research has established that it can be reliably observed and that it significantly relates to desired pupil outcomes.

Shavelson and Dempsey, in a more recent, unpublished paper (1977) give qualified support to Glass' promotion of teacher process variables. Reviewing teacher process behaviors, identified as important in their own right or linked in previous studies to desirable pupil outcomes, Shavelson and Dempsey conclude that the stability of teacher behavior depends on teaching conditions, or "facets" (i.e., grade level, subject matter, and type of students). While the variables "teacher presentation," "positive and neutral feedback," "probing," and "direct teacher control" were relatively stable across facets, the consistency of six additional variables was unclear. In other words, some process variables are stable and some unstable.

It appears, then that the stability of both product and process behaviors is open to question. While the findings of Rosenshine, Brophy, and Shavelson and Dempsey (1975) suggest that pupil gain is moderately stable under certain conditions, these findings must be interpreted cautiously in light of the criticisms of Glass. In turn, Glass' endorsement of process behaviors as the most stable index of teacher effectiveness must be qualified by the more recent (1977) findings of Shavelson and Dempsey.

Most of the studies reviewed by these authors have assumed that pupil achievement is caused by teacher behavior; that is, if pupil gain was inconsistent, it was assumed that teacher behavior was unstable. Of course, it is possible for pupil performance to be unstable, regardless of teacher behavior. Or it can be stable in spite of teacher behavior. It should be clear, then, that we must first determine the stability of teacher process variables in order to make inferences about teacher behavior from pupil achievement. Though we would like to believe that stable teacher behavior leads to stable pupil achievement and that unstable teacher behavior leads to unstable pupil achievement, research indicates that we cannot make such assumptions. We must instead study teacher behavior separately and then

relate our findings to pupil achievement, thereby including both process and product measures in our assessment.

Guideline 3: Performance Measured Should Match Objectives Planned

Teacher and pupil performance measures that are unrelated or only tangentially related to the teacher's goals and objectives are virtually worthless for both experimental and evaluative purposes. Yet, it is not uncommon to find behaviors and skills selected for investigation that are incongruent with the objectives of the teacher.

To maximize the probability of congruence between the teacher's objectives and the performance measured, a table of behavioral specifications can be developed. Such a table is developed according to the following process: first, the educational goals of teachers, administrators, and the community are recorded; second, the teaching behaviors implied by these goals are determined and appropriate measures of teaching process selected; and, third, affective and cognitive outcomes of pupils are logically derived from the identified teaching behaviors. The educational goals of teachers and others are used to select the teaching behaviors to be measured and, in turn, these teaching behaviors are used as a basis for extracting desired pupil outcomes. A comparison of pupil outcomes and instructional goals serves as a logical check on the appropriateness of the assessment tools and related outcome behaviors used in the teacher research or appraisal study. This process is shown in Figure 7.

Insert Figure 7 about here

Guideline 4: Objectives Planned Should Match Objectives Taught

Even when instructional objectives are congruent with instrumentation, factors outside the teacher's control can disrupt planned instruction. Teacher behaviorists must take into account such factors as extreme ability

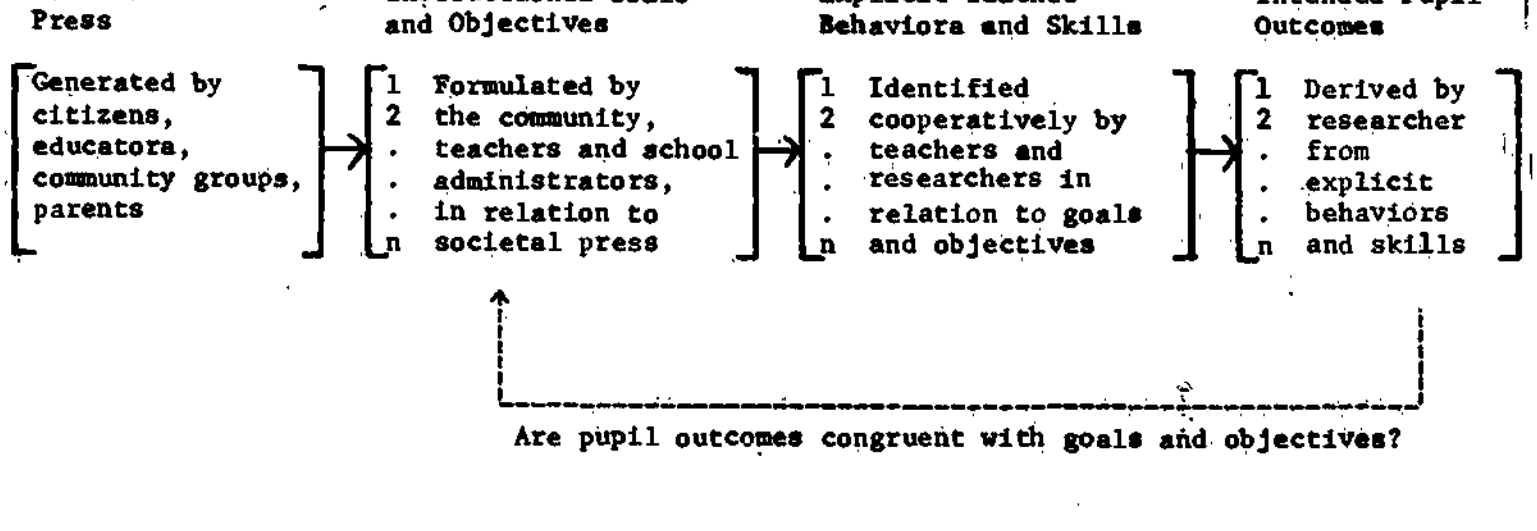

Figure 7. Matching teaching behaviors and skills with pupil outcomes.

differences among the pupils in a class, the teacher's familiarity with new materials, unexpected curriculum changes requested by department heads, and interruptions for nonacademic activities (sporting events, pep rallies, and classroom visitors), all of which may alter planned instruction and thereby deprive the teacher of opportunities to perform the behaviors that are to be observed.

To determine whether planned content was actually taught, the investigator must develop a systematic method of monitoring the congruence between instructional activities and effectiveness criteria throughout the assessment period. During the immediate stage of measurement, in particular, data can be collected to determine if practical constraints have prevented adequate instruction in relation to predetermined objectives. This may be accomplished by: (1) giving a list of objectives organized by various content areas to teachers, who then check those which they feel have been taught, and (2) giving a list of content areas arranged according to concepts and principles to pupils, who then check those areas for which they perceive instruction to have been provided. If both teacher and pupil checklists concur that particular content was not covered, research data collected relevant to that content should be eliminated from the study or left uninterpreted. If a large proportion of students (but not the teacher) agree that certain material was not covered, that material may be ignored or interpreted cautiously. In this case, pupils' perceptions may help identify content areas which need greater emphasis, clarification, or measurement.

Guideline 5: Short Measurement Periods Are Preferable to Long Measurement Periods

Teacher effectiveness research suggests that the length of a measurement period may be as short as a single lesson or as long as a full year of instruction. While both extremes are possible, a short measurement period,

incorporating a series of interrelated lessons, commonly referred to as a teaching unit, is preferable. A long measurement period may be appropriate, however, when higher-order objectives (those requiring the pupil to analyze, synthesize and evaluate content) are taught.

A factor favoring shorter measurement periods is the tendency of events outside of the classroom to interfere with the teacher's instructional plans. When measurement covers a relatively brief span of time, factors such as the pupil's home life, the instructional experiences he encounters outside of school (via the library, television, and his peer-groups), and the effect of holidays and vacations, are less likely to influence his performance. A longer measurement period, on the other hand, affords many more opportunities for factors outside the classroom to affect pupil outcomes. A relatively brief assessment period, linked to a specific area of instruction, reduces the chances that potent forces external to the classroom will interact with the behaviors and skills being measured.

Shorter measurement periods, though, require multiple assessments obtained at systematic intervals throughout the school year. Single assessments, while minimizing the influence of external factors on pupil performance, increase the chances of measuring teacher behavior which is atypical. Though assessments should cover a relatively brief period of time (the span of a lesson or a unit), they should be conducted repeatedly, throughout the school year. These can be planned randomly to obtain a general "picture" of teacher behavior, or systematically to capture behaviors or skills associated with particular content areas and teaching objectives.

Guideline 6: Adjusted Gain Rather Than Raw Gain Should Be Used for the Analysis of Pupil Growth

The term raw gain refers to the difference between a pupil's pre- and posttest score while the term adjusted gain refers to a considerably more

complex score derived from several intermediate calculations. Although raw gain scores are sometimes used to assess pupil change, adjusted gain is preferable. Raw gain scores suffer from several critical deficiencies which render them virtually uninterpretable.

Two of these deficiencies are unreliability and susceptibility to distortion by the regression effect. The regression effect refers to the tendency of scores which deviate considerably from the mean to approximate, or lean toward, the mean on subsequent assessments. This phenomenon affects the measurement of pupil change when a student's pretest score is subtracted from his posttest score in order to obtain a "difference score." Those pupils scoring high on the pretest tend to score lower on the posttest, and vice versa, regardless of the average gain or loss registered for the entire class. This regression effect is particularly distressing since it operates unequally on pupils. That is, one pupil's posttest score may be affected by his pretest score to a greater degree than another pupil's posttest score. This differential effect of the pretest upon the posttest distorts any meaning the raw gain score might have for determining pupil change.

To correct for this distortion, residual gain or a conceptually similar technique, analysis of covariance, must be used.[6,7] Residual gain is computed by correlating the pre- and posttest scores of all pupils, predicting a posttest score for each pupil on the basis of his pretest score, and subtracting it from his actual posttest score. This procedure creates a measure of gain which is independent of the pupil's initial standing and, therefore, more representative of the true change which has occurred during the measurement period. Analysis of covariance, which statistically holds constant the effect of the pretest scores, can be used to accomplish this same objective in a more efficient manner by offering greater detection power, i.e., reducing the probability of failing to reject a false null hypothesis (Type II error).[8]

The raw gain acore, besides being subject to distortion caused by the regression effect, is also notoriously unreliable. The use of two scores (pre- and posttest) in calculating raw gain assumes that any difference between the two is due to the effect of intervening instruction. This procedure also assumes that any gain from pre- to posttest indicates pupil improvement. As noted earlier, the researcher often overlooks the fact that the gain score is derived from two measures which are less than totally reliable. The raw gain score inherits unreliability from both the pre- and the posttest and is therefore considerably less reliable than either of the sources from which it is derived. For example, if the correlation between pre- and posttest is .70 and the reliability of each is .80 (coefficients which in practice are fairly common), then the reliability of the gain score would be .33. Clearly, raw gain scores are not sufficiently reliable to serve as indices of pupil change.

References

Alschuler, A. S. The effects of classroom structure on achievement motivation and academic performance. Educational Technology, August, 1969, 19-24.

Borich, G. D. The appraisal of teaching: Concepts and process. Reading, Mass.: Addison-Wesley, 1977 (in press).

Borich, G. D., & Madden, S. K. Evaluating classroom instruction: A sourcebook of instruments. Reading, Mass.: Addison-Wesley, 1977 (in press).

Brophy, J. Stability of teacher effectiveness. American Educational Research Journal, 1973, 10, 245-252.

Brophy, J., & Evertson, C. Process-product correlations in the Texas teacher effectiveness study: Final report (Res. Rep. 74-4). Austin, Texas: Research and Development Center for Teacher Education, 1974.

Chall, J. S., & Feldmann, S. C. A study in depth of first grade reading (U.S. Office of Education Cooperative Research Project No. 2728). New York: The City College of the City University of New York, 1966.

Christensen, C. M. Relationship between pupil achievement, pupil affect need, teacher warmth and teacher permissiveness. Journal of Educational Psychology, 1960, 51(3), 167-174.

Cronbach, L. J., & Furby, L. How should we measure "change"--or should we? Psychological Bulletin, 1970, 74, 68-80.

Cronbach, L. J., Glaser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Dacey, J. S., & Madaus, G. F. An analysis of two hypotheses concerning the relationship between creativity and intelligence. Journal of Educational Research, 1971, 64(5), 213-216.

DeBlassie, R. R. A comparative study of the personality structures of persistent and prospective teachers. Journal of Educational Research, 1971, 64(7), 331-333.

Doyal, G. T., & Forsyth, R. A. The relationship between teacher and student anxiety levels. Psychology in the Schools, 1973, 10, 231-233.

Duffey, J. B., & Martin, R. P. The effects of direct and indirect teacher influence and student trait anxiety on the immediate recall of academic material. Psychology in the Schools, 1973, 10, 233-237.

Dumas, W. Factors associated with self-concept change in student teachers. Journal of Educational Research, 1969, 62(6), 275-278.

Fuller, F. F. Concerns of teachers: A developmental conceptualization. American Educational Research Journal, 1969, 6(2), 207-226.

Glass, G. V. Teacher effectiveness. In H. J. Walberg (Ed.), Evaluating educational performance. Berkeley, California: McCutchan Publishing Corporation, 1974.

Horn, J. L., & Morrison, W. E. Dimensions of teacher attitudes. Journal of Educational Psychology, 1965, 56, 118-125.

Kahn, S. B., & Weiss, J. The teaching of affective responses. In R. M. W. Travers (Ed.), Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Kash, M. M., Borich, G. D., & Fenton, K. S. Teacher behavior and pupil self-concept. Reading, Mass.: Addison-Wesley, 1977 (in press).

Knoell, D. M. Prediction of teaching success from word fluency data. Journal of Educational Research, 1953, 46, 673-683.

Krasno, R. M. Teachers' attitudes: Their empirical relationship to rapport with students and survival in the profession (Research Mono. Tech. Report No. 28). Stanford, California: Stanford Center for Research and Development in Teaching, School of Education, Stanford University, June 1972.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley Publishing Company, 1968.

Loree, M. R. Shaping teachers' attitudes. In B. O. Smith (Ed.), Research in teacher education: A symposium. Englewood Cliffs, N.J.: Prentice-Hall, Inc., 1971.

Marjoribanks, K. Bureaucratic structure in schools and its relationship to dogmatic leadership. Journal of Educational Research, 1970, 63(8), 353-357.

McCallon, E. L. Teacher characteristics and their relationship to change in the congruency of children's perception of self and ideal-self. Journal of Experimental Education, 1966, 34(4), 84-86.

McDonald, F. J., Elias, P., Stone, M., Wheeler, P., Lambert, N., Calfee, R., Sandoval, J., Ekstrom, R., & Lockheed, M. Final Report of Phase II Beginning Teacher Evaluation Study. Prepared for the California Commission of Teacher Preparation and Licensing, Sacramento, California. Princeton: Educational Testing Service, 1975.

Neale, D. C., Gill, N., & Tismer, W. Relationship between attitudes toward school subjects and school achievement. Journal of Educational Research, 1970, 63(5), 232-237.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40(5), 647-662.

Rosenshine, B. Teaching behaviours and student achievement. London: International Association for the Evaluation of Educational Achievement, 1971.

Rosenshine, B. Classroom instruction. In N. L. Gage (Ed.), The NSSE 77th Yearbook, The Psychology of Teaching Methods, 1976.

Rutherford, W. L., & Weaver, S. W. Preferences of elementary teachers for pre-service and in-service training in the teaching of reading. Journal of Educational Research, 1974, 67(6), 271-275.

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

Shavelson, R. J., & Dempsey, N. K. Generalizability of measures of teacher effectiveness and teaching process (Tech. Report No. 75-4-2), Beginning Teacher Evaluation Study. San Francisco, California: Far West Laboratory for Educational Research and Development, 1975.

Shavelson, R. J., & Dempsey, N. K. Generalizability of measures of teaching process. In G. D. Borich, The appraisal of teaching: Concepts and process. Reading, Mass.: Addison-Wesley, 1977 (in press).

Soar, R. S. Assessment problems and possibilities. Journal of Teacher Education, 1973, 24, 205-212.

Soar, R. S., Soar, R. M., & Ragosta, M. Change in classroom behavior from Fall to Winter for high and low control teachers. Paper presented to the annual meeting of the American Educational Research Association, Chicago, 1973.

Solomon, D., Bezdek, W. E., & Rosenberg, L. Teaching styles and learning. Chicago: The Center for the Study of Liberal Education of Adults, 1963.

Stallings, J., & Kaskowitz, D. Follow Through Classroom Observation Evaluation 1972-1973. Menlo Park, California: Stanford Research Institute, 1974.

Travers, R. M. W. (Ed.). Second handbook of research on teaching. Chicago: Rand McNally, 1973.

Treffinger, D. J., Feldhusen, J. F., & Thomas, S. B. Relationship between teachers' divergent thinking abilities and thier ratings of pupils' creative thinking abilities. Measurement and Evaluation in Guidance, 1970, 3(3) 171-176.

Wallen, N. E. Relationships between teacher characteristics and student behavior: Part three. (U.S. Office of Education Research Project No. SAE OE 5-10-181). Salt Lake City: University of Utah, 1966.

Weiss, R. L., Sales, S. M., & Bode, S. Student authoritarianism and teacher authoritarianism as factors in the determination of student performance and attitudes. Journal of Experimental Education, 1970, 38(4), 83-87.

Yonge, G. D., & Sassenrath, J. M. Student personality correlates of teacher ratings. Journal of Educational Psychology, 1968, 49(1), 44-52.

Footnotes

[1]Some of my esteemed U.S. colleagues may differ with me on this point. While the issues raised by our differences are perhaps too complex to present in their entirety here, several should be mentioned. The tendencies to (1) report significant findings which fail to exceed the number expected by chance and (2) ignore differences in the operational definitions of purportedly similar constructs serve as examples of the problems which have either reduced the credibility of "significant" findings or led to the proliferation of "null" findings.

Rosenshine's review (1971) illustrates these problems. Rosenshine examined the findings of approximately 50 different studies in which over 200 separate teacher behaviors were investigated. On the basis of evidence from these studies, 11 behaviors were selected as potentially promising in relation to pupil performance. In interpreting the efficacy of these 11 behaviors, however, we must remember that they were derived, for the most part, from correlational, not experimental, studies. Therefore, causation cannot be inferred. Furthermore, these behaviors were derived from clusters of heterogeneous research studies which actually showed mixed results; some studies within a given cluster failed to confirm the efficacy of the variable in question. Also, variables were often operationally defined differently by different investigators. And finally, in some studies the number of significant findings failed to exceed that which could be expected by chance.

The problem of operational definitions is illustrated by the teacher variable clarity, which, Rosenshine points out, has been defined in three very different ways:

(1) whether "the points the teacher made were clear and easy to understand" (Soloman, Bezdek, & Rosenberg, 1963);

(2) whether "the teacher was able to explain concepts clearly... had the facility with her material and enough background to answer her children's questions intelligently" (Wallen, 1966);

(3) whether the cognitive level of the teacher's lesson appeared to be "just right most of the time" (Chall & Feldman, 1966).

The problem of chance significance is illustrated by a finding of my own which, I suspect, is not uncommon. I recently had occasion to analyze the extent to which process-product relationships in a large-scale teacher effectiveness study replicated over two consecutive years, during which time instrumentation and teacher sample remained constant. Of the 3,050 relationships my colleagues and I studied, only 24 were significant at $p < .10$ in the same direction for both years! A much more favorable result, of course, would have been expected on the basis of chance alone. Unfortunately, since few replications of this type are conducted, teacher behaviorists may never discover how unstable their findings actually may be.

[2]Studies by Stallings and Kaskowitz (1974) and by Brophy and Evertson (1974) indicate the large number of variables customarily studied in field research of this kind and the number of significant findings which are obtained before replication.

[3]It is important to note that only one study (Shavelson & Dempsey, 1977) has closely examined the stability of teacher behavior qua teacher behavior as opposed to inferring the stability of teacher behavior from its presumed effect on pupils. I will return to this point later to demonstrate the fallacy in this inference and the need for both types of stability studies.

[4]It is entirely possible that for some indices of teacher behavior the number of occasions and raters needed to reach an acceptable level of reliability would outstrip one's resources. In this case, it must be assumed that the behavior of interest is what I prefer to call logically unstable as opposed to psychometrically unstable. Also, this is the point at which the definition of reliability turns to one of generalizability, i.e., the generalizability of the construct measured over different conditions and raters. Cronbach et al. (1972) make an excellent case for designing studies which can assess an instrument's generalizability over different facets, or experimental conditions (i.e., sources of variance), as opposed to simply reporting the reliability of an instrument in a single context as our classical definitions of reliability (Lord & Novick, 1968) suggest.

[5]Only theoretically will it be the same, in which case we must assume no specific variance. Practically speaking, this is a near-to-impossible event.

[6]Residual gain, unfortunately, is not an entirely satisfactory correction for the regression effect. It requires adjustment, depending on the extremeness of posttest scores. A gain score is increased if the pretest score is high and decreased if it is low. In other words, pupils who score high on the pretest have points added to their posttests (because the regression effect has artificially pushed their posttest scores down, toward the mean), and pupils who score low on the pretest have points subtracted from their posttest (because the regression effect has artifically pushed their posttest scores up, toward the mean). Since the amount of adjustment depends on the position of the pretest score in relation to the mean, it varies from pupil to pupil. Unfortunately, the adjustment also depends on the characteristics of the pupils being tested, and this information is generally unavailable.

[7]While residual gain scores and analysis of covariance are repeatedly discussed in the literature (Rosenshine, 1971; Soar, 1973) as "parallel" techniques, they, in fact, are not. These different computational procedures are not mathematically equivalent and, therefore, can, in any given research effort, lead to quite different results, introducing the distinct possibility that the researcher may reject the null hypothesis with one technique and fail to reject it with the other. Generally, analysis of covariance is the preferred technique since its power to detect a significant finding, when one is present, exceeds that of the residual gain procedure. Hence, I refer to these techniques as conceptually similar because, though they are fallable (Cronbach & Furby, 1970), they both offer methods of dealing with pretest performance.

[8]The analysis of covariance procedure can be represented by the full model, $Y = a + b_1TB + b_2Pre + e$, and the restricted model, $Y = a + b_3Pre + e$, where TB is the teacher behavior of interest, Pre is the pupils' pretest achievement, and Y is pupil posttest achievement. The multiple correlation coefficient for the full model (R^2) minus the R^2 for the restricted model describes the relationship between teacher behavior and pupil posttest achievement with pretest held constant or, equivalently, $R^2_{full} - R^2_{res}$ is the squared part correlation between teacher behavior and pupil posttest performance, with pretest partialed out. See Porter, A., & Chibucos, T. Selecting analysis strategies. In G. D. Borich (Ed.), Evaluating educational programs and products. Englewood Cliffs, N.J.: Educational Technology Publications, 1974.