DOCUMENT RESUME

| | | |
|---|---|---|
| ED 107 600 | 95 | SP 009 223 |

| | |
|---|---|
| AUTHOR | Medley, Donald M; And Others |
| TITLE | Assessment and Research in Teacher Education: Focus on PBTE. PBTE Monograph Series No. 17. |
| INSTITUTION | American Association of Colleges for Teacher Education, Washington, D.C. |
| SPONS AGENCY | Office of Education (DHEW), Washington, D.C. |
| PUB DATE | Jun 75 |
| NOTE | 51p. |
| AVAILABLE FROM | Order Department, American Association of Colleges for Teacher Education, Suite 610, One Dupont Circle, Washington, D.C. 20036 ($3.00) |
| EDRS PRICE | MF-$0.76 HC-$3.32 PLUS POSTAGE |
| DESCRIPTORS | Affective Behavior; Behavioral Objectives; Effective Teaching; Evaluation Criteria; *Performance Based Teacher Education; *Performance Criteria; *Program Evaluation; Program Improvement; Student Behavior; *Teacher Behavior; *Teacher Evaluation |

ABSTRACT

This monograph presents a diagram which distinguishes four different assessment levels in the teacher's professional development. Level 1 refers to assessments of the training experience, level 2 to assessments of the teacher's behavior, level 3 to assessments of pupil behavior, and level 4 to assessments of instruction. The diagram explains that level 4 is influenced by level 3, level 3 by level 2, and level 2 by level 1. It is explained that program decisions based on assessments were traditionally made at level 1. The validity of levels 2, 3, and 4 are then examined. It is held that level 2 is the level at which teaching should be evaluated. It becomes necessary, however, to determine competencies for this evaluation. The document presents a hierarchy of relevant competencies, including broad concepts, general characteristics, and specific behavior items. Program evaluation is also examined, and the statement is made that it must be demonstrated that teacher education programs can produce the kinds of teacher behaviors which in turn produce more growth in pupils. It is then necessary to use pupil behavior in program evaluation. The problems this causes in teacher evaluation can be avoided, however, through the use of large samples. The monograph also discusses lack of knowledge as a weakness in specifying competencies and developing programs. (PB)

# ASSESSMENT AND RESEARCH IN TEACHER EDUCATION: FOCUS ON PBTE

by

Donald M. Medley
Professor of Education
University of Virginia
Charlottesville, Virginia

Robert S. Soar
Professor of Education
University of Florida
Gainesville, Florida

Ruth Soar
Florida Educational Research
and Development Council
Gainesville, Florida

for the AACTE
Committee on Performance-Based Teacher Education

June 1975

American Association of Colleges for Teacher Education
One Dupont Circle, Suite 610
Washington, D.C. 20036

# PBTE COMMITTEE MEMBERS

*Lorrin Kennamer*, Chairman; Dean, College of Education, University of Texas at Austin, Austin, Texas 78712

*Patricia Cabrera*, Director, Teacher Corps Project, School of Education, Phillips Hall of Education, University of Southern California, Los Angeles, California 90007

*Patrick L. Daly*, Vice President of AFT and Social Studies Teacher, Edsel Ford High School, 20601 Rotunda Drive, Dearborn, Michigan 48124

*William Drummond*, Professor of Education, Department of Curriculum and Instruction, College of Education, University of Florida, Gainsville, Florida 32601

*L. Harlan Ford*, Deputy Commissioner for Programs and Personnel Development, Texas Education Agency, Austin, Texas 78701

*Tommy Fulton*, President of Oklahoma Education Association and Art Teacher, Jarman Jr. High School, Midwest City, Oklahoma 73110

*David Krathwohl*, Dean, College of Education, Syracuse University, Syracuse, New York 78712

*Margaret Lindsey*, Professor of Education, Teachers College, Columbia University, Box 135, 525 W. 120th Street, New York, New York 10027

*Donald Medley*, Chairman, Department of Research Methodology, School of Education, University of Virginia, Charlottesville, Virginia 22903

*Gilbert Shearron*, Chairman, Elementary Education and Early Childhood Education, 427 Aderhold Building, University of Georgia, Athens, Georgia 30601

## STAFF

*Karl Massanari*, Associate Director, AACTE; Director, PBTE Project

*Shirley Bonneville*, Program Associate, PBTE Project

*Nancy Hoagland*, Manager, PBTE Information Center

*Jane Reno*, Technical Editor

*Sharon DeVeauuse*, Secretary

*Brenda Belton*, Secretary

*Mona Chase*, Secretary

The activity which is the subject of this report was supported in whole or in part by the U.S. Office of Education, Department of Health, Education and Welfare, through the Texas Education Agency, Austin, Texas. However, the opinions expressed herein do not necessarily reflect the position or policy of the U.S. Office of Education or the Texas Education Agency and no official endorsement by the United States Government should be inferred.

Library of Congress Catalog Card Number: 75-4308

Standard Book Numbr r: 910052-88-3

# Preface

The American Association of Colleges for Teacher Education (AACTE) is pleased to publish this paper as one of a series of monographs sponsored by its Committee on Performance-Based Teacher Education. The series is designed to expand the knowledge base about issues, problems, and prospects regarding performance-based teacher education as identified in the two papers on the state of the art developed by the Committee itself.[1,2]

Whereas these two papers are declarations for which the Committee accepts full responsibility, publication of this monograph (and the others in the PBTE Series) does not imply Association or Committee endorsement of the views expressed. It is believed, however, that the experience and expertise of these individual authors, as reflected in their writings, are such that their ideas are fruitful additions to the continuing dialogue concerning performance-based teacher education.

This monograph addresses one of the critical problems in designing and implementing performance-based teacher education programs, namely, the assessment of teacher performance. The problem, however, is not unique to PBTE. All of teacher education faces the problem of evaluating program effectiveness through the assessment of the performance of graduates. The design presented is a significant addition to the literature not only about PBTE but about all teacher education.

AACTE acknowledges with appreciation the role of the National Center for Improvement of Educational Systems (NCIES) of the U.S. Office of Education in the PBTE Project. Its financial support (provided through the Texas Education Agency) as well as its professional stimulation, particularly that of Allen Schmieder, are major contributions to the Committee's work. The Association acknowledges also the contribution of members of the Committee who served as readers of this paper. Special recognition is due Lorrin Kennamer, Committee Chairman; David R. Krathwohl, member of the Committee and chairman of its publications task force; and to Shirley Bonneville and Jane Reno of the Project staff for their contributions to the development of this publication.

*EDWARD C. POMEROY*
Executive Director, AACTE

*KARL MASSANARI*
Associate Director, AACTE
and Director, PBTE Project

[1] Stanley Elam, *Performance-Based Teacher Education: What Is the State of the Art?* (Washington, D.C.: The American Association of Colleges for Teacher Education, December 1971).

[2] AACTE Committee on Performance-Based Teacher Education, *Achieving the Potential of Performance-Based Teacher Education: Recommendations* (Washington, D.C.: The American Association of Colleges for Teacher Education, February 1974).

# Introductory Note

Does the PBTE movement have its feet on the ground? If so, it has a couple of very important Achilles' heels: 1) the problem of measuring or assessing the performances or competencies and 2) once they can be measured or assessed, their validation as behaviors that make a difference in student learning. The PBTE Committee, which sponsors this monograph series, has stressed these problems in the recommendations of Monograph No. 16, which summarized its first three years of work. Often incorrectly perceived as an advocate of PBTE, the Committee in reality is concerned that PBTE be properly implemented as *one* means of teacher education. *Then* let's see what it car contribute. Such an appropriate trial *cannot* come about unless the two problems mentioned above are much closer to solution than at present.

Do these problems mean that we should not try to implement such programs until then? Not at all. However, the existence of the problems suggests that making everyone conform to a PBTE mode is unwarranted. But, through carrying program development as far as we can and doing the best possible job of evaluation, we can begin to build research into these programs that will help us to determine what characteristics make a difference.

Therefore, this monograph is viewed by the Publications Subcommittee as one of its most important. The exploration of the psychometric realities of assessment, which is in the early part of the monograph, is a problem that workers in the field intuitively sense, but I know of no place where it has been laid out as it is here.

Here are some quotations to suggest what is in store for you in this monograph:

> It seems probable that there is a negative relationship between the social importance of a specific goal and the length of time it takes the average pupil to show appreciable growth toward it.

On the reliability of measures of class gain to assess a teacher:

> the (research) data are sparse but consistent.....(and indicate that) we would have to test each teacher in at least 20 different classes ....to obtain...minimally acceptable reliability.....

On observer measures:

> ....it is critical to establish empirically that....(the) items do belong together....because it is quite likely that some items that appear to belong together....(conceptually) will not hang together empirically.....This is what happened to us when we tried to assess teacher control.....

Enough to intrigue you? There is a lot more!

The monograph contains important conceptual points and, from three of the very best researchers in the field, much sage advice for successful work in assessing and validating teacher behaviors.

A topic of this kind is not easy to develop without assuming some background on the part of the reader. The authors have assumed as little as possible. We have tested their monograph for readability and find that students with one course in tests and measurements have not had trouble in comprehending it.

For all these reasons and more, we commend it to your attention.

*DAVID R. KRATHWOHL,* Member of the
PBTE Committee and *Chairman* of its
Task Force on Publications

## Contents

Page

Figures

# Introduction

If there is a single word that describes the role of assessment in a performance-based teacher education (PBTE) program, the word is *crucial*. The very name implies that the decision base in such a program is performance, that is, demonstrated competence. Decisions about the routing and progress of a student through a PBTE program and out of it into the public schools are, then, by definition based on assessments of teacher performance.

This fact is generally accepted; in two seminal publications of the American Association of Colleges for Teacher Education's PBTE Committee (Elam, 1971, pp. 6-7; The AACTE Committee, 1974, pp. 7, 30), the cruciality of the role of teacher performance assessment is emphasized as part of the definition of performance-based instruction itself.

In the development of most of the PBTE programs in current operation, much more attention has been paid to such problems as those of developing modules and reorganizing instruction than to the development of adequate assessment procedures. The probable consequences of this neglect are also cited by the Committee (*op. cit.*, pp. 40-41); a particularly forceful statement by Krathwohl says that:

> One can predict that performance-based teacher education (PBTE) is certain to fail to reach its ultimate objective if it continues on its present course. This failure will be caused by the almost complete lack of attention given to the assessment of teaching competencies, a core concept of PBTE. (Merwin, 1973, p. v)

We suspect that this neglect of the assessment problems is not entirely the result of conflicting demands on program developers' time. Our contacts with these harried individuals reveal that they have a strong reluctance to tackle these problems because of feelings of inadequacy arising from the lack of knowledge of adequate approaches to solutions. It will be the purpose of this monograph to make some suggestions designed to allay these feelings. Little enough is known about the assessment of teacher competence to indicate that even the modest proposals we feel qualified to make may be useful to program developers.

We will begin by presenting a paradigm or cognitive map of the area and by attempting to make the task more manageable by breaking it down into smaller subtasks. Specifically, we shall deal separately with the three principal areas: (1) developing techniques for assessing teacher performance, (2) specifying the competencies to be assessed in measurable terms, and (3) validating the program by validating the competencies it develops in its graduates.

Each of these tasks is too complex, too difficult, to be treated adequately in the space available. Nor, for that matter, is enough known about any one of the topics to justify an attempt at definitive treatment. All that we can possibly claim is that we have tried to shed some light on each by drawing on our own experience as researchers in teacher behavior.

## A Simple Paradigm

It will be convenient to distinguish four different levels in the teacher's professional development at which the teacher may be assessed, as shown in Figure 1.

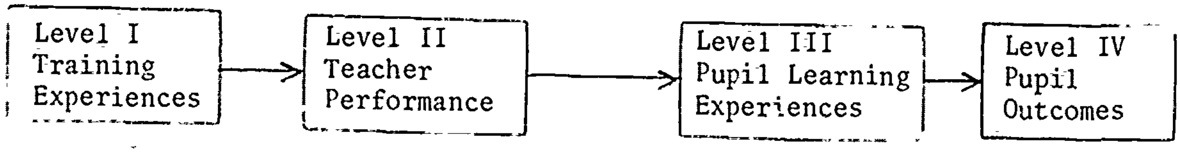

Level I Training Experiences
Level II Teacher Performance
Level III Pupil Learning Experiences
Level IV Pupil Outcomes

Figure 1. Assessment Levels in Teacher Education

Level I refers to assessments of the *training experiences* the teacher has had: What courses has he taken? What modules has he attempted? Which ones has he mastered? Which has he bypassed on the basis of having demonstrated mastery of its objectives beforehand?

Level II refers to assessments of the *teacher's behavior* while he is attempting to fulfill the role of a teacher. What kinds of questions does he ask in interaction with pupils? How does he organize his class for instruction? How does he determine the objectives of instruction?

Level III refers to assessment of the *behaviors of pupils* under the guidance of the teacher being assessed--assessments of the experiences they have which we all know must form the basis for any learning that takes place. What kinds of tasks do the pupils perform, in or out of class? How much time do they spend in active participation in class discussions? How often does each child receive reinforcement and for what?

Level IV refers to assessments of the *outcomes of instruction*--of those changes in behavior that it is the purpose of education to bring about. How well does the pupil read? What are his attitudes toward independent learning in adult life? What kind of a citizen does he become?

Each of these four assessment levels is represented in the diagram by a rectangle and the rectangles are joined by arrows representing lines of influence or of cause-and-effect. Thus outcomes (Level IV) are seen as influenced by, in part the result of, pupil behavior (Level III), of learning experiences the pupil has while in school; these experiences in turn are seen as at least partly determined by what his teacher does (Level II); and, finally, the way the teacher behaves--"teaches"--is affected by the experiences he has had during training (Level I). All of these things are, of course, strongly influenced by other factors not shown in the figure, e.g., community, school, pupil, and teacher characteristics.

The whole enterprise of teacher education is, of course, based on the assumption that, despite these extraneous factors, the influences that are assessed at each stage are potent enough to have an appreciable effect not only on the level immediately following but on all subsequent levels. The concept of *teacher effectiveness*, in particular, is based on the notion that pupil learning outcomes (Level IV) are affected by teacher behavior (Level II). And the justification for the very existence of teacher education is the presumption that what happens to a teacher in training (Level I) can somehow increase his effectiveness, that is, affect pupil learning outcomes (Level IV).

## Implications of the Paradigm

The very existence of intermediate Levels II and III suggests that the effects of training on teacher effectiveness are subject to attenuation and are, therefore, difficult to establish. Any impact on pupil learning (Level IV) of teacher behavior (Level II) must be achieved through pupil behaviors (Level III). Two teachers who behave identically will achieve the same outcomes only if their pupils also behave in the same way. If we attempt to relate teacher behavior to pupil learning without paying some heed to what the pupils are doing, we should not be surprised to find that the correlations tend to be low.

In the same way, the relationship between teacher education (Level I) and effective instruction (Level IV) depends on what teachers and pupils do in Levels II and III. We must train teachers to behave in such a fashion that their pupils will behave in such a fashion that the pupils will learn more!

## The Concept of PBTE

We now see that the essential innovation involved in performance-based teacher education is a simple one: when program decisions are based on assessments, they should be made at some level higher than Level I. In the past, decisions about when a teacher education student is ready to graduate or to be certified, promoted, to receive merit pay, or the like, have been based mainly on Level I assessments: on what courses the teacher has had, what degrees or other evidence of training he can present, or (on the assumption that experience is the best teacher of teachers) on how much experience he has had. The PBTE notion is that such decisions must be based on demonstrated competence rather than on evidence of training or experience supposedly related to competence. The problem is: how and at what level should competency be assessed? One controversial issue is whether competence can be assessed at Level II or whether it must be assessed at Level IV. It is to this question that we shall next address ourselves. Let us first discuss the feasibility of Level IV assessment.

### Assessing Teacher Competence in Terms of Pupil Outcomes (Level IV)

It seems to us that much of the enthusiasm for PBTE manifested in the past is based on the impression that Level IV assessment will be used.

Claims that a PBTE program would graduate only teachers who have actually demonstrated their competence to teach have been taken as meaning that only teachers of proven *effectiveness* will be turned out in PBTE programs.

This concept has a very attractive sound to the school administrator, the state certification officer, the legislator, the taxpayer. The factory that manufactures television sets releases only functioning sets; quality control insures this. Why should the college of education not do the same?

In a period of time when accountability has become a catchword, such a proposition seems particularly reasonable. Why should we who train teachers not require each and every candidate for certification to demonstrate that he can teach something to real pupils and have them learn it?

It is our contention that, attractive though it may sound, assessment at Level IV in teacher education is not a viable strategy. To support this claim, it is only necessary to ascertain the degree to which Level IV measures are likely to possess the three essential characteristics of a useful test or other measuring device: *validity, reliability,* and *practicality.* Let us examine each of these terms as they apply to Level IV assessments as the basis for decisions in performance-based teacher education and certification.

We shall begin by discussing the validity of such measures of teacher effectiveness.

## Validity of Level IV Measures

Before we go any further, let us make clear what we mean by a Level IV measure which we shall call a measure of *teacher effectiveness.* Such a measure is based on pupil gains on a test or other measure of the outcomes of instruction. Typically, a group of pupils is pretested, taught by the teacher for a prescribed period of time, and posttested. The mean gain, usually adjusted statistically to eliminate such influences as pretest and ability, is taken as the measure of outcomes and, in this case, of the effectiveness of the teacher. Teacher effectiveness must be measured in terms of effects on pupils.

The validity of so direct a measure of a teacher's ability to get pupils to learn seems self-evident and it may be for the limited type of learning usually assessed in this manner, but not necessarily for any other type. Most serious attempts to use Level IV assessments--or "teaching tests"--as devices for measuring teachers (cf. Flanders, 1974; Popham, n. d.; 1971) have for obvious reasons used relatively short periods of instruction--a few hours or a few days at the most. This obviously limits them to measurements of the kinds of effects on pupils that can be detected in a relatively short time.

We know of no systematic research into the length of time it takes to produce measurable gains toward various types of objectives of instruction. Such research is badly needed. In any case, it seems likely that much less time is needed to teach some facts, especially if they need to be retained only long enough to pass a unit test, than to help ρupils become self-directed learners, to serve, or to become responsible citizens. It also seems probable that there is a negative relationship between the social importance of a specific goal and the length of time it takes the average pupil to show appreciable growth toward it.

This means that "teaching tests" of this type can validly measure how effective a teacher is in achieving only short-term goals, which are almost certainly the least important goals of education. The adoption of this type of measure of teacher competence would systematically select as "best" those teachers who are good at teaching facts to pupils, facts the pupils may well forget as soon as they have passed the unit test; while the teacher who sacrifices this kind of learning for the sake of achieving more important "higher level" outcomes might tend to be eliminated or at least scored as less effective. If such an approach could and did work, its effects on public education could be disastrous.

Little is known about the structure of teacher effectiveness; it may be that the teacher who is most effective in teaching pupils facts is also most effective in teaching them to analyze and synthesize, to appreciate literature, etc.; but this suppcsition is at best doubtful. We shall later cite some evidence that it is not so. For the present, let us conclude that the validity of "teacher tests" of ability to achieve short-term outcomes as predictors of overall teacher effectiveness is by no means self-evident. If Level IV measures are to be used, they should be based on measures of multiple outcomes and on a period of teaching long enough to detect progress toward long-term goals--a minimum of a semester or two--before we may assume them to be valid.

Short-term teacher tests seem likely to measure something more like coaching or cramming skill than teacher effectiveness as usually conceived. In *any* case, their validity as predictors of overall effectiveness must be empirically demonstrated before their use is justified.

When standardized achievement tests are used to measure long-term pupil gains as a basis for teacher evaluation, additional problems emerge. For one, test validity is threatened by the well-known tendency to "teach to the test", that is, to emphasize the specifics measured by the test. The content validity of such a test is based on the assumption that the items on the test sample a large domain of items, so that the performance of the pupil on the test (his obtained score) is an unbiased estimator of his performance on all items in the domain (his true score). When a teacher teaches to the test, this assumption is untenable and the validity of the test is destroyed and with it the validity of the measure of teacher effectiveness based on it.

A recent illustration of the power of this effect is provided by an OEO study of performance contracting in which what looked like evidence of success vanished when control was exercised to eliminate "teaching for the test." (Page, 1972, 1973).

Unless such controls are used, evaluation of teachers based on mean gains of pup' s is likely to identify as most competent a type of teacher nobody wants.

The suggestion has been advanced that instead of using mean gains on achievement tests we use the number of pupils in a class who achieve mastery of the tested material at some specified minimum level of achievement, but this presents problems as well. Small (1972) has documented from the history of accountability in England a century ago the fact that when teachers are evaluated on this basis, they tend to focus their efforts on pupils at or near the specified level to the detriment of pupils at either higher or lower levels. Once again we run the risk of rewarding the wrong kind of teaching.

The general problem is that attempts to evaluate teachers on the basis of pupils' test performance tend to focus the teaching too narrowly on the specifics measured by the test.

## Reliability of Level IV Measures

Before commenting on this topic, let us agree on what the term *reliability* means. By a reliable measure we mean one that yields an obtained score quite close to the true score of the person measured. In the case of a Level IV measure, the true score of a teacher would be the adjusted mean gain score of all pupils in some population of pupils all of whom had been taught by the teacher being assessed for the prescribed length of time under the prescribed conditions, etc. Clearly, the principal score of error of measurement in this instance arises from differences among pupils. A teacher who can teach something to one child with ease might have difficulty teaching the same thing to another child, particularly if the two children differed in ability, interest, sex, race, or socioeconomic status. At present we tend to train teachers to teach anybody. We may limit the grade or subject a teacher is to teach, but we do not, as a rule, restrict a teacher to pupils at a certain level of IQ, of a certain sex or race, level of interest, etc. The population of pupils on which the "true" score under discussion is based must, then, be regarded as quite heterogeneous in these respects.

Suppose, now, that each teacher to be assessed is required to teach a certain unit to one class of pupils and that his effectiveness with that class has been ascertained. How reliable is such a score? Its measure of stability could be estimated by having each of the teachers teach not just one but two classes regarded as randomly drawn from the total population of pupils and correlating the two sets of mean gain scores.

Fortunately, some data reporting just such stability coefficients exist. Rosenshine (1970) has reviewed five published reports of such studies, and Veldman and Brophy (1974) report some new data. The median reliability coefficient in the studies reviewed by Rosenshine is *.32*; the median coefficient reported by Veldman and Brophy is *.27*.*

*The value of .32 is a crude median of 13 coefficients reported in Rosenshine's Table 1. The value of .27 is the crude median of 30 coefficients in Veldman and Brophy's Table 5; because these are reliabilities of the mean of 3 measures per teacher, the computed median (.52) was reduced by the Spearman-Brown formula to estimate the reliability of a single measure. More refined methods yielded very similar estimates.

These data are sparse, but surprisingly consistent and imply that if we use the mean gains of one class of pupils taught by a teacher as a measure of teacher effectiveness, our "teaching test" may be expected to have a reliability coefficient of about .3. This means that more than 90 per cent of the variance in such scores must be attributed to unknown influences--to chance. The competence of the teacher assessed accounts for less than one-tenth of the variance. Not a very good basis for program decisions!

Most textbooks in tests and measurements recommend that the minimum reliability for a test to be used to measure individuals should be .90 or .95. To achieve this standard, we would have to test each teacher in at least 20 different classes in order to obtain a measure of minimally acceptable reliability (.90) with a teaching test--a procedure which is out of the question.

Quite aside from the question of their validity, Level IV measures of teacher effectiveness are of doubtful value, then, because of their extremely low reliability.

## Practicality of Level IV Assessments

The practicality of a measuring device has to do with such matters as how much it costs; how long it takes to administer and score it; what its use requires in terms of materials, personnel, and the like; and the availability of alternate forms.

The ideal measure of teacher competence for use in a performance-based teacher education program should be usable not only in the terminal or summative evaluation of a teacher, but also in the formative stages to provide a basis for routing the student through the program. In a modular program in particular, there is need for instruments designed to measure the competence each module is intended to develop, instruments which can be used as pretests to determine whether the teacher should enter or bypass any given module and also as posttests to ascertain whether he has achieved the goal of a module before he passes on to another.

It should be clear that a Level IV measure of teacher effectiveness, one based on pupil outcomes, is not a very practical device for such purposes for at least two reasons. One is that, by its nature, such a test does not focus on a single competence--the complexity of the teaching act means that a number of competencies are involved in teaching the simplest concept, even to the smallest group of pupils. If it were possible to simplify the teacher's task to such a degree that one competency alone were used, the situation would be so artificial that the intrinsic validity of a Level IV measure would be lost.

The other factor limiting the practicality of Level IV measures is that they are far too cumbersome to use. Even in its simplest form, this type of "teaching test" takes a number of hours to administer and requires the time not only of the student but also of se" ral pupils, who must be on call and available whenever a student reaches that point in his own individual progress at which he needs them. Normally, the same children may not be used more than once in the same test so that a truly gigantic pool of pupils must be available. There is no need to elaborate further; the impracticality of such an approach should be obvious.

If the teaching performance test we have been discussing has any use at all, it may be of use at the end of the individual's training. After a student has completed a program and may, therefore, be presumed to have acquired all of the competencies needed to be certified, such a test might be administered to find out whether he is able to put them all together--to deploy what he has learned effectively in dealing with a teaching problem. But except for this specific purpose, we suggest that anything defensible as a Level IV measure is not practicable enough to be useful; even in this one application, the validity and reliability problems are such as to make its utility very doubtful.

## The Morality of Level IV Assessments

Quite aside from the pragmatic questions discussed above, we have some philosophical reservations about Level IV assessments. Use of such devices effectively makes the advancement of one human being--the teacher--dependent on the behavior of another human being--the pupil. The teacher's future depends on events which are not, and should not be, entirely under his control. The teacher's self-interest requires him to manipulate pupils so that they will behave in ways that will result in a favorable evaluation of the teacher. The resultant pressures on the pupils are all the more repugnant in that the pupils may be unaware of them, and constitute no less of a threat to their human rights. The pupil has the ultimate right not to learn--not to behave in the fashion prescribed for him by the teacher or school. And in order to be evaluated as competent, the teacher is virtually forced to violate this right.

## Assessing Teacher Competence Based on Pupil Behavior (Level III)

Even though the relations between teacher behavior and pupil behavior (Level III) are likely to be higher than those between teacher behavior and pupil outcomes, it seems to us that pupil behavior should also be dismissed from consideration as a basis for evaluating teaching.

Assessment at Level III involves the same problem of morality as assessment at Level IV, as well as others. It seems to us that the ultimate responsibility of the teacher is to provide pupils with the *opportunity* to learn, not to "make" them learn. The competent teacher would be the one who could maximize the opportunity afforded each pupil under his care to learn what he needed to learn--the one who (1) could diagnose pupils' needs and capabilities, (2) prescribe appropriate learning activities, that is, those most likely to result in learning for each pupil, and (3) work with the pupil in such a way that he would be most likely to experience those activities. But these three activities are teacher behaviors and a part of Level II. Because Level III shares so many of the problems which Level IV has, it seems to us that neither of them is an appropriate level for assessment.

In summary, pupil outcomes are not a satisfactory basis for evaluating teaching. Although the use of relatively short time periods for evaluating teaching has been advocated, this procedure is of questionable value because the results of short teaching periods are not known to relate to those of longer periods and the material taught in a brief time is likely to be simple and factual rather than more complex or abstract. We do not know that teaching facts and more complex material require the same skills and there is some evidence that they do not. The results of short-term teaching units, therefore, risk being irrelevant or even misleading.

Nor does the use of year-long time periods solve this problem. The evidence determined by correlating the gains made by two classes taught by each of a series of teachers indicates that the data frcm about 20 classes of pupils would be required to reach the minimum standards of reliability usually required for making decisions about individuals.

Finally there are the problems of preventing the teacher who is to be evaluated from teaching the test and of the likelihood that the teacher will concentrate her efforts on teaching pupils near the criterion, if the number of pupils meeting some minimum standard is the measure of teacher competence. Either result would be too narrowly focused.

All in all, evaluating teaching by testing pupils seems unsatisfactory, or even damaging, and it is hard to see how modification could make it functional.

## Assessing Teacher Competence Based on Performance (Level II)

Since Level II assessment, assessment based on *measures of teacher performance*, seems at present to be the only viable course open to us, we shall proceed on the assumption that Level II assessments will be the principal basis for decision-making within a PBTE program. As we shall see, there is reason to believe that such measures can be practical, reliable, and objective enough to meet the assessment needs of such a program. Their adoption does, however, leave us with a clear responsibility for establishing the validity of the measures.

## Some Characteristics of Adequate Performance Measures

There are three distinct steps involved in obtaining an accurate measure of teacher performance: (1) A sample of the relevant behavior must be obtained; (2) a scorable record of the behavior must be made; and (3) the record must be quantified or scored.

In order to obtain a *relevant behavior sample*, we must put the teacher in a situation in which he has an opportunity to use the competency in question. Perhaps the best strategy is to put him in something very like the "teaching test" situation described under Level IV assessment, that is, give him a teaching task to perform, or a teaching problem to solve, which clearly requires the use of the competency to be assessed.

Since the assessment is to be based, not on pupil outcomes but on whether the candidate follows the best known practice--uses the strategy our best knowledge identifies as optimal--there is considerably more latitude permissible in setting up the test situation than would be permissible if we were attempting Level IV assessment. Role-playing, micro-teaching, and other forms of simulation may sometimes be appropriately employed.

The problem of obtaining an accurate scorable record of the performance may be attacked with the aid of the extensive experience gained in research (done mostly during the last two decades) using recently developed techniques for coding classroom behavior on the basis of direct observation. (Cf., for example: Medley and Mitzel, 1963; Simon and Boyer, 1967, 1970; Boyer, Simon, and Karafin, 1973.) While it is not likely that these techniques can be adapted to measure all of the competencies we need to assess, the methods used can be adapted to the construction of instruments that will.

*Scorability of the behavior records* depends to a great extent on the amount of inferential judgment required of the observer in making the record. The typical rating, which is used in too many PBTE programs today, is not scorable in the sense we mean because it does not yield a behavior record at all. Ratings record judgments or evaluations for which the relevant behavior has only been registered mentally. Thus, the rater observes, remembers occurrences which seem relevant to him, combines them in some unspecified way to form a composite picture, and forms an evaluation based on his own conception of good teaching or whatever measure is being rated. However, only the evaluation is recorded, not the behavior. The idiosyncrasies, the subjectivity, the biases, and the errors of judgment of the raters are interposed between the behaviors to be assessed and the evaluation which is recorded.

The crucial and difficult task too often neglected is that of *specifying the competency in question in behavioral terms*--in terms of what the teacher does and how often (with the necessary contingencies also specified) rather than in terms of how "well" he performs or how "appropriate" his behavior is. We shall have more to say later about how to go about the specification task. Once it is completed, the development of the procedure for observing and recording a sample of behavior is greatly facilitated. (The reader is referred to the discussion of category and sign systems in Medley and Mitzel, *op. cit.*, pp. 298-305.)

*Scoring the behavior record* should be a mechanical procedure, that is, it should be possible to have the scoring done routinely by a clerk or a computer. Use should be made of mark-sensing recording forms or of the devices now available which make it possible for the recorder to record directly on magnetic tape.

We have tacitly assumed in the discussion above that the competency to be assessed is a skill manifest in classroom interaction because such competencies seem to give the most trouble. The same requirements apply to assessments of other competencies, of course, but tend to be easier to fulfill.

## Establishing Validity of Level II Measures

As we have pointed out, the teacher education enterprise must ultimately be defended on the grounds that it somehow results in more pupil learning in the schools. We must validate teacher education by showing that the lines of influence in Figure 1 exist, that is, that teachers' training experiences (Level I) may be expected to affect pupil outcomes (Level IV).

Figure 2 provides a basis for discussing strategies for establishing the existence of these lines of influence. The strategies proposed are represented by the dotted arrows in the figure. Let us first define them from the standpoint of the research worker.

Research attempting to establish empirically the existence of relationships between teacher behavior (Level II) and pupil outcomes (Level IV) may be called *research in teacher effectiveness*. Research attempting to relate pupil behaviors in school, or learning experiences (Level III), to outcomes (Level IV), may be called *research in classroom learning*. Research attempting to relate teachers' training experiences (Level I) to their teaching behavior (Level II) may be referred to as *training research*.

These three types of research may all be defended as viable and useful strategies. A fourth type which has been proposed in the past would attempt to relate training experiences (Level I) to pupil outcomes (Level IV), and may be called *research in teacher education*--perhaps. This strategy has been advocated as a means of program validation and perhaps for non-PBTE programs it is all we have. But when assessments in the program are based on teacher performance (Level II), *validation of a performance-based program and research in teacher effectiveness are identical processes*.

Or, to put it differently, if a PBTE program is defined in terms of Level II performance competencies of its graduates, validation of those competencies is *de facto* validation of the program. Research in performance-based teacher education then becomes a two-step process. *Training research, which relates Level I (training) to Level II (performance), becomes the same thing as program evaluation*. Program validation as defined here becomes exportable--and importable; and all the literature on research in teacher effectiveness--for what it is worth (see Morsh and Wilder, 1954; Rosenshine, 1971; Rosenshine and Furst, 1971, 1973; Dunkin and Biddle, 1974) becomes relevant. And whatever efforts at program validation are carried on in the local program augment the knowledge hitherto developed only as research in teacher effectiveness. A false distinction disappears and we have in the making a true symbiotic relationship between the researcher and the program evaluator. No longer may the researcher look down on the evaluator nor need the evaluator take a back seat at AERA meetings. We can hear the teacher educator declare: "We have met the researcher, and he is us!"

Assuming that the reader who is still with us agrees with our conclusion that Level II assessment, assessment of teacher performance, is the central concern both of evaluation and research in PBTE, we now propose to discuss the two practical problems whose solutions are critical

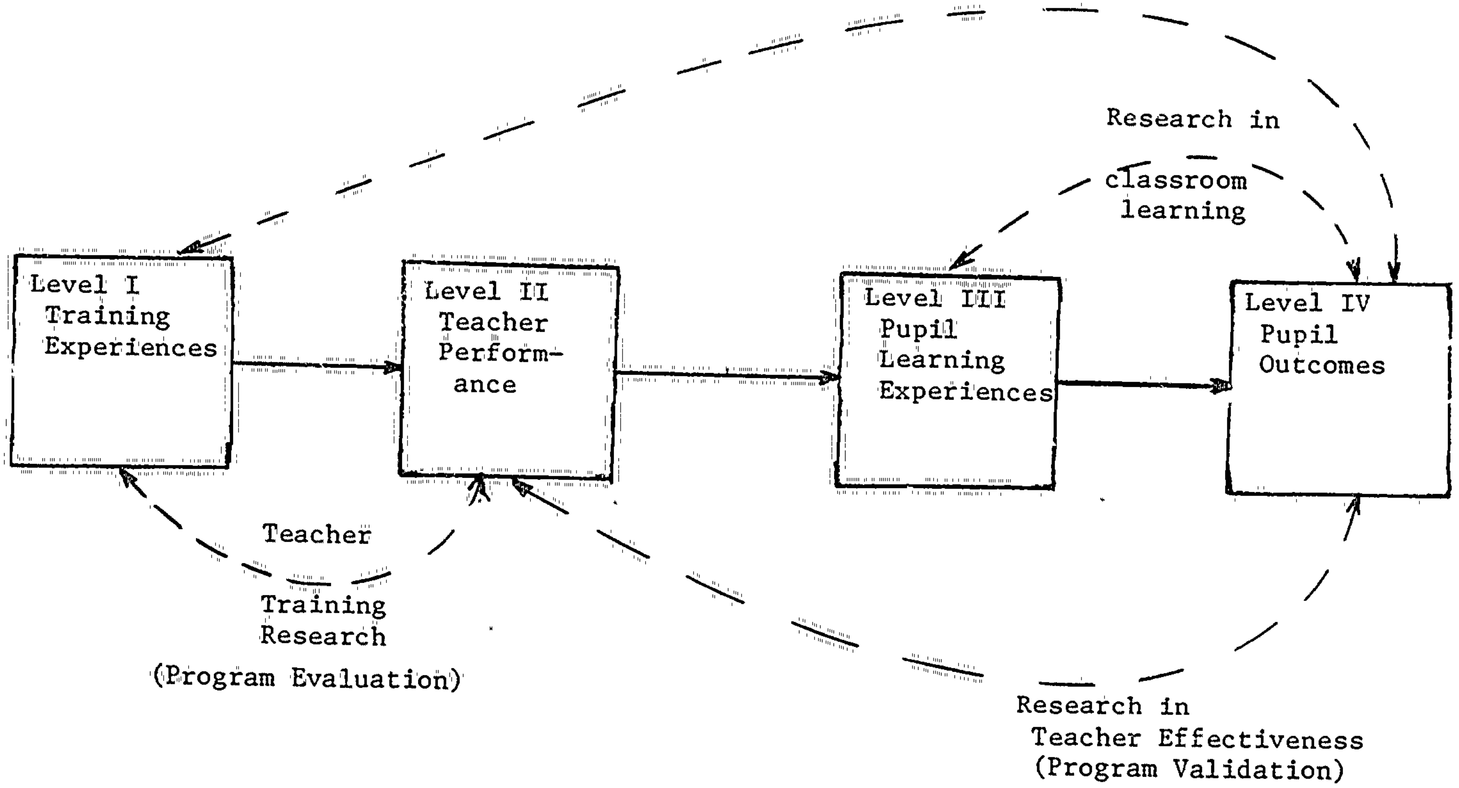


Figure 2 Research and Assessment in Teacher Education

to the success of a PBTE program: The problem of *specifying the set of competencies* the acquisition of which is the enabling goal of any given program and that of *validating the program* by showing that the development of these competencies will make a teacher more effective in promoting pupil learning. The discussion will draw heavily on our own experiences, in the first instance in working with program personnel attempting to specify a set of competencies and in the second instance in doing research in teacher effectiveness.

## Suggestions for Specifying Competencies in Assessable Terms

Specifying competencies in behavioral, low-inference terms is an early task and a central one in the development of a PBTE program. Admittedly, it is a difficult one, but it is necessary if the nature of the competency is to be identified with such precision that an observer or a teacher can know whether the competency in question has been demonstrated. It is also very useful in helping decide what a curriculum, a course, or a training module should contain.

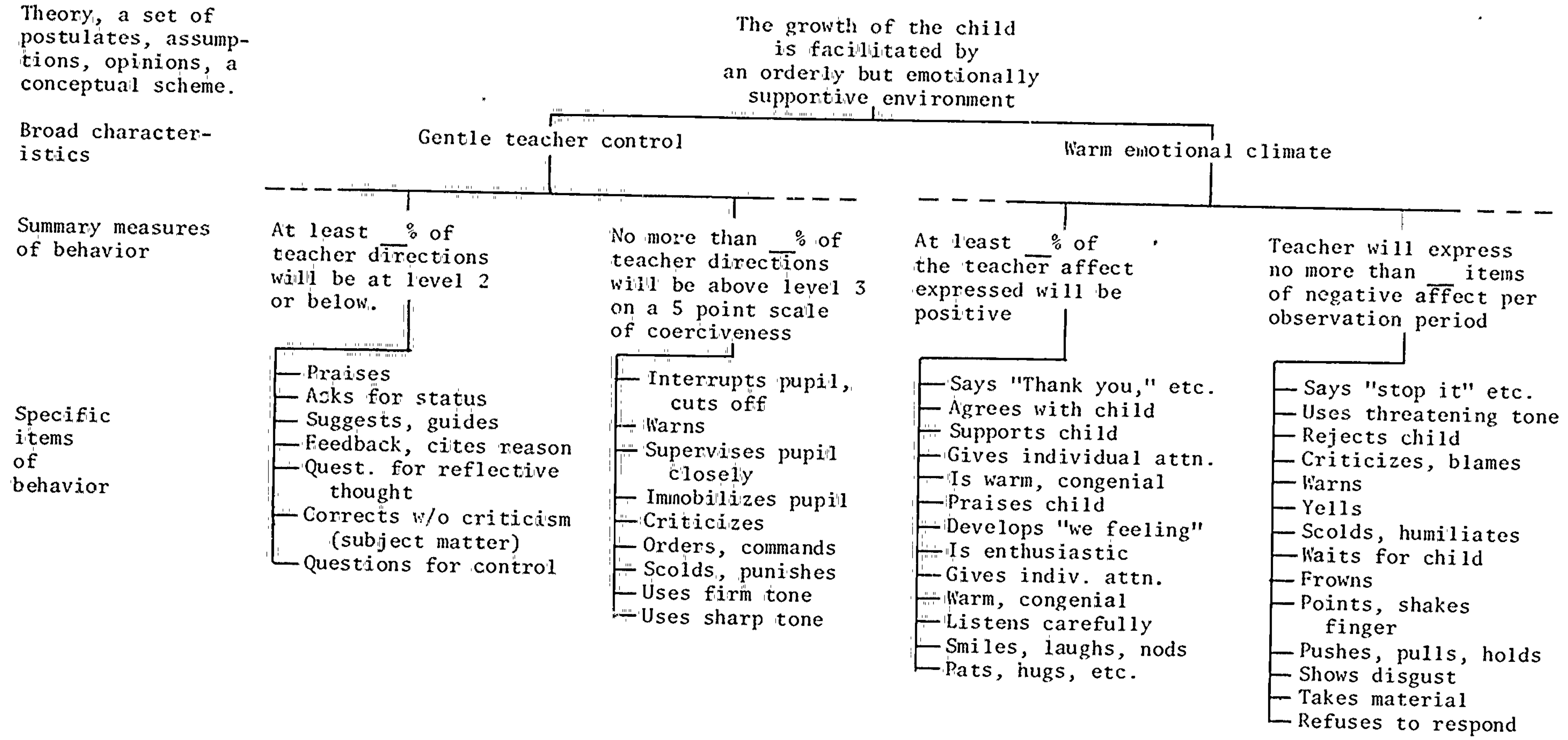
In our experience in helping groups with the task of specifying the competencies to be developed in a program, we have encountered a major communication problem that has seemed quite widespread. Competency definers seem to fall into two groups. One group tends to produce a set of competencies that is long and specific; the other tends to produce a set of competencies that is brief but abstract. Members of neither group seem to like or even to understand what the other produces and when either group approaches the task of designing measures of its own competencies, it runs into problems.

Members of the first group find that they need something like 657 measures--one for each competency--and face an instrument construction task that is to all intents and purposes impossible to accomplish. Members of the second group find it almost impossible to define ways of assessing the competencies that are public or even relatively objective, but must fall back on broad, general rating scales.

### A Hierarchical Organization

One way to simplify both the task of specifying competencies behaviorally and that of developing measures of them is to attempt to arrange both general and specific descriptions of behavior in a common hierarchical structure so that the set of more closely specified behaviors is seen as part of each broadly defined competency. Figure 3 is a simple example of what we mean.

The figure is not meant to represent criteria recommended for use; rather, it is an example of how one might go about organizing such criteria. It was developed at first to help members of the two groups get together to verify that we were talking about different levels of complexity or abstraction for what were really the same objectives. What the figure represents is a way of organizing competencies that seems to be useful. The first row across the top of the chart would represent the highest level of abstraction--relatively broad statements about what is



Theory, a set of postulates, assumptions, opinions, a conceptual scheme.

The growth of the child is facilitated by an orderly but emotionally supportive environment

Broad characteristics

Gentle teacher control

Warm emotional climate

Summary measures of behavior

At least ___% of teacher directions will be at level 2 or below.

No more than ___% of teacher directions will be above level 3 on a 5 point scale of coerciveness

At least ___% of the teacher affect expressed will be positive

Teacher will express no more than ___ items of negative affect per observation period

Specific items of behavior

- Praises
- Acks for status
- Suggests, guides
- Feedback, cites reason
- Quest. for reflective thought
- Corrects w/o criticism (subject matter)
- Questions for control

- Interrupts pupil, cuts off
- Warns
- Supervises pupil closely
- Immobilizes pupil
- Criticizes
- Orders, commands
- Scolds, punishes
- Uses firm tone
- Uses sharp tone

- Says "Thank you," etc.
- Agrees with child
- Supports child
- Gives individual attn.
- Is warm, congenial
- Praises child
- Develops "we feeling"
- Is enthusiastic
- Gives indiv. attn.
- Warm, congenial
- Listens carefully
- Smiles, laughs, nods
- Pats, hugs, etc.

- Says "stop it" etc.
- Uses threatening tone
- Rejects child
- Criticizes, blames
- Warns
- Yells
- Scolds, humiliates
- Waits for child
- Frowns
- Points, shakes finger
- Pushes, pulls, holds
- Shows disgust
- Takes material
- Refuses to respond

Figure 3. An Example of Hierarchy in Educational Thought and Practice

Note: Items defined in the Florida Climate and Control System: Observer's Manual important for a teacher to be, do, or accomplish, statements on which most people would agree. One example is the statement shown: "The growth of the child is facilitated by an orderly but an emotionally supportive environment."

Disagreement with that statement as an objective is not likely. Perhaps this reflects the fact that it does not really have enough concrete meaning so that you could either agree or disagree with it at this level of abstraction. You can get agreement on almost anything if you go high enough up the abstraction ladder; if you become sufficiently abstract, you are using words that mean different things to different people and everyone is agreeing with his own meaning. Just the same, it is useful, almost necessary, to begin with such a common base.

In the next row of the figure, in which we begin to break the broad statements down, we begin to face differences within the group which must be recognized and dealt with. It is here that differences must be reconciled before the program gets off the ground.

Returning to the example, we have listed two clear aspects of the initial statement on which we will assume agreement has been reached. One is that the order must be obtained by gentle or noncoercive means if the environment is to remain supportive. The other is that there must be a warm emotional climate. These two statements are listed on the second level of the hierarchy, called Broad Characteristics. Although more concrete than the statement above them, these behaviors are still not objectively measurable. Observers will still differ, for instance, in how warm is "warm." These behaviors are recognized to be characteristics of desirable classroom behavior, but they are still not specified in enough detail to be objectively assessable.

At the third level, we begin to identify what may be called summary measures of behavior that are recognized as components or aspects of the broader characteristics. Each summary measure is made up in turn of specific items of behavior and each behavior item is defined with the care necessary for inclusion in a manual of instruction for classroom observers. Items must be specified in enough detail so that an observer can be trained to recognize reliably each item of behavior that he sees demonstrated during a period of observation and so that the teacher himself knows whether he has exhibited any one of them. This is the fourth level of the hierarchy, the most specific; it provides the basis for obtaining a scorable record of behavior.

As an example at this most specific level, the item "Suggests, Guides" includes such teacher statements as "How about putting it over there, Jimmy, OK?" "I wonder if you would shut the door for us, Bill." "Bob, would you mind moving so John can sit down?" These statements are suggestions for change in behavior that have the characteristic of being "softened" by a "please" or "OK?", or by being phrased as questions (Soar, Soar and Ragosta, 1971).

Another item at the gentlest level of control, "Feedback, Cites Reason", is coded when the teacher gives information which implies a change in behavior without directly asking for it. For example, the teacher comment "I'm having trouble hearing" is not a direction, but it does give information that implies and probably produces a change in pupil behavior. An instance was observed in a first-grade classroom during the first week of class. The number one and number two trouble-makers (two boys who could be so identified in five minutes) were together in the back row of a small group, jostling, nudging, and pinching. The teacher looked at one of them and said, "John, I think there's room for you up here," gestured toward a spot by her feet, and waited. In two or three seconds, John came to sit by the teacher's feet. The other pupils did not seem to see it as a put-down; but it did create a big gap between the two trouble-makers and the problem was solved.

At an intermediate level of control (not shown in the figure) would be statements such as: "OK, anyone who wants to go to the bathroom, get in line," which is not very coercive, but clearly a direction with a reason. Other examples would be "Get out your arithmetic books and open them to page 27."; "When you've finished, put your papers on my desk."

At the coercive end of the scale would be a statement like, "Jimmy, stop that!" which would be coded "Orders, Commands," and probably "Sharp Tone" as well.

Using a hierarchy like the one described above let us talk about particular aspects of a rather vague concept like "orderly but emotionally supportive environment" in terms of a series of fairly objective statements about teacher behavior. Because each specific behavior is defined with sufficient care so that agreement between observers can be made acceptably high, we can reliably assess this aspect of behavior, and a teacher can know whether his behavior meets the requirements of the competency.

It is not our purpose to initiate a semantic argument over whether competencies are narrow or broad, and we would propose that varying degrees of generality are useful for different purposes, but we suggest that the term competency be used at the level of the teacher's ability to integrate the specific behaviors necessary to produce one of the Broad Characteristics of the classroom. The Summary Measures and Specific Items of Behavior would define the competency, give it behavioral meaning, and thus make it assessable.

Moving Up and Down the Hierarchy. No less important is the aid to communication provided by the ability to move up and down this abstraction ladder at will. The person who, when asked to name a competency, gives a theoretical statement or a broad, high-inference label for a classroom behavior can be encouraged to move lower in the hierarchy by being asked questions like, "What would this behavior look like if you saw it happen?" "What would a teacher who is doing this do that one who is not doing it would not do?--what kind of behaviors would differentiate them?" Questions such as these encourage people to be more explicit about what they mean and at the same time generate the statements lower in the hierarchy needed to make objective measurement possible.

The person who, when asked to define a competency, makes a statement like, "The teacher should avoid direct commands and orders," can be encouraged to move higher in the hierarchy by questions like, "Why would you care about that?" "Why is that important?" "How does it make a teacher more effective?" What typically happens is that the person moves up the scale to give a broader conception of how these behaviors relate to reality and to his scheme of values which puts the competency in larger perspective. And at the same time he suggests how the items should be combined into composites or clusters that are internally consistent.

Thus there are two advantages to this sort of hierarchical arrangement: (1) communication is clarified and agreement on the competencies to be adopted as program goals is facilitated, because the generally valued broad competency has operational meaning given to it by the items which it includes. And (2) at the same time, the items for an objective measuring instrument of each broad competency are specified so that the competency can be measured.

Figure 3 illustrates a part of one of many possible hierarchies that might be developed for assessing teacher performance. Others might relate to the nature and frequency of cognitive questioning, the manner of structuring learning activities, the use of experimental techniques--any of the broad goals for which modules and programs might be developed. But the important point is that the behaviors relevant to whatever broad goal or concept is specified have been identified and defined and the performance has been made assessable.

The Empirical Test. There is a critical issue which arises whenever we combine items on *a priori* grounds into clusters intended to represent broad competencies. How can we be sure that the specific items of behavior belong together in the real world as well as in the world of theory? Usually, the items will have been selected as developed to measure behaviors which are believed to represent a single aspect of good teaching. This procedure is similar to the way a series of items on an achievement test is selected to sample knowledge in a homogeneous subject-matter area. But it is critical to establish empirically that the items do belong together, that the cluster is internally consistent, because it is quite likely that some items that are believed to go together in the conceptual scheme will not hang together empirically. We do not know that much about the dynamics of teaching yet.

This is what happened when we tried to assess gentle teacher control (Soar, 1973): In addition to the verbal items described earlier, there was also a smaller set of nonverbal items grouped with the verbal items under the assumption that they both fell on a single dimension extending from gentle to harsh teacher control. However, factor analysis indicated that this assumption was not so. The verbal gentle control items did hang together, but most of the nonverbal gentle control items fell into a different subset which was fairly consistent internally, but not closely related to the verbal scale. We had oversimplified the area.

Interestingly enough, there was some crossover. Teacher smiling, which we classed as a nonverbal behavior, belonged in the "verbal" composite instead of the "nonverbal" one. Probably the reason is that people smile as they talk, using their faces as an additional source of stimulus or feedback. Other nonverbal behaviors, such as "Touches" or "Gestures" tended to occur more independently of verbal behavior. This made sense after the fact, but these results were not anticipated. It is not at all unusual for the facts to contradict our best theories in this way.

After we had discovered this clustering in the analysis of the data, we remembered a teacher and an aide in a first-grade classroom. The teacher often smiled at pupils, praised them, and was very warm and supportive, but she never touched a child. Her aide, on the other hand, seldom smiled or praised a child, but she rarely passed one without ruffling his hair or giving him a pat. She almost never sat down without a child on her lap and very often had one on each knee. We thought at the time that it was a nice example of differentiated staffing! But we see it now as an example of the behaviors fitting together the way the empirical analysis says they do.

The important point to learn from these examples is that if behaviors are put together that are not at least moderately intercorrelated, whatever meaning they are supposed to have is destroyed, and with it the discriminating power of the measuring instrument. On our original scale, this teacher and her aid, different though they were, would probably have had similar scores on our composite--scores somewhere in the middle. That this kind of empirical check is critical whenever items are combined should be obvious from this example.*

Importing Past Research

In specifying competencies and developing measures for them it seems important to pursue leads from past research on teacher effectiveness, not only for the obvious reason that the research may suggest dimensions of behavior which might otherwise be neglected, but also because it may suggest specific items of behavior that reflect those dimensions. This is really an extension of the concept of "importation" of teacher effectiveness research for program validation discussed earlier--its extension to the initial selection of competencies and program development. It would be wasteful for programs not to make this kind of use of present knowledge developed in teacher effectiveness research, as well as using such knowledge as it becomes available from program validation studies.

A number of reviews of this literature have been cited earlier: Morsh and Wilder, 1954; Rosenshine, 1971; Rosenshine and Furst, 1971, 1973; and Dunkin and Biddle, 1974 Joyce (1974) and Kay (1975) have also discussed the various possible conceptual bases from which competency lists may be derived.

*There are two ways of conducting such an empirical check. One is to do an internal consistency item analysis of each composite and "purify" it by eliminating nonconsistent items. The other approach is to factor analyze the entire set of measures and see whether the factor structure that results corresponds to the *a priori* structure. Both methods have advantages and disadvantages. We would offer one suggestion: if you use factor analysis, view the results with caution. The sample sizes available in this kind of study are usually smaller than the minimum recommended by experts and numbers of influences can affect the results.

Several concepts from our research seem to have implications for competency specification. One such concept is the usefulness of distinguishing three areas within which the teacher exercises control that are often confused (Soar and Soar, 1973): (1) control of the behavior of pupils, (2) control of choice of subject matter, and (3) control of the thinking processes which the pupils use. Our data indicate that when these three types of control are distinguished, their effects on pupil growth are found to differ. Some data (collected in grades three to six) support the idea that the teacher's control of behavior may have quite different consequences for pupils than the teacher's control of thought processes. A factor called *Indirectness vs. Silence and Confusion* appeared to represent a "freeing" yet orderly style of teacher-pupil interaction, and showed a significant positive relationship with pupil gain in creativity. Another factor, called *Freedom of Physical Movement*, however, showed a significant negative relationship with creativity gain (Soar, 1966, p. 178).

Subjectively, these differences in teacher control can be recognized in occasional classrooms. In a second-grade classroom, there was close control of both subject matter and pupil behavior (the teacher had taken over an unruly class in the middle of the year). There was little talking between children and pupil movement was brief and task-related, with teacher and pupil talk so quiet it was hard to hear half-way across the room, even during recitation. Yet the teacher did not restrict thought processes; rather, she supported complex thinking by pupils. In reading groups, her questions were divergent, encouraging pupils to infer meanings and motives; in arithmetic, she sought alternative ways of solving problems; among pupil reports of a field trip, she valued a poetic description as highly as a reportorial one. The teacher closely controlled pupil behavior and allowed no choice of subject matter, but worked hard at freeing pupil thinking.

In a kindergarten classroom a different combination of controls existed. The teacher required pupils to choose an activity from the many materials set out in interest centers and refused even to suggest alternatives when asked. When she was asked a question which was subject matter related, she was likely to answer with a question. Yet it was clear that there were well-established rules of behavior such as "No running", "No loud talking", "Hands to yourself", and "Don't interfere with other people's work." In this classroom, choice of subject-matter and thought processes were free, but behavior was controlled.

Although these distinctions seem reasonable, once proposed, the data indicate that most teachers do not make the distinctions as they manage behavior and teach.

A different concept is an empirically derived distinction between "structure" and "control", two terms which may seem at first to be closely related. In the sense intended here, structure represents the set of standard operating procedures which the teacher and pupils understand in common, such as the sequence of activities which is followed daily, and the limits of behavior which pupils understand and accept. In contrast, control is made up of the moment-to-moment, face-to-face interactions between teacher and pupils intended to modify the behavior of pupils. The data suggested that teachers who provide the least structure in their classrooms may feel the need for more controlling behaviors than ones who provide more structure (Soar and Soar, 1972).

This distinction was later observed clearly in a classroom that was part of a program in which classrooms were intended to be "open" and in which the teacher seemed to feel obligated to give pupils greater freedom than she was comfortable with; as a result, the activity and noise level in the classroom continued to build. After a while, the noise reached a threshold at which the teacher apparently could not bear it any longer and she stepped in firmly (not critically, but firmly), brought things to a grinding halt, and reestablished quiet. As soon as she had done this, she withdrew and, after a bit, the cycle repeated itself again and again. It seems likely that the problem arose because the teacher, and perhaps the program she represented, did not recognize the need for more structure and less control--more structure would decrease the need for such frequent direct intervention by the teacher. Further, if the teacher had recognized the distinctions among the three areas, it seems likely that she could have structured limits for behavior while freeing choice of subject matter and thought processes.

The data indicated that this alternation of close control and the absence of control was likely to be destructive in terms of pupil outcomes. For pupils to be in a setting in which most of the time it was all right to do almost anything, then all of a sudden it was not all right, one in which they never really knew which set of rules applied or when the rules would change, was not supportive of learning.

A third concept from past research which might be considered is the concept of "wait-time" reported by Rowe (1974). She defines two kinds of "wait-time": (1) the length of time the teacher waits for a pupil to answer a question before she intervenes and (2) the length of time the teacher waits after a pupil response or statement before she reacts. There seems to be a rather clear threshold in each case and waiting this minimum time period seems to be associated with a number of desirable changes in pupil activity. Pupils interact more, the number of inferences and conditional statements increases, pupils suggest ideas with greater apparent confidence, the number of appropriate responses increases, and greater numbers of experimental tests of hypotheses are proposed, compared with classes where wait-time is below the threshold.

These are only a few examples of dimensions of behavior suggested in the teacher effectiveness literature which are probably not widely known, but which should be considered in specifying competencies. They illustrate the value of importing knowledge from teacher effectiveness research into program development.

## Program Validation

### The Need for Program Validation

We are all concerned by the fact that we are spending so many millions of dollars and so much effort in the development and implementation of programs, with so little knowledge of what kinds of teacher behaviors really are associated with increased growth of pupils. The possibility that all this effort may be wasted is a real one. A critic of educational research observed some years ago that virtually all the research on teacher effectiveness could be summarized by one general conclusion--*nothing makes any difference.* We are not as pessimistic as that, but must agree that we do not know very much and that some portion of what we "know" may not be true.

There is no alternative to moving ahead with building today's programs on the basis of what we know today. We cannot wait for the researchers to answer all the questions because they are not going to answer them any time soon. Those of us who call ourselves "the researchers" are probably surer of this than anyone else. But in a situation in which we so often find that the things we thought we knew are not true, there is a real and present danger in building a complex program without beginning as soon as possible to find out whether the behaviors we "know" will produce desired changes in pupils will, in fact, do that. As hard as it has been in the history of educational research to show that anything teachers do makes a difference for pupils, it seems critical to begin early in this whole sequence to find out whether what we are going to considerable time, trouble, and expense to train teachers to do makes a difference to the pupils these teachers ultimately teach. The public and the legislators will want to know and they have a right to know. They are going to want evidence that what our teachers do leads to more growth by pupils and that we, in our programs, can produce the kinds of teacher behaviors which in turn produce more growth in pupils.

It is critical to keep the two steps of program evaluation and validation separate for the reasons enumerated earlier in this paper; but both steps must be accomplished, and now.

It would be reasonable to ask, at this point, "Why is it undesirable to use pupil measures to evaluate the teacher, but necessary to use them to validate teacher behaviors?" The two activities have a number of basic differences. In program validation, pupil gain measures are not needed for every teacher in every program. Such measures are needed only for samples large enough to enable us to learn and verify relationships between teacher behavior and pupil gain. The considerable amounts of error in the measurement process do not spuriously injure or benefit individual teachers; they only weaken the relationships found between measures. Greater time and resources may be given to collecting measures and analyzing relations between them in program validation, which needs to be done only once, than in evaluating teachers which must be done over and over again. These differences make program validation by pupil outcomes feasible, even though teacher evaluation by pupil outcomes is not.

## Some Examples of Weakness of Current Theory

Although program development must proceed now, it seems important to recognize the weakness of the theory on which it is based.

Perhaps the two innovations most often currently advocated are behavior modification (or contingency management, or behavior analysis, depending on the label used), on the one hand, and the movement for open classrooms on the other. Both claim support from theory and research. But if each of these is taken as a complete educational program in itself, (and that seems to be what is advocated) how could both be right when the programs are as different in character as they are? Some disquieting questions arise about how much help current knowledge is in formulating effective programs and reinforce the need for program validation studies whose outcomes can be fed back into program revision.

There is some research which relates to this question. We now have four sets of data which raise questions about the usefulness to pupils of extreme amounts of teacher control--either high (as in behavior modification) or low (as in open classrooms). Figures 4, 5, and 6 present data from different pupil groups and different observational measures. The behavior measures in all three have in common the fact that they represent teacher control in some way, primarily control of thought processes. Each curve represents a nonlinear relation which may be described as an inverted "U", since it indicates that, as teacher control increases from the minimum, pupil growth increases for the outcome measure used, but only up to a point. Beyond this point, increasing teacher control is associated with *decreasing* pupil growth. That is, there appears to be an optimal amount of teacher control for a given growth measure, which is neither the most nor the least control in most cases. Parenthetically, not only achievement gain, but also self-concept and several personality measures show this tendency.

Relationships such as these should be considered in specifying competencies.

Another illustration of the weakness of current theory is the emphasis given to increasing the cognitive level of teacher questions. Yet Taba, Levine and Elzey (1964) concluded that unless the teacher first spent sufficient time with pupils at the lower cognitive levels, the pupils were unable to sustain higher level thinking. Dunkin and Biddle (1974, p. 243) review a study by Rogers and Davis showing "a significant negative relation between higher-level questioning and pupil performance on test items of the analysis type." Three sets of our data also indicate that too much of the interaction in the classroom can be at too high a cognitive level (Soar and Soar, 1972, 1973). Several dimensions of classroom interaction were scored which represented the frequency with which relatively abstract interaction took place between teacher and pupils, following Bloom's Taxonomy of the Cognitive Domain and a Deweyian approach to teaching. These measures tended to be negatively associated with gain in both pupil achievement and self-concept. In some cases, the negative relationships held for the total pupil group, but, in other cases, it appeared to be true for disadvantaged pupils but not for advantaged pupils.

This is not to conclude that a teacher should interact only at the lower levels, but it does imply that it is possible for a teacher to interact too often at too high a level. The need for a "match" between where the pupil is and where the teacher is seems obvious, but how often do theoreticians raise caution about teachers working at too high cognitive levels?

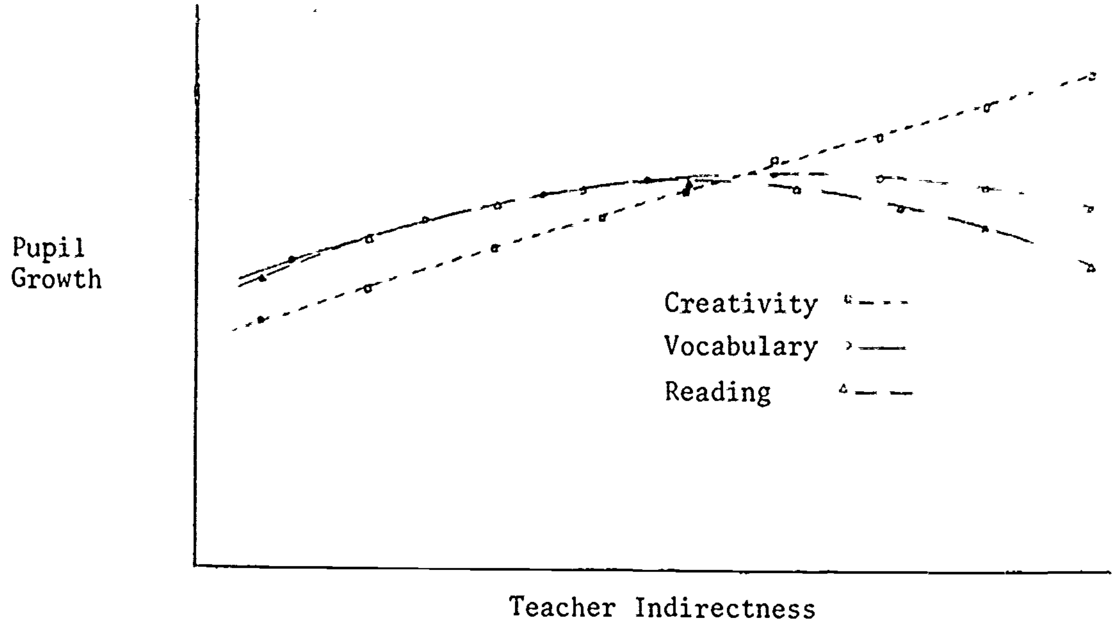


Figure 4.
Teacher Indirectness Related to Pupil Growth

For 55 Classrooms, Grades 3-6
(After Soar, 1968)

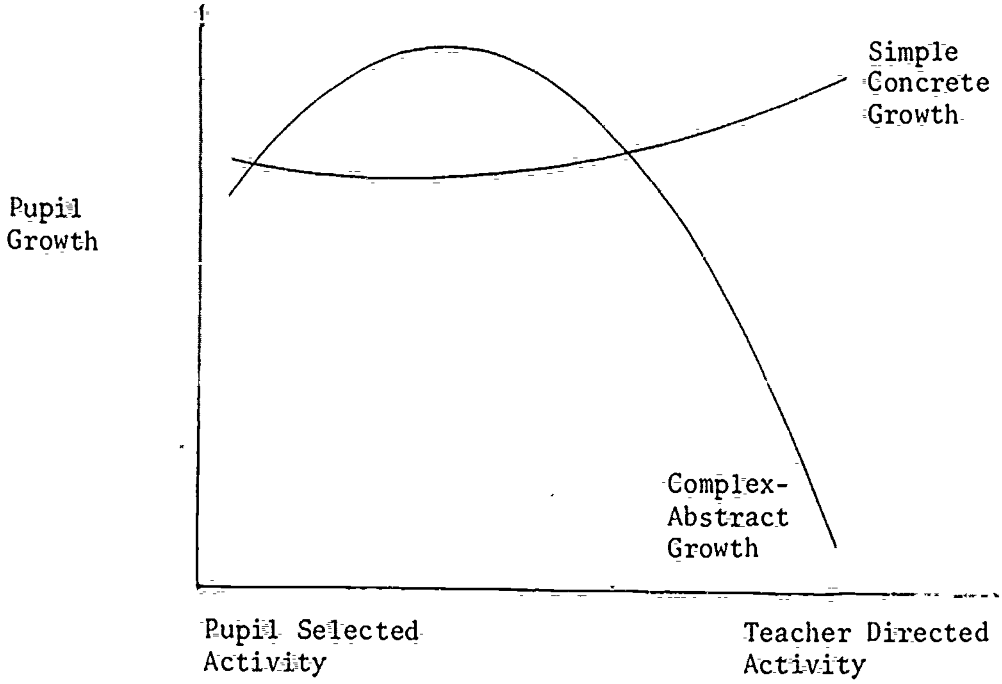


Figure 5.
Relation Between A Teacher Practices Observation Record Control Factor and Pupil Growth for 20 Follow Through First Grades

(After Soar and Soar, 1972)

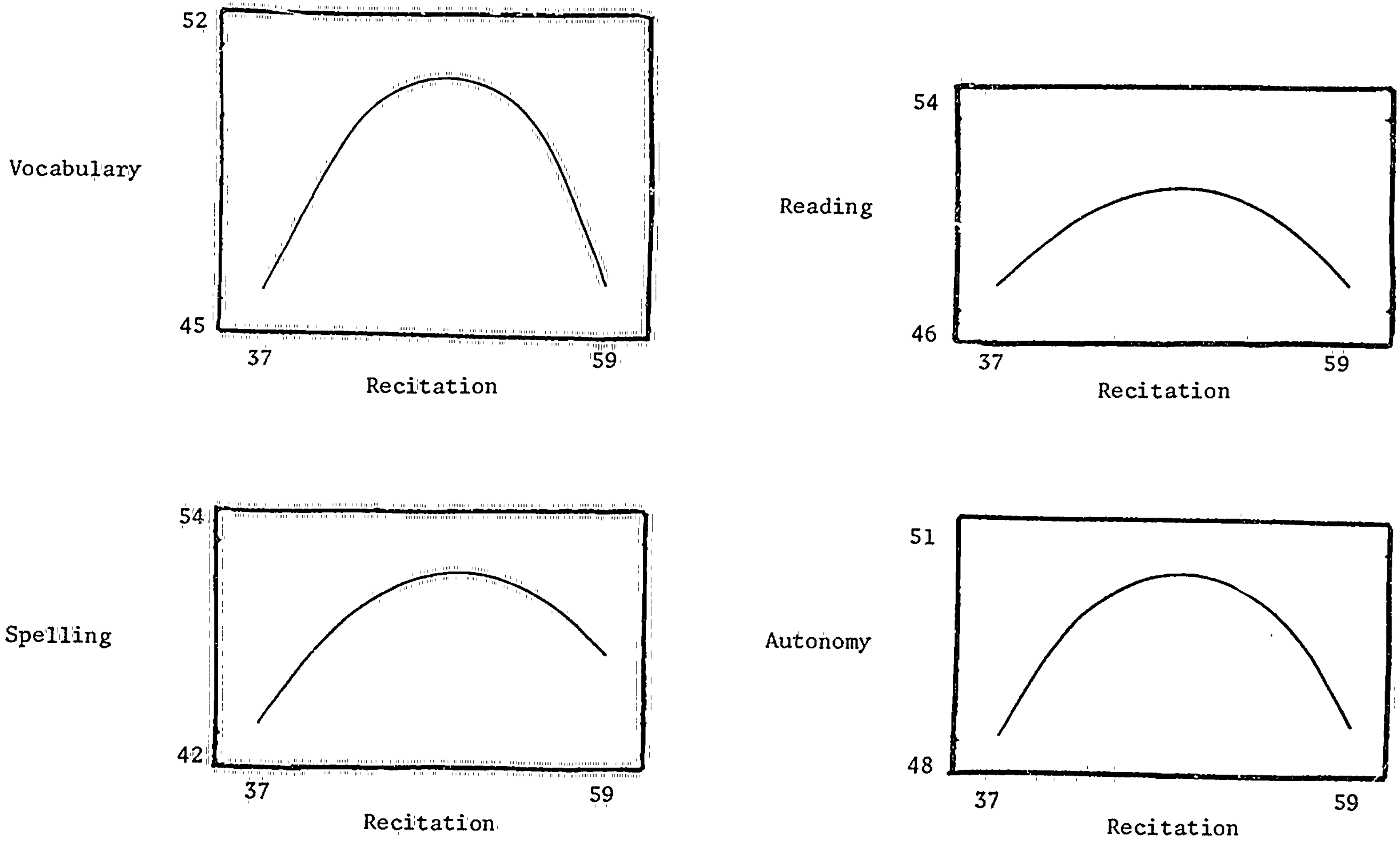


Figure 6
Relations Between Recitation and Four Measures of Pupil Gain
for 59 Fifth Grade Classrooms
(After Soar and Soar, 1973)

Findings such as these raise questions about the soundness of the theory on which recommendations for teacher behavior (and probably specification of competencies) are based and emphasize the usefulness of importing findings from teacher effectiveness research. There seem to be two aspects to the problem of defining a competency: what is the nature of the behavior and how much of it is desirable?

Another example of the inadequacy of current knowledge is our occasional ignorance even about which variables we should study. We stumbled across one such variable by accident. We had set up an analysis for looking at school year gain and wanted next to change the focus of the analysis to look at pupil gain the following summer. This change would have been easy except that we found we were one variable short on our data cards, so we put in what we thought was an extraneous variable as a "placeholder"--one which represented the number of pupils on which the mean for each classroom was based. You can guess what happened--it turned out to be a moderately powerful predictor of gain in that analysis. After the fact, this made sense. The variable represented the number of pupils who were present through three days of testing on three different occasions. The examiners had gone back for three make-up periods each time, but there were still a number of pupils we lost.

Our guess was that if the pupil came from a home in which school was valued, he did not have any choice; he was there. But if there was not much concern at home, he was able to drop out as the testing went on and consequently he dropped out of our data set. So probably this variable was an unobtrusive measure of attitude toward education in the home. This interpretation is somewhat uncertain because attendance is often assumed to represent the pupil's attitude toward teacher and classroom. But analyses of the data suggested that pupil attendance as well as "survival rate" related to gain over the summer but not during the school year; and it seems likely that family influences are stronger during the summer than the school year, whereas, if attendance had reflected pupil attitude toward the classroom, it would have been likely to show an effect during the school year.

In summary, these are only a few examples of the weakness of current theory as a basis for specifying competencies and developing programs. There is no question of the need to proceed with program development now, but these examples emphasize the need to use the empirical knowledge which does exist in teacher effectiveness research and to feed back new knowledge into program modification as rapidly as possible.

## Some Issues in Validation

Past research in teacher effectiveness suggests a number of issues which may be important in program validation. There appear to be two classes of issues--one dealing with the limits for a given validity relationship, or the terms under which it is true; the other with the statistical analyses.

Validity Specifications. The question is not, "Is the teacher behavior valid?" The question is instead, "For what is it valid?" For what kind of outcome, e.g., a complex one or a simple one? Unless the teacher behavior patterns which facilitate growth toward both kinds of objectives are the same, then the teacher must know how to produce both kinds of behavior and when. There is reason to suspect that one pattern involves open, accepting, clarifying, reflective behaviors while the other is based on more tightly structured, reinforcing behaviors. The behavior which is "valid for" (related to) one extreme of this range of outcomes may not be valid for the other extreme. Figures 4 and 5 illustrate this possibility.

We not only need to know what kind of objectives a teacher behavior is valid for, but we also need to know the kinds of pupils for whom it is valid. Current data indicate that this may be a critical question. The social status of the pupil, for instance, sometimes makes a difference in the kind of teacher behavior which is associated with most growth for him.

Figure 7 illustrates the relation found between pupil gain in reading and teacher control by means of coercion and negative affect, plotted

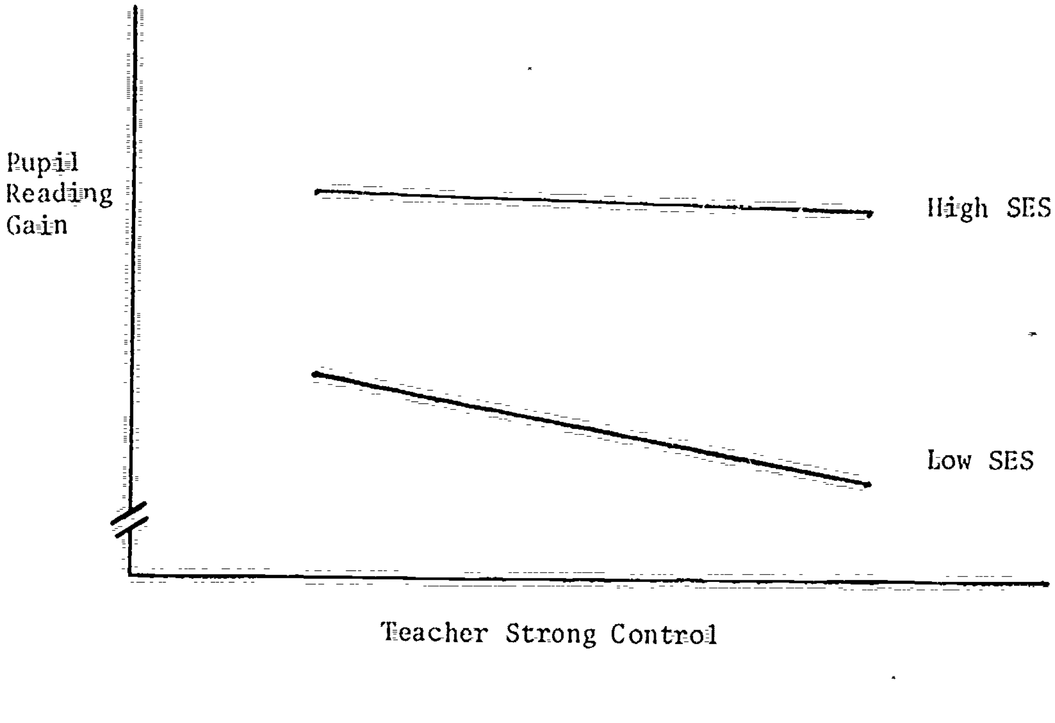


Figure 7
Teacher Strong Control in Relation to Pupil Gain in Reading by Socio-Economic Status

separately for groups of pupils of low and high socioeconomic status. For high status pupils, the measures were essentially unrelated; but for low status pupils, increases in negative teacher control were associated with decreases in pupil growth in reading (from data reported in Soar and Soar, 1973). While not strong, the difference in relationship for the two groups was statistically significant.

The fact that the greater *decrease* in gain occurred in the low social status group was opposite to expectations. We had expected that lower class pupils would have met negative affect more often, would have adapted to it, and that it would have had less impact on them; but these conditions were apparently not true. Instead, what may have happened was that the middle or upper class child was more likely to have support from home which compensated to a degree for an unfortunate classroom and which made him less dependent on it, but that the growth of the lower class pupil was more dependent on the nature of the classroom.

Parallel findings have been reporte⁴ by Brophy and Evertson (1974).

In addition to the questions of *valid for what* and *valid for whom*, we need to know, *valid for how long?* As we have indicated earlier in relation to multiple learning objectives, it may be that effective teaching for immediate learning may be different from effective teaching for long-term objectives. At any rate, it seems risky to assume that conclusions can be generalized from one to another without empirical evidence that the generalization is sound.

The problems cited so far represent a formidable challenge for program validation, but there is still another consideration which needs to be taken into account--the characteristics of the student teacher. We need to know what behaviors are a "best fit" for what kind of teacher, because probably not all kinds of behavior can be used effectively by all candidates in teacher education.

This degree of complexity in the nature of effective teacher behavior greatly increases the difficulty of validation; but if we fail to take it into account we risk producing in program validation the same inconsistent and often nonsignificant results which have been common in teacher effectiveness research.

The Need for Complex Analysis. Perhaps one of the reasons that research on teacher effectiveness has not been more productive in the past is that it has used an inappropriate model. The true nature of the relationship between teacher behavior and pupil growth appears to be very different from that implicit in most of the research which has been carried out. Most past research has sought a small number of large effects but it seems to us that an appropriate model would look for many small effects which are probably cumulative. If our hypothesis about large numbers of small effects is accepted, then the nature of the research changes considerably. We need to know many more things about the behavior of a teacher and the characteristics of pupils in order to identify the specific teacher behaviors which are effective for particular pupils.

Another defect in past research in teacher effectiveness has been the use of analyses which examined only linear relationships and assumed that more is always better. But a fairly common recent finding is that relations between classroom behavior and pupil growth are often nonlinear as illustrated in Figures 4, 5, and 6 and in similar findings reported by others (Solomon, Bezdek and Rosenberg, 1963; Coats as cited by Flanders, 1970; Brophy and Evertson, 1974).

This finding of nonlinearity seems intuitively reasonable. What does not seem reasonable is that we should have expected to find, without limit, many kinds of teacher behavior that increase pupil growth as they are increased. But whenever we calculate a linear correlation we implicitly assume that this is so.

Another aspect of the relationship between teacher behavior and pupil growth that is often ignored is the possibility that variables may interact in the statistical sense. The question in its simplest form is, "What is the simultaneous effect of two aspects of behavior? Is it different from the effect of each considered alone?" Because classroom behaviors do not occur in isolation, but rather occur in a context of other behaviors, it seems intuitively sound to assume that the effect of one may be moderated by the presence or absence of another.

An example of such a statistical interaction occurred in our research when a variable, the proportion of classroom activities in which the problem had been chosen by the teacher, was found by itself to be unrelated to pupil gain. Another variable, the overall amount of recitation, also showed no linear relation with pupil gain. But when recitation and teacher choice of problem were examined simultaneously, it was found that greater achievement gain took place in those classrooms where one or the other of these variables occurred with considerable frequency, but not both (See Figure 8). It did not seem to matter which one, but frequent occurrence of either one or the other was important. On the other hand, if both occurred with considerable frequency, or if neither did, pupils gained less. Similar effects were found for other pairs of variables.

It may be possible that this interaction can be related to the findings of nonlinearity for teacher control. If an intermediate amount of teacher control is associated with maximum pupil gain, then this intermediate amount of control can be produced in various ways: either by an intermediate amount of one dimension of classroom controlling behavior or by a combination of one kind of behavior which represents control and another kind of behavior which provides some freedom. Two controlling kinds of behavior at high levels result in too much control; neither being present results in too little control for pupil growth, just as in the inverted "U's" described in Figures 4, 5, and 6.

So far, the concept of statistical interaction has been discussed in relationship to the simultaneous effects of different combinations of classroom behaviors on pupil learning. Another example of the same concept would examine the interaction of entering pupil characteristics and classroom behavior because both are related to outcomes, as illustrated in Figure 7. This is the logic of the aptitude-treatment interaction (ATI) studies. The extent to which such interactions can be made use of

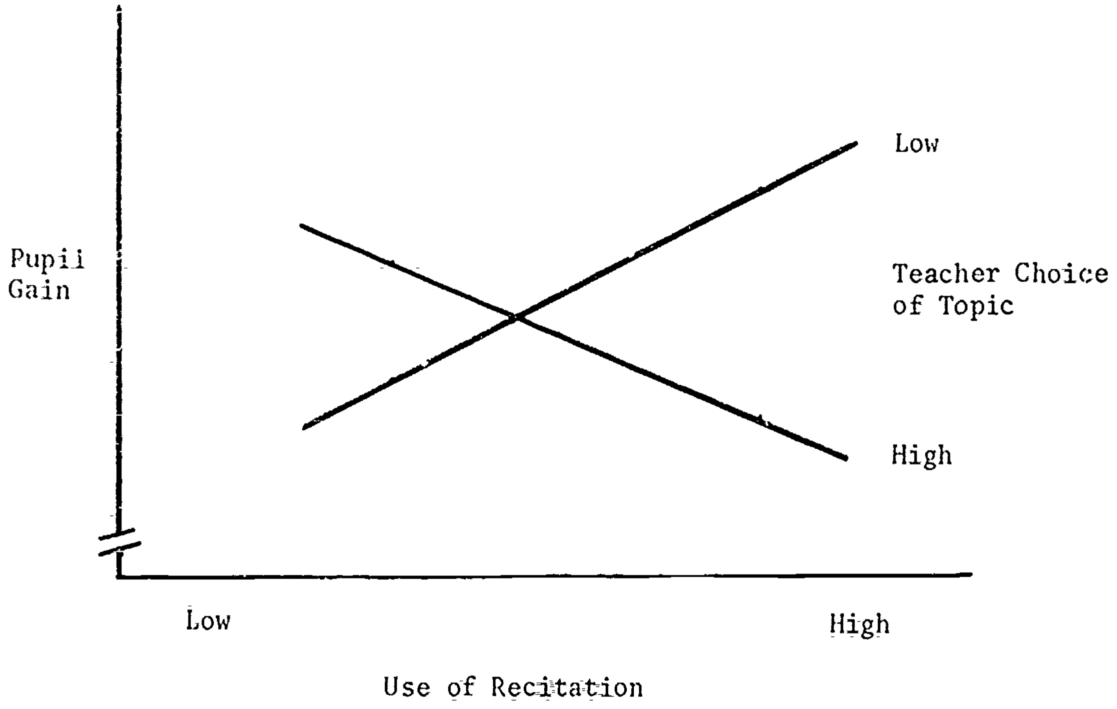


Figure 8
The Relation of Two Classroom Behavior Measures with Pupil Gain in Arithmetic Concepts

in teacher education or in program development may be limited; the task of teaching is complex enough without them. But it does seem important to take them into account in the process of program validation where they are more easily dealt with.

There are a number of pupil characteristics which seem to determine in part which teaching behavior is most functional in a given instance. For example, one of our findings indicated that pupils who were high in motivation grew more in classrooms in which they were frequently allowed to carry out assigned work at their seats and then were free to select other activities when the assignment was finished while for pupils who were low in motivation growth was greater in classrooms where this activity occurred infrequently.

Socioeconomic status has already been cited as another pupil characteristic which must be considered in identifying which teacher behaviors will be most functional, as shown in Figure 7.

These findings of nonlinear relationships and interactions, which seem to make theoretical sense, indicate the need for complex statistical analyses in program validation because of their potential for dealing with the reality of the phenomena more adequately than the simple linear correlations or t tests used in the past. As evidence of this, in two sets of our data the number of significant linear correlations was about one and a half times the number expected by chance, but the number of significant nonlinear relations and interactions was between three and four times the expected number.

Past teacher effectiveness research, then, seems to have been asking questions that were too simple to represent the reality of the situation. At an intuitive level it is clear that the classroom is a highly complex place and to try to understand it by asking simple questions is probably not a productive way to go.

A Suggestion for Analysis. It is possible that some of our readers may be deterred from using complex methods for analyzing the relations between teacher behavior and pupil outcomes, including nonlinear relations and interactions, because of the apparent difficulty of the data analysis, even though they may find the rationale for doing so quite compelling. There are recent developments in this area that simplify the task considerably by making it possible to use multiple regression procedures to answer questions which used to be answered by analysis of variance. This approach offers a degree of convenience and flexibility which is an important practical advantage, as well as offering some theoretical advantages (Cohen, 1968; Kelly, Beggs and McNeil, 1969; Walberg, 1971; Kerlinger and Pedhazur, 1973). A widely available program suitable for such analyses is the Step-Wise Multiple Regression, BMD02R, from the Biomedical Computer Program Library (Dixon, 1974), which is available at most computer centers.

There are also more complex approaches, forms of multivariate analysis, which may be even more informative, but they require greater skill and more degrees of freedom to exploit their power.

In an undertaking like program validation, in which most of the effort and money goes into collecting the data and reducing it to an appropriate form, the statistical analyses represent a small portion of the total. To settle for an analysis which does not fully exploit the data would be foolish economy. Complex phenomena need complex analysis if they are to be understood.

## Concluding Comments

We have presented a simple paradigm which separates the educational process into identifiable points for assessment: teacher training, teacher behavior or performance, pupil behavior, and pupil outcomes. The paradigm permits specifying the point at which assessment may take place and identifies lines of influence between these points. It makes clear, for example, why studies which simply examine relations between teacher training and pupil outcomes are not likely to be productive because there are too many unidentified steps (and too many variables) in between. It indicates that the major difference between PBTE and past practice is that it shifts evaluation from the training program itself to the behavior of the teacher who is graduated from that program. Teacher behavior becomes the output of the training program and the input into the real-life classroom. Program evaluation, then, examines the relation between training and teacher behavior and program validation examines the relation between teacher behavior and pupil behavior or pupil outcome.

Problems in using pupil outcomes as criteria for evaluating teachers were presented and the conclusion was drawn that the problems are disabling. Measures of even year-long growth of pupils are so unreliable that data based on gains of about 20 classes per teacher would be required to evaluate teachers with the customary minimum standards of reliability.

In the interest of practicality, teaching periods of a few hours or a few days are frequently used. But the problem of whether short-term gain relates to long-term gain, the probability that short-term gain will be measured at lower cognitive levels, and our lack of knowledge about whether teacher behaviors which promote learning at low cognitive levels also promote learning at high cognitive levels all raise question about these short-term approaches. There are the further difficulties that the objectives of education are many and complex, so that repeated measures for multiple outcomes would be required. Finally, and most importantly, to evaluate the teacher on the growth of the pupils is to base the fate of the teacher on the behavior of others over which he neither has effective control nor, it seems to us, should have it. These are some of the reasons why evaluating teachers by measuring the gains of their pupils is impractical and probably immoral as well.

Rather, the better procedure for evaluating teachers would be the *measurement of teacher behavior*, which is under his control to a greater degree, although even this measurement is neither simple nor easy. Some of the problems of specifying and measuring competencies may be eased, however, by using a hierarchical organization in which specific behavioral items and the need to talk about large numbers of specific competencies, is bypassed for most purposes. But when this procedure is followed, it is critical to assure empirically that the items in each group do, in fact, belong together.

While we recognize that program development will need to proceed on the basis of current theory and knowledge, even though current knowledge is weak, it is important to utilize the research knowledge which does exist. Not only does it show that some concepts are weak and others in error, but it suggests additional concepts and measures which go beyond present theory and should be considered in the specification of competencies.

But important as program development and evaluation are, we will not in the long run know whether they have advanced education until we know whether they make a difference to pupils--until we validate our programs. We need to know empirically that the teacher behaviors which programs teach do, in fact, produce the desired outcomes for the pupils taught. This is the step in the process which has most often been omitted. But the weakness of theory and research on which programs are now based, coupled with the high cost of program development and the increasing concern by the public for accountability in education, leave no alternative to moving ahead without delay in this critical area.

The cost in time and money involved in the conduct of validation research will be reduced materially if it is done in the context of an ongoing program rather than separately from such a program. There is some evidence that complex studies produce meaningful results when they bring together information about pupil, school, and community, along with multiple low inference measures of teacher and pupil behavior, and multiple outcome measures, but only when the data are analyzed by procedures which can exploit their richness. While not simple or easy, such studies are feasible with the present "state of the art." This will be a difficult and expensive process, but minor compared to the difficulty and expense of program development and operation which we take for granted.

While the assessment levels in the paradigm present fairly discrete points for assessment, taken as a whole, the paradigm provides a dynamic model in which training experiences or programs may be evaluated in terms of teacher behaviors and pupil behaviors and outcomes may be used to validate teacher behaviors. The results of both processes can then be fed back into program modification. This becomes a continuous process of train, evaluate, validate, feedback, modify, train, etc. In this sense, evaluation and validation represent a small investment whose potential return appears to be great. This process, with its empirical feedback loops, may be the key to a new era in education--one in which we really begin to *know* what a teacher can do to help pupils learn and what a program must do to teach these skills to teachers.

## REFERENCES

American Association of Colleges for Teacher Education, Committee on Performance-Based Teacher Education. Achieving the Potential of Performance-Based Teacher Education: Recommendations. Washington, D.C.: American Association of Colleges for Teacher Education, February 1974.

Boyer, E. Gil; Simon, Anita; and Karafin, Gail, eds. Measures of Maturation: an Anthology of Early Chil'hood Observation Instruments. Vols. 1-3. Philadelphia: Research for :ter Schools, Inc., 1973.

Brophy, Jere E. and Evertson, Carolyn M. The Texas Teacher Effectiveness Project: Presentation of Non-Linear Relationships and Summary Discussion, Report No. 74-6. Austin: Research and Development Center, University of Texas, 1974.

Cohen, Jacob. "Multiple Regression as a General Data Analytic System." Psychological Bulletin 70 (1968): 426-443.

Dixon, W.J., ed. BMD: Biomedical Computer Programs.Berkeley: University of California Press, 1974.

Dunkin, Michael J. and Biddle, Bruce J. The Study of Teaching. New York: Holt, Rinehart, and Winston, Inc., 1974.

Elam, Stanley. Performance-Based Teacher Education: What Is the State of the Art? Washington, D.C.: American Association of Colleges for Teacher Education, December 1971.

Flanders, Ned A. Analyzing Teacher Behavior. Reading, Massachusetts: Addison-Wesley, 1970.

____________. "The Changing Base of Performance-Based Teaching." Phi Delta Kappan 55 (1974): 312-315.

Joyce, Bruce R. Performance-Based Teacher Education Design Alternatives: The Concept of Unity. Washington, D.C.: American Association of Colleges for Teacher Education, 1974.

Kay, Patricia M. What Competencies Should Be Included In a C/PBTE Program? Washington, D.C.: American Association of Colleges for Teacher Education, 1975.

Kelly, Francis J.; Beggs, Donald L.; and McNeil, Keith A. Research Design in the Behavioral Sciences: Multiple Regression Approach. Carbondale, Illinois: Southern Illinois University Press, 1969.

Kerlinger, Fred A. and Pedhazur, Elazar J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, Inc., 1973.

Medley, Donald M. and Mitzel, Harold E., "Measuring Classroom Behavior by Systematic Observation," in N.L. Gage, ed. Handbook of Research on Teaching. Chicago: Rand McNally and Company, 1963.

Merwin, Jack C. Performance-Based Teacher Education: Some Measurement and Decision-Making Considerations. Washington, D.C.: American Association of Colleges for Teacher Education, 1973.

Morsh, J.E. and Wilder, Eleanor W. Identifying the Effective Instructor: A Review of the Quantitative Studies, 1900-1952. AFPTRC-TR-54-44. United States Air Force Personnel Training and Research Center, 1954.

Page, Ellis B. "How We All Failed at Performance Contracting." Phi Delta Kappan 54 (1972): 115-117.

__________. "A Final Footnote on PC and OEO." Phi Delta Kappan 8 (1973): 575.

Popham, W. James. "Alternative Teacher Assessment Strategies," in T. Andrews, ed., Assessment. Albany: Multi-State Consortium on Performance-Based Teacher Education.

__________. "Performance Tests of Teaching Proficiency: Rationale, Development and Validation." American Educational Research Journal 8 (1971): 105-117.

Rosenshine, Barak. "The Stability of Teacher Effects Upon Student Achievement." Review of Educational Research 40 (1970): 647-662.

__________. Teaching Behaviors and Student Achievement. London: National Foundation for Educational Research, 1971.

Rosenshine, Barak and Furst, Norma. "Research on Teacher Performance Criteria," in B. O. Smith, ed., Research in Teacher Education: A Symposium. Englewood Cliffs, N.J.: Prentice-Hall, 1971.

__________. "The Use of Direct Observation to Study Teaching," in R. M. W. Travers, ed., Second Handbook of Research and Teaching. Chicago: Rand-McNally, 1973.

Rowe, Mary Budd. "Wait-time and Rewards as Instructional Variables: Their Influence on Language, Logic, and Fate Control." Journal of Research in Science Teaching 11 (1974): 81-94.

Simon, Anita and Boyer, E. Gil, eds. Mirrors for Behavior: An Anthology of Classroom Observational Instruments. Vols. 1-6. Philadelphia: Research for Better Schools, Inc., 1967.

______________________, eds. Mirrors for Behavior: An Anthology of Classroom Observationa. Instruments Continued. Summary and Supplemental Volumes A and B. Vols. 7-14. Philadelphia: Research for Better Schools, Inc., 1970.

Small, Alan A. "Accountability in Victorian England." Phi Delta Kappan 11 (1972): 438-439.

Soar, Robert S. An Integrative Approach to Classroom Learning NIMH project numbers 5-R11 MH 01096 and 7-R11 MH 02045. Philadelphia: Temple University, 1966. (ERIC Document Reproduction Service No. Ed 033 749).

____________. "Optimum Teacher-Pupil Interaction for Pupil Growth." Educational Leadership: Research Supplement 26 (1968): 275-280.

____________. Follow Through Classroom Process Measurement and Pupil Growth, 1970-1971. DHEW OE Contract No. OEG-0-8-522394-3991 (286). Gainesville, Florida: Institute for Development of Human Resources, University of Florida, 1973.

Soar, Robert S. and Soar, Ruth M. "An Empirical Analysis of Selected Follow-Through Programs: An Example of a Process Approach to Evaluation," in I. J. Gordon,ed., Early Childhood Education. Chicago: National Society for the Study of Education, 1972.

______________________. Classroom Behavior, Pupil Characteristics, and Pupil Growth for the School Year and for the Summer. NIMH Project Numbers 5 ROI MH 15891 and 5 ROI MH 15626. Gainesville, Florida: Institute for Development of Human Resources, University of Florida, 1973. JSAS Catalog of Selected Documents in Psychology, in press.

Soar, Robert S; Soar, Ruth M.; and Ragosta, Marjorie J. Florida Climate and Control System: Observer's Manual. Gainesville, Florida: Institute for Development of Human Resources, University of Florida, 1971.

Solomon, Daniel; Bezdek, William E.; and Rosenberg, Larry. Teaching Styles and Learning. Chicago: The Center for the Study of Liberal Education of Adults, 1963.

Taba, Hilda; Levine, Samuel; and Elzey, Freeman F. Thinking in Elementary School Children. Coop. Res. Proj. No. 1574, Office of Education, U. S. Department of Health, Education and Welfare. San Francisco: San Francisco State College, 1964.

Veldman, Donald J. and Brophy, Jere E. "Measuring Teacher Effects on Pupil Achievement." Journal of Educational Psychology 66 (1974): 319-324.

Walberg, Herbert J. "Generalized Regression Models in Educational Research." American Educational Research Journal 8 (1971): 71-91.

# ABOUT AACTE

The American Association of Colleges for Teacher Education is an organization of more than 860 colleges and universities joined together in a common interest: more effective ways of preparing educational personnel for our changing society. It is national in scope, institutional in structure, and voluntary. It has served teacher education for 55 years in professional tasks which no single institution, agency, organization, or enterprise can accomplish alone.

AACTE's members are located in every state of the nation and in Puerto Rico, Guam, and the Virgin Islands. Collectively, they prepare more than 90 percent of the teaching force that enters American schools each year.

The Association maintains its headquarters in the National Center for Higher Education, in Washington, D. C. -- the nation's capital, which also in recent years has become an educational capital. This location enables AACTE to work closely with many professional organizations and government agencies concerned with teachers and their preparation.

In AACTE headquarters, a stable professional staff is in continuous interaction with other educators and with officials who influence education, both in immediate actions and future thrusts. Educators have come to rely upon the AACTE headquarters office for information, ideas, and other assistance and, in turn, to share their aspirations and needs. Such interaction alerts the staff and officers to current and emerging needs of society and of education and makes AACTE *the* center for teacher education. The professional staff is regularly out in the field--nationally and internationally--serving educators and keeping abreast of the "real world." The headquarters office staff implements the Association's objectives and programs, keeping them vital and valid.

Through conferences, study committees, commissions, task forces, publications, and projects, AACTE conducts a program relevant to the current needs of those concerned with better preparation programs for educational personnel. Major programmatic thrusts are carried out by commissions on international education, multicultural education, and accreditation standards. Other activities include government relations and a consultative service in teacher education.

A number of activities are carried on collaboratively. These include major fiscal support for and selection of higher education representatives on the National Council for Accreditation of Teacher Education--an activity sanctioned by the National Commission on Accrediting and a joint enterprise of higher education institutions represented by AACTE, organizations of school board members, classroom teachers, state certification officers, and chief state school officers.

The Association headquarters provides secretariat services for two organizations which help make teacher education more interdisciplinary and comprehensive: the Associated Organizations of Teacher Education and the International Council on Education for Teaching. A major interest in teacher education provides a common bond between AACTE and fraternal organizations.

AACTE is deeply concerned with and involved in the major education issues of the day. Combining the considerable resources inherent in the consortium--constituted through a national voluntary association--with strengths of others creates a synergism of exceptional productivity and potentiality. Serving as the nerve center and spokesman for major efforts to improve education personnel, the Association brings to its task credibility, built-in cooperation and communications, contributions in cash and kind, and diverse staff and membership capabilities.

AACTE provides a capability for energetically, imaginatively, and effectively moving the nation forward through better prepared educational personnel. From its administration of the pioneering educational television program, "Continental Classroom," to its involvement of 20,000 practitioners, researchers, and decision makers in developing the current *Recommended Standards for Teacher Education*, to many other activities, AACTE has demonstrated its organizational and consortium qualifications and experiences in conceptualizing, studying and experimenting, communicating, and implementing diverse thrusts for carrying out socially and educationally significant activities. With the past as prologue, AACTE is proud of its history and confident of its future among the "movers and doers" seeking continuous renewal of national aspirations and accomplishments through education.

## ABOUT THE TEXAS TEACHER CENTER PROJECT

The AACTE Committee on Performance-Based Teacher Education serves as the national component of the Texas Teacher Center Project. This Project was initiated in July, 1970, through a grant to the Texas Education Agency from the Bureau of Educational Personnel Development, USOE. The Project was initially funded under the Trainers of Teacher Trainers (TTT) Program and the national component was subcontracted by the Texas Education Agency to AACTE.

One of the original thrusts of the Texas Teacher Center Project was to conceptualize and field test performance-based teacher education programs in pilot situations and contribute to a statewide effort to move teacher certification to a performance base. By the inclusion of the national component in the Project, the Texas Project made it possible for all efforts in the nation related to performance-based teacher education to gain national visibility. More important, it gave to the nation a central forum where continuous study and further clarification of the performance-based movement might take place.

While the Texas Teacher Center Project is of particular interest to AACTE's Performance-Based Teacher Education Committee, the services of the Committee are available, within its resources, to all states, colleges and universities, and groups concerned with the improvement of preparation programs for school personnel.

## AACTE BOARD OF DIRECTORS

Executive Committee:

*John Dunworth*, President and Chairman of the Board, AACTE; President, George Peabody College for Teachers,
Nashville, Tennessee 37203

*Sam P. Wiggins*, Immediate Past President, AACTE; Dean, College of Education, The Cleveland State
University, Cleveland, Ohio 44115

*Frederick R. Cyphert*, President-elect, AACTE; Dean, College of Education, The Ohio State University,
Columbus, Ohio 43210

*Dean C. Corrigan*, Dean, College of Education and Social Services, The University of Vermont, Burlington,
Vermont 05401

*Bert L. Sharp*, Dean, College of Education, University of Florida, Gainesville, Florida 32601

Ex Officio Member: *Edward C. Pomeroy*, Executive Director, AACTE, One Dupont Circle, Washington, D.C.
20036

*Sister Maria Petra Dempsey*, Chairwoman, Department of Elementary and Early Childhood Education, Xavier
University of Louisiana, New Orleans, Louisiana 70125

*Daniel E. Griffiths*, Dean, School of Education, New York University, New York, New York 10003

*C. Glen Hass*, Professor of Education, College of Education, University of Florida, Gainesville, Florida
32611

*Robert Heideman*, Director of Educational Placement, University of Wisconsin-Madison, Madison, Wisconsin
53706

*Henry J. Hermanowicz*, Dean, College of Education, The Pennsylvania State University, University Park,
Pennsylvania 16802

*Asa G. Hilliard*, Dean, School of Education, San Francisco State University, San Francisco, California
94132

*James Kelly, Jr.*, Dean, School of Education, University of Pittsburgh, Pittsburgh, Pennsylvania 15213

*Loretta P. Konecki*, Assistant Professor, Department of Education, Regis College, Denver, Colorado 80221

*Paul B. Mohr, Sr.*, Dean, College of Education, Florida Agricultural and Mechanical University, Tallahassee,
Florida 32307

*Curtis E. Nash*, Dean, School of Education, Central Michigan University, Mount Pleasant, Michigan 48859

*J. T. Sandefur*, Dean, College of Education, Western Kentucky University, Bowling Green, Kentucky 42101

*Betty B. Schantz*, Assistant Dean, University-School Relations, Temple University, Philadelphia,
Pennsylvania 19122

*Bob G. Woods*, Dean, College of Education, University of Missouri-Columbia, Columbia, Missouri 65201

Liaison Members:

*Dale Garland*, Associate Director, Instruction and Professional Development, NEA, 1201 Sixteenth Street,
N.W., Washington, D.C. 20036

*Rolf W. Larson*, Director, National Council for Accreditation of Teacher Education, 1750 Pennsylvania
Avenue, N.W., Washington, D.C. 20006

# AACTE ORDER FORM FOR OTHER RECENT AACTE PUBLICATIONS

Number of Copies

## YEARBOOKS - Annual Meeting Sessions

______ *Strengthening the Education of Teachers* 1975 (available after June 1975)

______ *Ferment and Momentum in Teacher Education* 1974 105 pages $4.00

## POSITION PAPERS

______ *Teaching Centers: Toward the State of the Scene* - Allen Schmieder, Sam J. Yarger, 1974, 50 pages $3.00

______ *Accreditation Problems & the Promise of PBTE* - Rolf W. Larson, 1974, 29 pages $3.00

## TEACHER EDUCATION CONCEPTUAL MODELS

______ *Obligation for Reform* George Denemark, Joost Yff, 1974, 68 pages $2.00

Number of Copies

## THE CHARLES W. HUNT LECTURES

______ *Strengthening the Education of Teachers* - C. E. Gross 1975 $1.50

______ *Ferment and Momentum in Teacher Education* - Margaret Lindsey 1974, 23 pages $1.00

*Journal of Teacher Education* (Quarterly)

______ One-year subscription - $10.00
______ Three-year subscription - $25.00
______ Back issues available - $ 3.00 ea.
(Specify date)______

## INTERNATIONAL-MULTICULTURAL EDUCATION

______ *Multicultural Education through Competency-Based Teacher Education* William A. Hunter, Editor, 1974, 288 pages $6.00

BILLED ORDERS: Billed orders will be accepted only when made on official purchase orders of institutions, agencies, or organizations. Shipping and handling charges will be added to billed orders. Payment must accompany all other orders. There are no minimum orders. A 10 percent discount is allowed on purchases of five or more publications of any one title.

Payment enclosed______ Amount______

Purchase Order Number______

NAME______
(Please print or type)

ADDRESS______

______ZIP CODE______

Ask for our complete list of AACTE publications on teacher education.

Send orders to: Order Department, American Association of Colleges for
Teacher Education, Suite #610, One Dupont Circle,
Washington, D.C. 20036

AACTE SPECIAL SERIES ON PBTE

| Number of Copies | PBTE Monograph Series | |
|---|---|---|
| ________ | #1 | "Performance-Based Teacher Education: What Is the State of the Art?" by Stanley Elam @ $2.00 |
| ________ | #2 | "The Individualized, Competency-Based System of Teacher Education at Weber State College" by Caseel Burke @ $2.00 |
| ________ | #3 | "Manchester Interview: Competency-Based Teacher Education/Certification" by Theodore Andrews @ $2.00 |
| ________ | #4 | "A Critique of PBTE" by Harry S. Broudy @ $2.00 |
| ________ | #5 | "Competency-Based Teacher Education: A Scenario" by James Cooper and Wilford Weber @ $2.00 |
| ________ | #6 | "Changing Teacher Education in a Large Urban University" by Frederic T. Giles and Clifford Foster @ $3.00 |
| ________ | #7 | "Performance-Based Teacher Education: An Annotated Bibliography" by AACTE and ERIC Clearinghouse on Teacher Education @ $3.00 |
| ________ | #8 | "Performance-Based Teacher Education Programs: A Comparative Description" by Iris Elfenbein @ $3.00 |
| ________ | #9 | "Competency-Based Education: The State of the Scene" by Allen A. Schmieder (jointly with ERIC Clearinghouse on Teacher Education) @ $3.00 |
| ________ | #10 | "A Humanistic Approach to Performance-Based Teacher Education" by Paul Nash @ $2.00 |
| ________ | #11 | "Performance-Based Teacher Education and the Subject Matter Fields" by Michael F. Shugrue @ $2.00 |
| ________ | #12 | "Performance-Based Teacher Education: Some Measurement and Decision-Making Considerations" by Jack C. Merwin @ $2.00 |
| ________ | #13 | "Issues in Governance for Performance-Based Teacher Education" by Michael W. Kirst @ $2.00 |
| ________ | #14 | "Performance-Based Teacher Education Design Alternatives: The Concept of Unity" by Bruce R. Joyce, Jonas F. Soltis, and Marsha Weil @ $3.00 |
| ________ | #15 | "A Practical Management System for Performance-Based Teacher Education" by Castelle Gentry and Charles Johnson @ $3.00 |
| ________ | #16 | "Achieving the Potential of Performance-Based Teacher Education: Recommendations" by the AACTE Committee on Performance-Based Teacher Education @ $3.00 |
| ________ | #17 | "Assessment and Research in Teacher Education: Focus on PBTE" by Donald M. Medley, Ruth and Robert Soar @ $3.00 |

---

| Number of Copies | Technical Assistance Paper Series | |
|---|---|---|
| ________ | #1 | "What Competencies Should Be Included in a C/PBTE Program?" by Patricia Kay @ $2.50 |

BILLED ORDERS: Billed orders will be accepted only when made on official purchase orders of institutions, agencies, or organizations. Shipping and handling charges will be added to billed orders. Payment must accompany all other orders. There are no minimum orders.

DISCOUNTS: A 10 percent discount is allowed on purchase of five or more publications of any one title. A 10 percent discount is allowed on all orders by wholesale agencies.

Payment enclosed________________________ Amount________________________

Purchase Order No.________________________

NAME________________________________________________________________

ADDRESS____________________________________________ZIP CODE________________

Please address: Order Department, American Association of Colleges for Teacher Education, Suite #610,
One Dupont Circle, Washington, D.C. 20036

## PBTE ADVISORY COUNCIL MEMBERS

The Committee Membership plus the following persons comprise the Advisory Council

*Elbert Brooks*, Superintendent of Schools, Metropolitan Schools, 2601 Bransford Avenue, Nashville, Tennessee 37203

*George Denemark*, Dean, College of Education, University of Kentucky, Lexington, Kentucky 40506

*George Dickson*, Dean, College of Education, University of Toledo, Toledo, Ohio 43606

*William Jenkins*, Vice President for Academic Affairs, Florida International University, Tamiami Trail, Miami, Florida 33174

*J. W. Maucker*, Vice President for Academic Affairs, Academic Affairs Office, Kansas State Teachers College, Emporia, Kansas 66801

*Donald McCarty*, Professor of Educational Administration, University of Wisconsin, Madison, Wisconsin 53706

*Michael Shugrue*, Dean, Community Relations and Academic Development, Richmond College, City University of New York, 130 Stuyvesant Place, Staten Island, New York 10301

*John Skinner*, President, Student NEA, 1201 16th Street, Room 501, Washington, D.C. 20036

*Atilano Valencia*, Chairman, Department of Education, New Mexico Highlands University, Las Vegas, New Mexico 87701

LIAISON MEMBERS:

*Theodore Andrews*, Associate in Teacher Education and Certification, New York State Department of Education, Albany, New York 12201 *Multi-State Consortium*

*James Collins*, Assistant Dean, College of Education, Syracuse University, Syracuse, New York 13210 *National Consortium of CBE Centers*

*Norman Johnson*, Chairman, Department of Education, North Carolina Central University, Durham, North Carolina 27707 *Southern Consortium*

*Donald Orlosky*, Professor of Education and Director of LTI, University of South Florida, Tampa, Florida 33620 *Leadership Training Institute for Educational Personnel Development*

*Allen Schmieder*, Chief, Support Programs, Division of Educational Systems Development, U.S. Office of Education, Washington, D.C. 20202 *Division of Educational Systems Development.*

*Emmitt Smith*, Research Professor, West Texas State University, Canyon, Texas 79075 *Texas Center for the Improvement of Educational Systems Project*

*Frank Sobol*, Deputy Director for Programs in Teaching and Curriculum, Office of Research, National Institute of Education, Department of Health, Education and Welfare, Washington, D.C. 20208 *National Institute of Education*

*James Steffensen*, Acting Chief, Program Development Branch, Teacher Corps, 400 Maryland Avenue, S.W., Washington, D.C. 20202 *Teacher Corps*