DOCUMENT RESUME

ED 117 110 SP 009 813

AUTHOR Roth, Robert A.
TITLE The Nature of and Alternatives for Teacher Competency Statements and Implications for Assessment Techniques.
INSTITUTION Michigan State Dept. of Education, Lansing. Teacher Preparation and Professional Development Services.
PUB DATE Sep 75
NOTE 51p.

EDRS PRICE MF-$0.76 HC-$3.32 Plus Postage
DESCRIPTORS Educational Accountability; *Evaluation Criteria; *Evaluation Methods; Performance; *Performance Based Teacher Education; *Performance Criteria; *Teacher Education; Teacher Evaluation

ABSTRACT

Competency based teacher education has been defined in various ways, but there is general agreement on at least two basic elements. The first essential characteristic is the specification of teacher competencies which form the basis of the entire program. The second is the design of assessment techniques directly related to the specific competencies. Competencies have been written in a variety of ways and have been related to various domains or competency areas. In each of the competency domains the form of the competency must be examined to determine appropriate assessment techniques. There are a number of assessment factors which need to be considered in the evaluation of competencies. The nature of the standards, or criterion selection, is essential. Other concerns are comprehensiveness and fidelity of the assessment system, validity and reliability of data, and general utility of the process. Assessment of knowledge competencies can be accomplished through paper and pencil testing. Assessment of teaching behaviors or performances, however, requires observation of the individual demonstrating the skill. This may be accomplished by rating scales or structured observation systems. Utilizing sampling and student achievement have also been used, although it has been concluded that student learning measures cannot be fairly used to evaluate individual teachers at present. (An extensive list of references is included.) (RC)

THE NATURE OF AND ALTERNATIVES FOR TEACHER COMPETENCY STATEMENTS AND IMPLICATIONS FOR ASSESSMENT TECHNIQUES

Robert A. Roth

Michigan Department of Education
Teacher Preparation and Professional Development Services
September 1975

CONTENTS

THE NATURE OF AND ALTERNATIVES FOR TEACHER COMPETENCY STATEMENTS AND IMPLICATIONS FOR ASSESSMENT TECHNIQUES

Competency-based teacher education is perhaps the most frequently discussed topic in education today. Close to 500 teacher education institutions (Sherwin, 1973, p. 3), and over 35 states (Roth, 1974) have become involved in either studying or developing such programs. Competency-based teacher education has been defined in various ways but there is general agreement on at least two basic elements. The first essential characteristic is the specification of teacher competencies which form the basis of the entire program. The second is the design of assessment techniques, directly related to the specified competencies, which are necessary in order to determine whether or not a student has achieved the competencies.

Competency Domains

In view of the critical role of these competencies it is important to review the nature of competency statements and their implications for assessment. Competencies have been written in a variety of ways and have been related to various domains or competency areas. The competency domains identified in the literature are knowledge, behaviors, affect, consequences, and experiences. Each of these needs to be examined to determine implications for possible assessment strategies.

Knowledge domain competencies refer to information and cognitive processes necessary for effective instruction and related activities. These include knowledge of: a subject area, planning for instruction, instructional strategies, child growth and development, human relations, etc. Knowledge in these areas deals with facts, processes, theories, and techniques. The scope of the knowledge competencies will depend upon what areas of the teacher education program (content area, liberal arts, professional education) are included in the competency-based program. Examples from various knowledge areas would be an ability to balance chemical equations, write behavioral objectives, identify a variety of instructional techniques, describe Piaget's stages of development, and relate counseling techniques appropriate to given situations (the specificity of these competencies will be discussed in a later section). These are usually evaluated by paper and pencil processes such as those utilized in current traditional teacher education programs.

Some educators have referred to an area of competence which usually is considered as being either in the area of knowledge or performance, and may belong somewhere between the two. These competencies have been identified as "outputs" and are described in the statement which follows:

> Teachers produce a variety of outputs which can be categorized as either Products, Events, or Conditions. Included among these categories of outputs are the following:
>
> A. Products - A product is a tangible, concrete, transportable outcome of work effort.
>
> Instructional units
> Lesson Plans
> Lists of objectives
> Guides, outlines, sets of directions
> Bulletin boards
>
> B. Events - An event represents an instance of occurrence of an observable transaction or set of behaviors.
>
> Class discussion
> Demonstration
> Presentation
> Field Trip
>
> C. Conditions - A condition represents an instance of a desired circumstance expected to endure and to influence a program.
>
> Parent acceptance of school program
> Classroom Climate
> School atmosphere
> Working relationships with other teachers (Morse, Smith, and Thomas, 1972, pp. 11-12)

The behavior domain refers to the performance competencies an individual demonstrates. These are the actual teaching acts considered necessary in order to enable students to learn. The performance of teaching skills is based on the previously acquired knowledge competencies, but requires a demonstration that the student can perform and utilize various strategies and techniques. Examples here include demonstration of a variety of questioning skills, introduction of a lesson, guiding students in discovery activities, etc.

The affective domain has been identified in the literature as the opinions, attitudes, emotions, and dispositions of the teacher. This covers a variety of specific factors such as sensitivity to needs of students, self-acceptance, professionalism, etc. Human relations training labs and interaction laboratories have been established to accomplish these competencies. It is important to note, however, that we may wish to distinguish between affective competencies, such as accepting student feelings, which are expected to be demonstrated in the classroom, and personality variables of the trainee, such as emotional security, which are more difficult to elicit and evaluate.

Consequence domain objectives relate to the influence the teacher has on pupils. In these competencies the criterion considered is the product; i.e., the behaviors or attitude and achievement gains of the pupils being instructed.

The consequence area, however, can be separated into at least two distinct categories, student behaviors and student learning. Student behaviors refer to those activities students engage in which are assumed necessary to attain the educational objectives. Some programs are placing a great deal of emphasis on evaluating this dimension of teacher competence. Examples of student activities include the following:

1) students being supportive and cooperative
2) students being attentive to class activities
3) students participating in verbal interaction
4) students following specific activities to completion
5) students using media and resources for study
(Hatfield, 1974, pp. 41,42)

An example of the second type, a pupil achievement consequence objective, is

> Given fifth grade pupils who have not mastered their multiplication facts, the pupils will be able to master all the facts (1-10) X (1-10) and be able to complete them on a paper and pencil test at a rate of 30 per minute. The criterion is 90% accuracy by at least two out of three pupils within four weeks.

Experience or expressive domain objectives have been described as activities an individual engages in which are outcomes in themselves. There are no specified outcomes which are to occur as a result of the experience, the objective is complete once the individual has experienced the activity. An example is "the student will read a story to a kindergarten child--while holding the child on his lap," or "the student will visit the home of each of his pupils (Weber, 1970)."

Competency Forms

There seems to be a variety of viewpoints as to how competencies should be written. One approach is to write them as general statements of behavior with some broadly defined expected level of achievement. An example of this approach is "the teacher is able to use a variety of teaching techniques, selecting those which are appropriate in particular situations." Note that the competency is general enough to cover a number of specific behaviors. Also, the standard of achievement "appropriate" is not very specific and provides for a more subjective evaluation. These are high inference types of competencies.

Merwin (1973), however, argues that PBTE is supposed to differ from current teacher education programs by the explicitness with which the competencies and the criteria used in assessing their mastery are stated. Further, this explicitness should leave little or no ambiguity regarding procedures for assessing the performance nor in arriving at a decision as to whether or not the individual possesses it.

In addition, Morse, et al. (1972), believe that evaluation goes beyond measurement of performance. Judgments have to be made in relation to those factors which give meaning to the performance information produced. Central to this judgmental process is a clear delineation of what it is the assessment is to assess.

Pursuing this line of thinking, another approach would be to develop specific performance objectives derived from the competency statement. These specific performance objectives are behaviors which must be demonstrated as evidence that one has attained the generic competency from which they were derived. In this situation, the evaluation focuses on the demonstration of the more specific behaviors and achievement of the competency is determined by whether or not most or all of the specific performances were demonstrated. This is a lower inference type of objective and is somewhat less subjective in nature. An example is

competency:	The teacher trainee is able to use a variety of teaching techniques.
performance objectives:	The teacher trainee will demonstrate ability to give a lecture by stating objectives clearly, using an audible voice, varying the pace, establishing eye contact, and summarizing key points.

The teacher trainee will demonstrate ability to conduct a group discussion by defining the topic, involving all students, summarizing key points, . . .

The teacher trainee will demonstrate ability to employ oral questioning. . .

The teacher trainee will demonstrate ability to give a demonstration . . . etc.

Competency statements may also be written as behavioral objectives. This is the type of competency statement most frequently believed to be associated with competency-based programs. In this approach, the behavior, mastery level, and conditions are specifically stated, with the criterion levels stated as frequencies, per cent accuracy, or other such measures. In this approach, competency statements can be used directly as assessment criteria. Examples of behavorial objectives are

Given examples of classroom management techniques (written descriptions or videotaped) the teacher trainee will identify by name at least five of six correctly.

Given a small group of students in a microteaching session the trainee will ask one knowledge, one application, and one synthesis type question as developed in his lesson plan within a twenty-five minute lesson.

In each of the competency domains cited, the form of the competency statement must be examined to determine appropriate assessment techniques needed to evaluate achievement. It should be noted that the assessment strategies are affected by a variety of variables related to competency statements. As each of the competency areas are examined in the following pages, variables such as context and specificity will be considered as they relate to the particular competency domain under discussion.

Assessment Factors

In order to determine implications and problems of assessment of competencies, an analysis of the literature was conducted to determine assessment practices and concerns. Remaining sections of this paper reflect these findings.

There are a number of factors relating to assessment of competencies in general. One such concern is the evaluation context. For example, if the individual is required to demonstrate that he has a particular skill, he might accomplish this by teaching to one or two peers, a small group of students, or an entire class. In each instance, he is demonstrating that he can perform the skill, and each of these alternative contexts may be appropriate.

In some cases, however, the competency may require that the individual not only be able to demonstrate a particular skill, but that he utilize this skill at the appropriate time or at a designated frequency as part of his normal teaching style over a period of time. This requires not only that the individual "can do" but "does do." This type of competency requires the classroom as a context for evaluation, as well as a longer period of time for observation.

The nature of the competency statement clearly has implications for the context required. On the other hand, the context in which assessment takes place has a direct bearing on the nature of the outcomes and the data collected in the assessment process. Context variables need to be considered when evaluating competencies, and some standardization is necessary (when possible) in order to make comparable evaluations.

As an example of this relationship, with particular reference to performance standards, Schalock, et al., (1974) discussed the competency "defining the objectives of instruction." He points out that there is nothing inherent in this competency that is addressed to the quality expected (standard) nor is there any reference to the context in which performance is to take place. Also

> Because of this interdependency of competency descriptor, the context in which competency is to be demonstrated, and the performance standard set for its demonstration, the task of becoming clear as to what the assessment system was to do and how it was to do it was more difficult than anticipated (Schalock, Garrison, and Kersh 1974).

Although the setting of standards is a key element in the design of assessment, it is not an easy task.

> Obviously, there is no source, other than judgment, to which one can refer to select appropriate standards. The question of standards is one which plagues all evaluation efforts. However, the nature of the competency, its relevance to instruction, its suspected impact on classroom learning and other such considerations should be weighed in setting the standard (Airasian, 1974, p. 16).

Some assistance in the criteria selection process is provided in the following statement. It should be noted, however, that this was written in terms of assessing teachers in general rather than assessing specific competencies.

Six Attributes for Discriminating Among Criterion Measures

1) Differentiates among teachers. There are decisions where we do not have enough knowledge merely by knowing that a teacher has met a minimal level of proficiency. Both administrators and researchers, for instance, often encounter situations where they need a measure sensitive enough to assess variance in teachers' skills.

2) Assess learner growth . . . emphasize the necessity to produce criterion measures which can be used to assess the results of instructional process, not merely the process itself. In certain limited instances we may not be interested in the outcomes of instruction as reflected by modifications in the learner; but these would be few in number. Certain classes of criterion measurers are notoriously deficient with respect to this attribute.

3) Yields data uncontaminated by required inferences. An attribute of considerable importance is whether a measure permits the acquisition of data with a minimum of required extrapolation on the part of the user. If all observations are made in such a way that beyond human frailty they have not been forced through a distorting inferential sieve, then the measure is better. A classroom observation system which asked the user to record the raw frequency of teacher questions would possess the attribute more so than a system which asked the user to judge the warmth of teacher questions.

4) Adapts to teachers' goal preferences. A measure of teaching skill will be more useful for given situations if it can adapt to such dissimilarities in goal preferences.

5) Presents equivalent stimulus situations. There are times when we might like to use a measure which would permit the measurement of teaching proficiency when the stimulus situations were identical or at least comparable.

6) Contains heuristic data categories. In a sense this final attribute is the reverse of attribute number three above which focused on the collection of data uncontaminated by required inferences. At times we want data that simply state what was seen and heard in the classroom. At other times it would be useful to gather information--interpretations--which illuminate the nature of the instructional tactics. For the unsophisticated individual, in particular, measures which would at least in part organize his perceptions regarding strengths and weaknesses in teaching would in certain situations be most useful (McNeil and Popham, 1973, pp. 238-239).

In selecting assessment procedures the influence of the nature of the competency statement in this process has been stressed. Some general points to consider in designing assessment are:

A. Objective instruments development vs. subjective.
B. Effect of assessment on process.
C. Selection of acceptable indices.
D. Establishing validity and reliability (Baird and Yorke, 1971, p. 5).

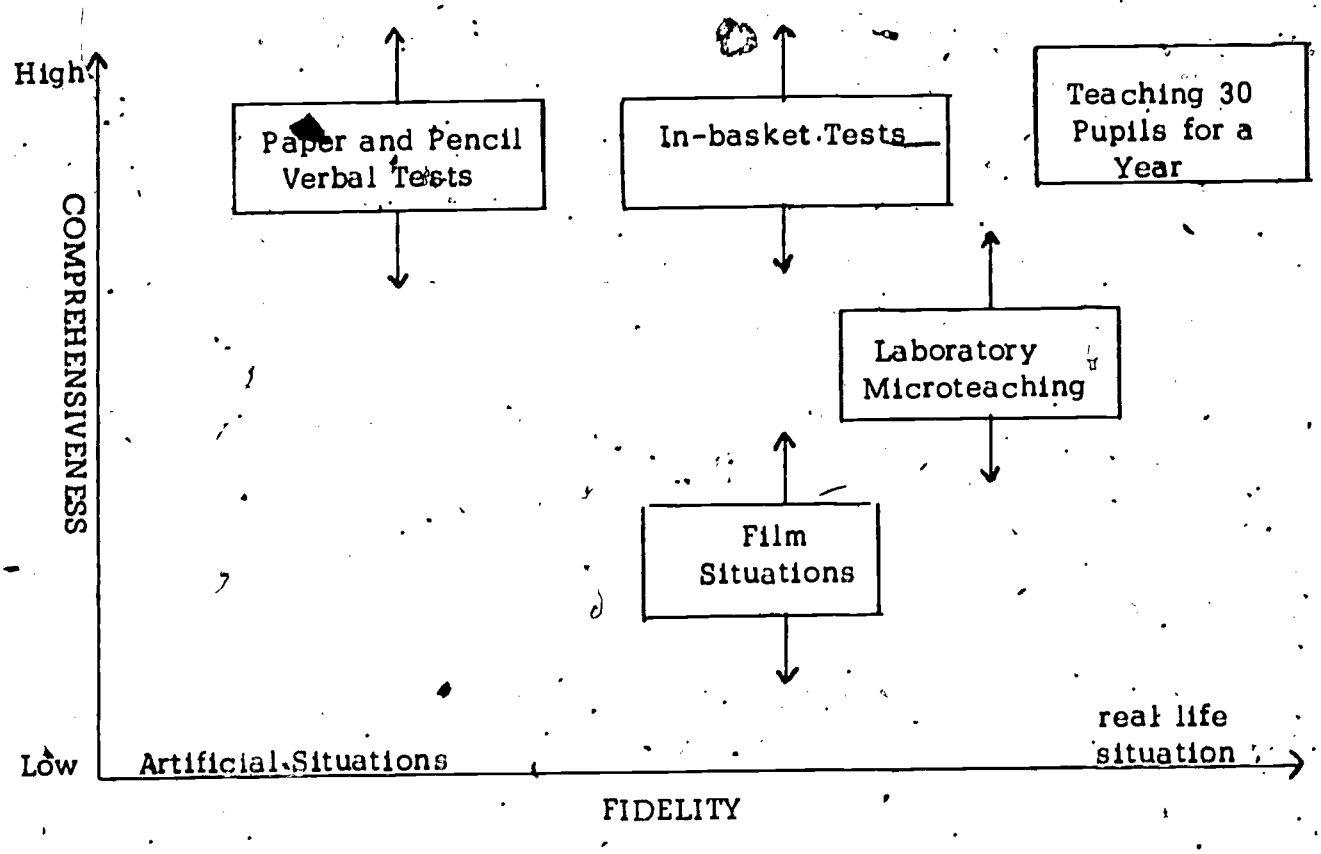
The validity and reliability of measurement instruments are, of course, important considerations. These will be discussed at length in appropriate sections, and therefore only briefly here. The evaluator should decide which of the validations are pertinent to the instrument being used. According to Young (1973) face validity is the most common form, and is concerned with the instrument agreeing with the mode of responding (written or verbal) whether it measures process or product, and the level of responding (memory, conceptualization, etc.). Also, content validity may be estimated by expert ratings of each item, construct validity is estimated by giving the evaluation to a group of persons possessing the trait and to a group not possessing the trait, and predictive validity is determined between levels of evaluations or in different time periods.

Some general recommendations regarding procedures are provided by the following:

> Actual data gathering techniques to evaluate knowledge and practice competencies are not complex. For knowledge competencies paper and pencil tests, oral examinations, and the like are appropriate. For practice competencies, studies of performance in classroom, microteaching, or other similar situations can be evaluated by one, or preferably more, judges on the basis of checklists, or overall performance (Airasian, 1974, p. 17).

> Stimulus and response modes could be specified from a number of available alternatives.. Prospective teachers might respond to paper and pencil directions in a videotaped mode--or vice-versa. Films, audiotapes and actual demonstrations are other possibilities. Responses need not be limited to overt teacher behavior as the only teacher "product" but could include products such as lesson plans, teacher-made tests, reports to parent, and other record-keeping and planning outcomes. And of course, one stimulus may produce a series of responses in various modes (Kay, 1974, p. 276).

An overview of some possible techniques has been developed in terms of two criteria, comprehensiveness and fidelity. Fidelity refers to the degree of realism of the test compared to the criterion situation.



Fidelity and Comprehensiveness of Different Types of Tests, (Quirk, 1974)

Another general concern, no matter what the competency domain or assessment technique, is that of utility. This asks of each data gathering effort whether the costs of time, money and effort can be justified by the extent to which they reduce risk fordecision makers. According to Merwin (1973) there are two ways to apply this criterion. One is to ask the extent to which the added information provided has reduced risks in selecting among alternatives, and the second involves comparing the costs of this particular means to getting the information with costs in using another means to the same information or equally predictive information highly correlated with it (e.g., indirect vs. direct assessment).

One means of viewing competencies and their assessment has been developed by Turner and should be mentioned at this point. His six criterion levels for evaluation are as follows:

Criterion Level 1. At the highest level, the criterion against which teachers (or teaching) might be appraised consists of two parts. The first part is observation of the acts or behaviors in which the teacher engages in the classroom. The observations must be conducted with a set of instruments which permit classification of teacher behaviors in both the cognitive and affective domains. The second part is systematic analysis of the level of outcomes achieved by the teacher with the pupils he teaches. Outcomes in both the cognitive and affective domains must be included. Because of variation in the entry behaviors of students and variations in teaching contexts, the residual outcomes in pupil behavior (the terminal behaviors corrected for entry behaviors and moderating variables) should be used as the criterion measures. To be placed at criterion level 1, the above two-part appraisal of teacher performance must be conducted over a relatively long period of time, probably at least two years (on a time sampling basis), with both the observational and residual pupil behavior components assessed during each of the years. The reason for the two-year period is that both teacher and pupil behavior are open to some random fluctuation and care must be taken to obtain a sufficient sample of behavior from both sources to assure fair conclusions.

Criterion Level 2. This criterion level is identical to criterion level 1 except that a shorter performance period is involved.

Criterion Level 3. This criterion level differs from criterion levels 1 and 2 in that pupil performance data are eliminated from the criterion. Judgments about competence or proficiency are thus based on the observable behaviors of the teacher rather than on the pupil outcomes associated with these behaviors.

Criterion Level 4. This criterion level differs from criterion level 3 in that both the teaching context and the range of teacher behavior observed are restricted. The context might be a typical microteaching context involving a few pupils or even peers acting as students. The teacher behavior observed would be restricted to a few categories in the cognitive or in the affective domain.

Criterion Level 5. This criterion level differs from criterion level 4 in that the teacher need not perform before live students (simulated students would be satisfactory). He must, however, be able to produce or show in his behavior at least one teaching skill; e.g., probing.

Criterion Level 6. This level differs from criterion level 5 in that the teacher need not engage in producing a performance, but rather, only show that he understands some behavior, concept, or principle germane to teaching (Turner, 1972, p. 3).

The relationship between these levels and assessment techniques will be identified at appropriate points throughout this paper.

A general overview of this relationship is provided by the following chart:

WHEN	WHAT	HOW
pre-practicum	Level 6--Trainee shows that he understands some behaviors, concepts, or principles germane to teaching--usually in a paper and pencil exercise.	paper and pencil tests; interviews
pre-practicum	Level 5--Trainee demonstrates his possession of teaching "skills", however, he need not do so with students. He may interact with case studies or other simulated materials.	case studies; simulation
pre-practicum	Level 4--Trainee demonstrates teaching behaviors in a micro-teaching context with a few students or peers.	microteaching; interaction analysis
practicum	Level 3--Trainee is judged on the basis of his ability to demonstrate "teaching behaviors" in the classroom.	videotape; observation forms; questioning pupils; interaction analysis
practicum and on the job	Level 2--Short-range outcomes achieved by the trainee with the pupils he teaches.	all tools used to assess public school pupils' growth (including above)

on the job	Level 1--Long-range outcomes achieved by the trainee (now a certified teacher) with the pupils he teaches.	all tools used to assess public school pupils' growth (including above)

(Baird and Yorke, 1971, p. 7)

Knowledge Assessment

Assessment of knowledge competencies generally can be accomplished through paper and pencil testing. This can easily be done in the preservice college classroom requiring very little in the way of special settings, instrumentation, or techniques. In addition, there are other ways of assessing knowledge, such as mediated stimulus-response techniques.

As an example, Okey and Humphreys (1974) suggest audio recordings of classroom discussions used to teach and assess the skill of identifying different types of teacher questions. Also, they suggest videotaping a classroom to teach and assess the ability to use reinforcement.

In another example, Popham (1974, p. 54) suggests alternative assessment approaches for the competency statement "Teachers must be able to both select and generate defensible instructional objectives." One procedure requires teachers to generate a set of measurable objectives, then have these judged by others using criteria of significance, suitability for learners, etc. Also, a teacher could select a specified number of objectives from a larger pool, and these could be judged according to established criteria. Popham further suggests that the teacher could describe, in an exam-type setting, alternative procedures for selecting and generating defensible objectives.

The knowledge category, you may recall, refers to facts, processes, theories, techniques, etc., encompassing a variety of cognitive processes. It has been noted by Dziuban and Esler (1974) that many learning tasks are inherently complex because of the interaction of their components and thus do not lend themselves to being dissected into very small parts. In structuring a laboratory problem, for instance, a student may have wide latitude in formulating hypotheses, structuring experimental procedures, and interpreting data.

Of all the assessment areas, the knowledge area is perhaps the most developed.

> For three-quarters of a century, decision makers of one kind or another have wanted to assess what candidates for teaching positions know. Measurement technology for

> for assessing academic knowledge thus became highly developed. Consequently, we now have widely available tests of knowledge of subject matter and of knowledge about teaching methods. (McDonald, 1974, p. 23)

In spite of this prodigious effort and its advanced status, there are a number of problems to consider, particularly when developing assessment for instructional units in teacher education programs. Since each module or course has its own objectives, existing tests of knowledge may not be applicable.

Also, in a program that has specific objectives and mastery levels, as competency-based programs purportedly do (particularly in the knowledge domain), the assessment is related to the specific objectives. Its purpose is to determine whether an individual has attained mastery of the objective as specified by a criterion level, not how he compares with a group of peers. This requires criterion-referenced testing.

In shifting to criterion-referenced testing, however, one encounters a problem in applying traditional psychometric characteristics of tests. Definitions of these characteristics, such as reliability and validity, involve assumptions not consistent with criterion-referenced tests.

> Many of these definitions involve equality of form and content among items as well as considerations of equivalent item difficulty. These characteristics produce instruments of extreme homogenity and low variance. Additionally, criterion referenced tests derive their meaning from the relationships they describe between the items and predetermined criteria (Dziuban and Esler, 1974, p. 4).

Another previously mentioned problem inherent in competency-based programs is the need to establish mastery levels for each of the competencies. There are several factors to consider in this process. Quirk (1972) states a number of cautions in using criterion scores, indicating they should take into consideration the number of test items per objective, the level of difficulty of these items, and a statement of the minimum performance level. Quirk also cites three factors related to setting cutoff scores to indicate "mastery" including 1) standard error of measurement, 2) the "x-percent correct" phenomenon and 3) the multiple cutoff model. Quirk notes that a test with low reliability would have a very large error of measurement in trying to estimate a score that represents "mastery." In referring to the "x-percent correct"

phenomenon, Quirk states that the percent of items that any given candidate answers correctly depends on the content of the items, and the difficulty level of the items in the test as well as his personal state during the test. If alternate forms of the test are to be used, the forms need to be equated statistically.

Some consideration has been given to describing a teacher candidate's overall ability by developing a competency profile, with competencies along the horizontal axis and degree of achievement along the vertical axis of a graph. It has been suggested that such an approach would assist employers in identifying better qualified teachers and those with skills which are particularly suited to their schools. This is a type of multiple cutoff or parallel stalk model as referred to by Quirk, and he has expressed some concerns. For example, if a candidate were to perform better on one objective than another, and the two objectives were highly correlated, the reliability of the difference scores would be quite low, even if the reliability of both measures were high. This same concern applies in evaluating the performance of the same candidate on two different objectives, or on the retesting of the same objective.

Also, according to Hills (1971), such scores can be set arbitrarily without adequate evidence on the validity of the variable that is being used for selection, as well as the validity of the available measure.

An additional concern in the area of reliability relates to retesting (this concern was cited earlier). Some competency-based programs are achievement rather than time based. Students progress as they complete competencies only, not by accomplishing as much as they can in a course restricted by time. Students are allowed to be retested until they achieve mastery. Also, for modules, pre and post tests are provided on each objective. Such situations require an examination of the reliability of the difference score. The reliability of this difference score is likely to be quite low.

Another set of considerations relate to behavioral objectives. In citing the long lists and number of behavioral objectives in competency-based programs, Quirk (1974) states the main measurement problem to be the reliability of the individual measures. Dividing the performance of a prospective teacher into finer elements could produce an unsatisfactory reliability figure. Also, according to Dziuban and Esler (1974) practical considerations often dictate testing competencies which are only indirectly related to the true goals of the behavioral objectives. This same discrepancy, however, has long been noted in norm referenced instruments.

Although Quirk has offered several criticisms of criterion-based testing in competency-based programs, disagreement with his arguments are also found in the literature. Cox (1974), for example, argues that many of the traditional measurement principles, such as the standard error of measurement, the reliability of the difference scores, and predictive validity, have been developed for norm-referenced tests and are probably not applicable to criterion-referenced measurement. A number of psychometricians have studied criterion-referenced test reliability (e.g., Livingston, 1974; Carver, 1974; Hambleton and Novick, 1973) and offered their analysis. According to Haladyna (1974) "each differs, and each suffers from a paucity of empirical studies either confirming or disconfirming the respective approaches."

Teacher Outputs

Previously in this paper teacher outputs were identified as possibly being a unique group of teacher competencies as opposed to being classified under the knowledge or performance category. A rationale for considering this area of teacher competence and its implications for measurement are provided by Morse, Smith, and Thomas (1972). Outputs, as they define them, represent primary, observable dimensions of teacher productivity, and serve as a bridge for connecting teacher behavior with learner outcomes.

> As a result of the performance of tasks various outputs will be produced. The outputs teachers produce are achievements for which they can be held directly accountable. They are defined as the sole means by which teachers perform their responsibilities toward learners. Teachers can control the outputs they produce, they can also predict with varying degrees of accuracy the effects their outputs are likely to have on learners. By distinguishing between teacher outputs and learner outcomes, one can give substance to the technical outcomes of teaching behaviors. This procedure emerges from and is consistent with the position that in order to nurture certain learner outcomes the teacher must do something. The things done include systematically using or developing materials, providing various experiences, and creating various climates or conditions thought to be conducive to learning. To that extent, it is these things: i.e., outputs, for which we can hold teaching behavior responsible or accountable. The teacher's responsibility includes assuring the relevance of those outputs to meeting the individual and collective needs of pupils.

> The kind, quantity, and quality of outputs that teachers produce can be measured. These measures constitute the basic data to be collected in any effort at assessing competence. The eventual linking of this data to data gathered about learner outcomes should provide a rich base of information from which to draw in making judgments about the competence of teachers (Morse, Smith, and Thomas, 1972, p. 11).

Performance Assessment

Teaching behaviors or performances require observation of the individual demonstrating the skill. This may be done by personal observation or use of recording equipment, with or without the utilization of systematic observation scales. Why evaluate teaching performance, why not deal with the ultimate criterion of effectiveness, pupil learning? Much more will be written on this in the section on pupil achievement, however, the following rationale has been noted in the literature.

> Measuring teacher effectiveness by measuring change in pupils is probably only feasible for simpler, lower level objectives.
>
> For the attainment of higher level objectives, or more slowly developing objectives, the more appropriate procedure appears to be to measure the behavior of the teacher and compare it to behavior which is thought to be related to the development of higher level objectives in pupils. Such a procedure appears feasible, both for the assessment of competence of individual teachers and for the certification of programs (Soar, 1973, p. 210).
>
> Similarly, the teacher appears to be more fairly evaluated if the judgment is made on what he does, rather than on the outcome of what he does. The first is under his control and the second is not (or at least not nearly so much so) (Soar, 1973, p. 209).

In reference to Turner's criteria, Merwin (1973, p. 12) notes that Turner's lower criterion levels involve assessing teaching behavior which is supposed to bring about a desired change in pupil behavior. He argues, however, that such a substitution can only be justified on the basis of a demonstrated reliable relationship between the assessed teacher behavior and change in pupil behavior that would be measured using the direct assessment approach. Currently, both traditional and competency-based programs must operate without such validation.

Some general concerns related to assessment of teacher performance have been identified by Merwin. He cites 1) error due to a lack of comparability in conditions under which the measure is taken; 2) errors in observing and recording behavior; and 3) inaccuracies in the matching of the observed behavior against the criterion behavior in attempting to arrive at the yes-no decision regarding achievement of competency (Merwin, 1973, p. 10).

As noted in the introductory pages, the selection of evaluation techniques depends partially upon the specificity of the competency. Examples of teacher competencies in the performance area may be useful at this point to illustrate some of the problems encountered due to level of specificity. The following examples were derived from The Florida Catalog of Teacher Competencies (Dodl, 1973).

1) Identify a student's instructional needs on basis of errors.

2) Involve students in teacher-pupil planning.

3) Structure opportunities to develop health and safety habits.

4) Help students develop attitudes compatible with society and self.

5) Cause student to perceive relevance of learning.

6) Use variety of media in course of teaching lesson or unit.

Merwin has analyzed these competencies and provided the following concerns.

> In the first example, "Identify a student's instructional needs on basis of errors," one assessor might well accept a simple oral questioning procedure while another might consider only careful classification of errors established on a theory of development as adequate. As evidence of "involving students in teacher-pupil planning" (example number 2) one judge might accept allowing students to say what they want to do, while another may feel that the observation is not complete until completion of what is jointly planned. The complexities, and alternative procedures that might be involved in

> determining whether a teacher has "caused" a student to perceive relevance of learning (example number 5) are almost unlimited. Whether what is needed to make these competency statements functional for directing measurement efforts is greater explicitness in the behavior to be observed, the need for adding criteria of acceptance, or both, it must be recognized that they do not provide an adequate base for designing assessments as they stand (Merwin, 1973, pp. 12,13).

> Even without the criteria statements needed to judge the adequacy of explicitness for unambiguously directing development of the measurement procedures to be used, a number of aspects of these statements pose assessment problems. For example, there are bound to be difficulties in designing procedures to determine the amount of "help" provided by a teacher in attempting to demonstrate his competency to help students develop attitudes compatible with society and self (number 4). The variety of media available and practical will vary widely from situation to situation in assessing a teacher's competency to use a variety of media (number 6) (Merwin, 1973, pp. 8,9).

The tenuous nature of criterion levels was examined in the preceding knowledge domain section, and these concerns apply to performance levels. Also cited earlier as a factor in the evaluation of teaching performance is the context called for in the competency statement. As one reads the above comments it is important to note that many of the concerns have relevance only within the context of an unstructured or uncontrolled (experimentally) environment such as a normal classroom. A very different context is provided by simulation situations where variables are controlled and the context is somewhat structured. This situation is analogous to Turner's levels four and five.

> Working under limited simulation procedures to assess teacher behavior during interaction with pupils as called for at level four allows more control of conditions, permitting greater objectivity and focus of observation of teacher performance at a cost of some realism. Level five simply provides further control of factors affecting the assessment of teacher performance at the cost of possibly a crucial element, use of live students. (Merwin, 1973, pp. 16,17).

In discussing performance assessment, a number of references to context have been made in the literature. Morse, Smith, and Thomas (1972) state that the nature of the context; i.e., the people who make decisions, the setting and the role being assumed, plays a crucial part in determining the way in which an individual will be judged as to competence. How

the competency is defined, the focus of investigation, the criteria and standards are all said to be a function of the context in which assessment is to take place.

Okey and Humphreys (1974) point out that performance outcomes are the doing skills of teaching, many of which require a classroom setting while they are learned and assessed. Furthermore, Garrison (1974) relates that, in his experience, in a program that defines competency in terms of the performance of teaching functions in an ongoing school setting the identification of the contexts in which competencies are to be demonstrated becomes as critical as the identification of the competencies themselves.

In addition to the context elements, referred to above, Merwin (1973) points to two other concerns related to context and performance assessment; namely, 1) the content under study and method of teaching, and 2) the background relative to the topic under study that the pupils bring to the learning experience. This latter concern as to the background of the pupils assumes that the task of the teacher will be different if the children are relatively homogeneous with few deficiencies, as opposed to a heterogeneous group of pupils, some having considerable deficiencies. Also the personal characteristics and attitudes toward school and learning of the pupils should be considered when evaluating the performance of the individual teacher.

Howell (1971) mentions a two-fold problem in terms of gathering data in evaluating performance. All factors likely to have major effects on the learning in question need to be described, as well as possible extraneous influences on pupil performance from which the data are obtained.

> The known sources of possible contamination can often be dealt with in designing the evaluation procedures, and unknown ones can be countered by sampling teaching performance generously and averaging results over a number of occasions or over many learners. But this may be expensive. The sample size, the sampling procedures, control over pupil situational variables to assure comparable conditions for the pre and post-learning performances, and recognition of interventions other than teaching--all are problems of the validity of the data, which are quite distinct from problems of the validity of the theoretical constructs or of the teaching purposes. . .(Howell, 1971, p. 21).

One competency-based teacher education program has described its approach to assessment which accounts for context.

> The approach taken to the measurement of individual teaching competencies was one of obtaining carefully delimited professional judgments, in the form of rating scale placements, as to the adequacy of a student's performance in a particular demonstration context. At least two separate professional judgments were obtained in relation to each competency demonstration, one from a student's college supervisor and one from his school supervisor. An evaluative judgment was also obtained from a content specialist if a student requested it. The ratings were designed so as to accommodate the impact of setting differences on competency demonstration (Garrison, 1974, pp. 65-66).

Merwin (1973) has cited several concerns related to assessing performance including difficulties in obtaining objective and reproducible observations, sampling problems involving elements of time, environmental factors surrounding the performance under observation, and characteristics of both the pupils and the type of learning involved.

Baird and Yorke have focused on problems of selecting context and the timing of assessment such as

> A. One setting, one time vs. many settings, many times.
> B. Early (in the day, week, semester, etc.) vs. late.
> C. Before (diagnostic) during (formative), or after (summative) instruction (Baird and York, 1971, p. 5).

The predictive validity of performance assessment, particularly in student teaching or similar type situations, is also a problem because the prediction of individual differences for future performance could be unreliable due to the limited range of performance observed. Yet it has been noted that

> ...what the student teacher does under a specific set of circumstances at a given point of time is of less concern than what the performance tells us about future performance--the validity of the assessment of predicting future effectiveness in helping pupils learn (Merwin, 1973, p. 22).

A note of caution, however, has also been provided. McDonald cautions that:

> We cannot treat teaching as if it were so different on each separate occasion that we can never evaluate it. The conflict between establishing reasonable expectations for teaching performance and the variety and complexity of the situations in which teaching occurs is one of the most important problems we have to solve. Until it is solved, our decisions about competence must necessarily be tentative (McDonald, 1974, p. 22).

A similar (or even synonymous) concern relates to sampling, and the relationship of an individual's performance at a given point in time to his actual ability to demonstrate a skill should he so choose. This relationship between "performance" and "competence" is, in a sense, a predictive validity issue affected by adequacies in sampling. Several writers have expressed concern over this issue;

> A major matter of concern revolves around sampling which will permit defensible generalizations (Merwin, 1973, p. 10).

> The extent to which evidence gathering situations permit students to manifest the behaviors inherent in the competencies is the extent which the evaluation is valid...Any testing situation provides only a sample of a student's behavior (Airasian, 1974, p. 17).

> A difficult problem associated with monitoring the activities assigned in a classroom is that of sampling. The drawing of reliable samples, of course, is a difficult problem regardless of the observational system that is being employed (Raths, 1973, pp. 20-22).

> There is the related problem of sampling. Does the absence of an item from a person's speech mean that he cannot produce it or merely that he has not found it necessary to produce it (Dill, 1974, p. 9).

Imbedded in this issue is the question of performance versus competence. Dill (1974) argues that teaching competence is not to be confused with teaching performance. Teaching performance is what the teacher actually does, and is based on knowledge of the instructional content and pedagogy as well as other factors such as memory, non-pedagogical knowledge and beliefs, distractions, fatigue, etc. In studying actual teaching performance one must consider a variety of factors and the underlying competence of the teacher is only one factor.

In the following comment teaching competence is viewed as only being observable in a very controlled situation where context variables have little influence. It should be noted again, however, that this assumes evaluation of "competence" as opposed to a specific "competency."

Evaluation of a teacher's competence in a student teaching situation requires accounting for a variety of factors, whereas evaluation of a specific competency in a microteaching session is less complicated and "performance" is more directly related to a competency.

> Only under idealized conditions can teaching behavior be taken to be a direct reflection of teaching competence. In actual fact, teaching performance cannot ever directly

reflect teaching competence. Observation of actual teaching behavior will show numerous false starts, deviations from plans, etc. Teaching competence, then is concerned with an ideal teacher, in a completely adequate classroom, who knows the pedagogy and content perfectly, and is unaffected by classroom conditions of crowding, inattention, distractions, etc. (Dill, 1974, pp. 29-30).

Howell (1971) has also distinguished between teaching competence, teaching competencies, and teaching performance in the following manner:

1) Teaching competence, as such, is not directly observable but is generally regarded as a more or less enduring personal characteristic;

2) A specific teaching competency, too, is presumed to be persistent and hence applicable to a whole series of situations within the limitations of its definition;

3) A teaching performance, however, is the observable manifestation of teaching competence, or competency, and is bound by time and place and other general situational variables, which define its setting or context (Howell, 1971, pp. 4-6).

Perhaps the most widely used method of assessing teacher performance is subjective rating, where an observer evaluates the teacher or trainee on the basis of his own criteria and interpretation of the situation. Problems with ratings again focus with the observer. Popham (1974) suggests that the difficulty may be due to different notions that raters (administrators, peers, students, etc.) have regarding what constitutes good teaching. Quirk (1972) suggests that one method used to avoid this problem, or at least modify it, is to train raters carefully on the definition of the items, show the raters examples of teacher behavior for each item, and check for the reliability of the ratings using actual classroom situations.

Another method of assessing teaching performance skills that has received considerable attention is the use of systematic observation techniques. Two systems are used to record behaviors, sign systems and category systems. The sign system uses a large number of behaviorally defined variables which are checked if they occur during a short; e.g., five minute, observation period. Category systems deal with fewer variables (categories) and are recorded continuously.

Many of the problems cited for teacher rating methods are eliminated or vitiated through the use of systematic observation instruments. By using systematic observation, the observer is made a recorder, insofar

as possible, rather than an evaluator (Soar, 1973). Also, according to Soar, this data tends to be "low inference" rather than "high inference" and stays closer to the original behavior. (It may be noted here, that although low inference measures stay with the behavior observed, higher inference measures appear to correlate better with some indices of student achievement; e.g., Rosenshine and Furst, 1971) This relates to the earlier discussion on the specificity of competency statements. Systematic observation techniques illustrate how general (broad) competency statements can be clarified by describing these in terms of several specific items. This appears to have several desirable effects when considering assessment. Some of these have been described by McNeil and Popham (1973). For example, instruments which require less inference from the observer have a greater agreement among users. Reliability is also enhanced when the dimensions are clearly defined and observers have had training, there is agreement on what is to be coded and there are fewer things for the observer to do during observation.

It is of interest to note

> Recent studies using ratings of intermediate levels of inference, such as "clarity" and "enthusiasm have produced more promising results than the earlier high inference ratings. However, before these results can be used maximally, the low inference behaviors which enter the ratings need to be identified (Soar, 1973, p. 208).

Merwin (1973) emphasizes that observation schedules must focus attention of the rater specifically on those aspects of performance relevant to the competency under judgment. Also, procedures for comparing the recorded behavior with the standards set for the competency must be clear and unambiguous. The degree of explicitness of the competency will be a large determiner of success in this process.

Although systematic observation techniques appear to have an advantage over rating systems, there are several factors to consider when utilizing such techniques. Reliability and validity are among these factors and have been treated in several ways in the literature. We will first consider validity.

It has been argued (McDonald, 1974; Abramson, 1971, among others) that measurement procedures used in the evaluation of teaching competency must have high validity. That is, there must be a demonstrated relationship between a teaching skill or performance and its effects upon students. However, in a number of studies that have attempted to relate pupil outcomes to classroom interaction variables, little relationship was found between pupil achievement and the observed

teacher classroom behavior. According to Abramson (1971) findings of such studies may result from incompatability of the achievement and observation data collected. Pupil achievement is collected on individuals, while observations are group data.

McDonald states, however, that the conclusion should not be drawn that we must defer the development of an evaluation system until all the relevant research has been done. He argues that there is already an abundance of ideas on pertinent teaching competencies which we can begin to measure--a necessary first step, and whose effect on teaching performance can be studied systematically as part of the process of developing evaluation systems (McDonald, 1974, p. 24).

Abramson (1971) considers the validity of observation systems in terms of content, concurrent, and construct validity. He defines these in the following manner:

1) Content validity is the degree to which the system provides information that is representative of the population of classroom behaviors that the system is meant to classify . . . It is essential that empirical evidence of the system's content validity be obtained . . .

2) The concurrent validity of two or more instruments is a function of the agreement between the measurements resulting from the application of these instruments. Typically, a new instrument is shown to be valid if the results obtained from its application are comparable to those obtained from a criterion measure, usually a more established instrument or a measurement with known validity. This validation process using two or more observation systems could also be followed providing the criterion against which the new instruments are to be validated is itself valid.

3) Construct validity is the degree to which the hypothesized outcomes of the practical application of the theory which gave rise to the instrument are borne out by the results of the appropriate experiments in which it has been used (Abramson, 1971, pp. 5-7).

According to Medley and Mitzel (1963), in order for an observational scale to be valid for measuring behavior, it must provide an accurate record of behaviors which actually occurred scored in such a way that the scores are reliable.

In addition

> The validity of measurements of behavior as the term is used here, depends then, on the fulfillment of three conditions: 1) representative sample of the behaviors to be measured must be observed. 2) An accurate record of the observed behaviors must be obtained. 3) The records must be scored as to faithfully reflect differences in behavior.
>
> The first condition would be fulfilled perfectly if the observed behaviors were a single random sample of the behaviors to be measured. Unfortunately, it is seldom feasible to obtain a random sample in practice, so it is necessary to use nonrandom samples which care to make them at least appear to be representative.
>
> The second condition--accurate record of behavior--and the third--meaningful scoring--are interdependent in the sense of how a record may be scored depends on how it is made. but they must be kept separate using a technique (Medley and Mitzel, 1963, p. 250).

Reliability has received more attention than validity in the literature, and also has been viewed from several perspectives. According to Abramson (1971) the reliability of assessment procedures needs to be established, reliability referring to replicability of the measurement and its underlying construct.

According to Quirk (1972) reliability is the sine qua non of the use of a measurement device. If the reliability of a performance or a judgment is low, the prediction of subsequent performance based on that measurement device is not likely to increase very much above chance level.

Abramson (1971) reviewed some of the literature that dealt with the reliability of observational measurements and concluded that there were essentially two major procedures normally used to establish the reliability of these data: 1) coefficients of observer agreement, and 2) an analysis of variance (ANOVA) technique first proposed and developed by Medley and Mitzel (1963). Most studies, including Flanders', have used the per cent agreement or Scott's (1955) coefficient of agreement between observers as their measure of reliability with fewer studies reporting reliabilities based on the ANOVA technique. According to Abramson, the coefficient of observer agreement and its variations may be thought of as roughly analogous to the test-retest or alternate forms reliability of most standardized tests because it provides a measure of comparability between two or more measurements of

samples drawn from a larger population of behaviors. However, the major advantage of the ANOVA technique, according to Abramson, results from its ability to partition the sources of variation inherent in the data into its component parts and thus yield error estimates as well as obtain estimates of true and total variance and calculate reliabilities using the classical definition $r=s^2true/s^2total$. It is thus possible to calculate reliabilities for the entire observation schedule and for the individual items which comprise it. These reliability coefficients and the error estimates for the different sources of variance may be extremely useful during the initial phases of item construction and revision because these data permit comparisons between the variances generated by items, observers, and teachers. Thus, through the ANOVA technique, it is possible to obtain inter-observer reliabilities as well as other useful information (Abramson, 1971, pp. 4-5).

Medley and Mitzel (1963) define reliability as the extent to which the average difference between two measurements independently obtained in the same classroom is smaller than the average difference between two measurements obtained in different classrooms. According to Medley and Mitzel unreliability can result from two measures of the same class differing too much due to the behaviors being unstable, lack of agreement among observers, different items lacking consistency, etc. It may also result from the differences between different classes being too small (Medley and Mitzel, 1963, p. 250).

Medley and Mitzel defined three terms useful in reliability determinations.

1) Reliability-coefficient refers to the correlation to be expected between scores based on observations made by different observers at different times.

2) Coefficient of observer agreement is the correlation between scores based on observations made by different observers at the same time.

3) Stability coefficient is a correlation between scores based on observations made by the same observer at different times.

Using these definitions, the following argument is presented:

> The true score pertains to the typical behavior that would be observed in a classroom over a period of time, only a sample of which is actually observed. Then a coefficient of observer agreement does not tell us how closely an obtained score may be expected to approximate a true score, because the two measures correlated are

> based on a single sample of behavior. The true score pertains also to the actual behavior which occurs, rather than to what some particular observer would see. Therefore, a stability coefficient does not estimate the accuracy of a score either, since it is based on a correlation between observations made by a single observer. The coefficient of observer agreement tells us something about the objectivity of an observational technique; the coefficient of stability tells us something about the consistency of the behavior from time to time. But only the reliability coefficient tells us how accurate our measurements are (Medley and Mitzel, 1963, p. 254).

Further study of reliability is provided by Brown et al.,

> Per cent or agreement between observers tells almost nothing about the accuracy of the scores obtained. It is entirely possible to find observers agreeing 99 per cent in recording behaviors on an instrument whose item or category consistency is very poor. Reliability can be low even though observer agreement is high for several reasons. For example, observers might be able to agree perfectly that a particular teaching practice occurred in a classroom, yet if that same practice occurs equally, or nearly so, in all classrooms, the reliability of that item as a measure of differences between teachers will be zero. Errors arising from variations in behavior from one situation or occasion to another can far outweight errors arising from failure of two observers to agree exactly in their records of the same behavior (Brown, Mendenhall, and Beaver, 1968, p. 4).

Although reliability and validity have received the most attention, there are a number of other concerns related to systematic observation techniques. McDonald (1974) suggests that we must develop information related to the reliability, validity, and the learnability of teaching skills. Also, according to McDonald, the information gathered must be uncontaminated by subjective biases and political processes, and the conditions of measurement must provide comparable information on groups of teachers. That is, the conditions under which teacher behavior is measured, must be standardized.

Flanders, whose work has been most influential in the development and utilization of systematic observation, has pointed out that choosing a particular system of interaction analysis tends to determine how one will conceptualize teaching (Flanders, 1974, p. 313).

Popham (1974) argues that assessment energy should be focused on the desired outcomes in learners, that is, assess the end results directly without encountering the measurement noise associated with the extra assessment step involved in systematic observation. However, the difficulties which are encountered using product criteria will be discussed in a later section.

Other problems identified by Popham (1974) are that

-deleterious factors may cancel out positive teacher behavior, and a manageable system could not pick up all negative process variables,

-observational approaches identify general classroom practices whereas teacher evaluation requires personal and particular decisions, and

-there is considerable danger that many teachers will "fake good."

Flanders has identified needed improvements in this approach to assessment. These are: the need for mathematical models to help guide the conceptualization of interactive phenomena and assist in establishing procedures for analyzing the data, attention to more effective methods of observer training and procedures for estimating the reliability of observation, and the development of multiple coding within a single time frame and analysis of longer chains (Flanders, 1974).

The preceding techniques have primarily been used to assess teaching performance competencies in actual classroom settings. Due to the variety of problems posed by context variables previously described, some assessment procedures have been devised for simulated situations. The reader may recall that the nature of the competency statement also determines whether or not live classrooms are required or if simulation is appropriate. A rationale for such an approach and some characteristics are provided in the following:

> Interaction skills are particularly difficult to measure. Attempts to do so with paper-and-pencil instruments have failed completely, mainly because no one has been able to devise test exercises which call for the kinds of abilities that determine success in face-to-face interactions--the ability to "read" behavior, relate it to professional knowledge, and react almost instantaneously, for instance.

Attempts to measure interaction skills directly--that is, by observing the teacher in action with a class--have been more successful, in the sense that it has been possible to identify some of these skills and to observe performances at various levels of skill. Such attempts must fail as measuring instruments, however, because it will never be possible to secure comparable samples of the behaviors of different teachers from which measurements can be derived. It has been impossible to confront any two teachers with the same problem (or equivalent ones) because no two pupils are alike--much less two classes--and no single pupil or class is the same after an experience as before it.

What is needed is a procedure for simulating the problems a teacher encounters when he interacts with a class, a procedure which can be duplicated over and over so that more than one teacher can be confronted with the identical problem.....
One approach that has been suggested and tried with limited success is to use a film or videotape recording of a class to simulate the real one. The strength of this method lies in the realistic stimuli it can present. When one sits or stands before the giant screen at Teaching Research in Oregon, where the Classroom Simulator was developed, and sees and hears the life-size, full color representation of a classroom before him, the approximation to confronting a live class is startlingly close. And when one intervenes--asks a pupil to stop doing something, perhaps--and the pupil responds appropriately, the effect is even more realistic.

Unfortunately, this does not always happen. Sometimes, the pupil's response is not so appropriate. Limitations of the equipment make it possible to offer only three alternative pupil responses per problem; and these three are not always perfectly synchronized. Nor can they include fully appropriate follow-up to all the wide variety of responses teachers might make. And, finally, each problem must be short in duration since only one intervention point can be provided.

Two basic problems confront us when we try to simulate classroom interaction. One has to do with the difficulty in constructing a model which can generate appropriate reactions no matter what the teacher response may be, and when it comes, providing pupil reactions which are lawful and predictable to all these possibilities. The other has to do with the difficulty of providing continuity because the number of alternate stimuli needed increases at a geometric rate each time the teacher responds, and each alternative has to be worked out in advance, filmed, and programmed (Medley, 1969, pp. 4-5).

McDonald has also offered some alternative simulation assessment strategies and is optimistic about their use. One type is a filmed simulation test that portrays a teacher conducting a class. The film is stopped periodically and the viewer is asked to say what he or she would do in this situation; in other parts of the test the viewer is asked to explain what is occurring in the class, and, in some places, he is asked what advice or suggestions he would give to the teacher. Another involves the teacher arranging the subject matter in the form of presentations or questions, and the experimenter responds whenver the teacher asks a question. This gamelike situation does discriminate sharply between deductive and inductive teaching styles (McDonald, 1974, p. 5).

The use of performance tests, such as those used in simulated situations, have also raised several measurement concerns. According to Quirk (1971) compared to the more popular paper-and-pencil multiple-choice tests, performance tests are much more complicated to administer, usually test only one individual at a time, require special training for the observers, are more difficult to score reliably, and are more expensive to administer and to score in terms of personnel time, equipment, and facilities. Test security is also a serious problem (Quirk, 1971, pp. 10-11).

Quirk (1974, p. 317) also cites what he calls a host of critically important research questions about microteaching tests or other simulated tests. For example, how consistent is the teacher's behavior over time? What is the effect of familiar versus unfamiliar pupils on the behavior of the teacher? What is the effect of pupil practice on the teacher? How is teacher behavior related to pupil learning? What are the correlations between simulated teaching tests and paper-and-pencil tests? So far, he asserts, these questions far outnumber the adequate answers.

Affective Assessment

The affective area is difffcult to assess, and this is usually not subject to formal evaluation in teacher education programs. A variety of procedures for developing affective domain competencies, however, have been developed and objectives of these activities have been established. Competencies stated in this area must be evaluated, but due to the nature of the area, competencies may be stated in broad terms and unique kinds of assessment strategies, such as unobtrusive measures and long term data, may be required.

Some general and somewhat "social" concerns voiced by Airasian (1974) include the question as to whom the judgments about a given student's values, personality, interests, and preferences be disseminated, in what form, with what guidelines, and for how long? DeMarte et al. (1975) report that

> Since there are no right answers to emotions, attitudes, or feelings, and because human beings tend to "second guess" experimenters, the accuracy of any affective assessment can be questioned, particularly paper and pencil instruments. Given this state of the art, affective instruments must be used with caution in teacher education (DeMarte et al. 1975, p. 2).

As indicated in the early pages of this paper, some would consider this competency area to have two parts, teacher personality characteristics and teaching behaviors in the affective domain. It is possible, it may be argued, that a teacher can and does demonstrate sensitivity to students' needs, utilize students' ideas, and accept their feelings, and yet does not possess the personality characteristics of warmth, sensitivity, or empathy. He may demonstrate the affective teaching skills because he has been trained to do so and believes it is a good teaching technique. Whether or not this is an acceptable dichotomy is, of course, a moot point, but these two components will nevertheless be examined here.

In terms of personality characteristics, it has been noted by Getzels and Jackson (1963) that very little is known for certain about the nature and measurement of teaching personality, or about the relation between teacher personality and teaching effectiveness.

Some approaches to assessment of personality are described by Sandefur (1970) such as the Minnesota Teacher Attitude Inventory (MTAI), the California F Scale, and the MMPI. The fakeability of such tests has been noted as a potential source of error, particularly when one can readily discern a preferred direction to fake, as on the MTAI.

In measuring noncognitive variables such as attitudes, several researchers have turned to attitude questionnaires, similar to those above. An example of a scale of this type was developed by Bogardus (1925) to measure social distance, or the closeness of the relationship to which the respondent is willing to admit members of designated social groups. Bogardus regarded degree of acceptance in terms of whether or not individuals would accept others: (1) to close kinship by marriage (2) to my club as personal chums, (3) to my street as neighbors, (4) to employment in my occupation of my country, (5) to citizenship in my country, (6) as visitors only to my country, and (7) would exclude from my country. A general tolerance score is obtained by averaging the step values (ranging from one to seven) assigned by the respondent to each of the groups he rated. Stern (1963) analyzed this type of attitude assessment and noted four issues when items are assembled and keyed arbitrarily in accordance with the opinions of the investigator:

1) Are all items relevant to the same measurement continuum?

2) Are the items in fact ordered as steps along that continuum?

3) Is the relative distance between the steps constant?

4) Are the responses actually a function of the attitude the items were intended to sample, rather than of some irrelevant process (Stern, 1963, p. 405).

Assessment of affective teaching competencies is also not very encouraging.

> Unquestionably the state of the art of affective assessment lags behind cognitive or psychomotor assessment. In the end, interpretive judgments based upon both formal and informal observations and discussions will probably provide the optimum means of gathering affective evaluative data about student progress. The lack of objectivity associated with such techniques in comparison to more formal paper-and-pencil techniques should not deter evaluation. One method of stressing the importance of affective aims is to diagnose and evaluate them (Airasian, 1974, pp. 17-18).

Among the devices used for assessment in this area are: systematic observation techniques (previously described) self-response questionnaires, Q-sort techniques, the semantic differential, and rating scales.

In terms of rating scales it has been noted that

> . . . the measuring device is not the paper form but rather the individual rater. Hence a rating scale differs in important respects from other paper-and-pencil devices. In addition to any limitations imposed by the form itself, ratings are limited by the characteristics of the human rater--his inevitably selective perception, memory, and forgetting, his lack of sensitivity to what may be psychologically and socially important, his inaccuracies of observation and, in the case of self-ratings, the well established tendency to put his best foot forward, to perceive himself in a more favorable perspective than others do (Remmers, 1963, p. 329).

Rating scales can be evaluated on the basis of the following criteria

1) Objectivity. Use of the instrument should yeild verifiable, reproducible data not a function of the peculiar characteristics of the rater.

2) Reliability. It should yield the same values, within the limits of allowable error, under the same set of conditions. Since basically, in ratings, the rater and not the record of his response, is the instrument, this criterion boils down to the accuracy of observations by the rater.

3) Sensitivity. It should yield as fine as distinctions as are typically made in communicating about the object of investigation.

4) Validity. Its content, in this case the categories in the rating scale, should be relevant to a defined area of investigation and to some relevant behavioral science construct; if possible, the data should be covariant with some other, experimentally independent index. These requirements correspond to the concepts of definitional, construct, concurrent and predictive validity (American Psychological Association, et al., 1954).

5) Utility. It should efficiently yield information relevant to contemporary theoretical and practical issues; i.e., it should not be so cumbersome and laborious as to preclude collection of data at a reasonable rate (Remmers, 1963, p. 330).

Guilford has categorized rating scales into given major groups: graphic, standard, accumulated points, and forced-choice. He also pointed out that any such classification is a very loose one, based on shifting principles (Guilford, 1954, pp. 263-301).

As in the other measurement devices considered in previous sections of this paper, reliability and validity must be considered.

Remmers (1963) states that using reliability statistics for sociometric data may be relatively meaningless and even misleading. For example, in test-retest coefficients there is a problem of distinguishing between effects of memory and those of real change. If there is too short an interval between testing, memory may play an important part in increasing consistency of responses, whereas if the interval is too long, there may be real changes in group structure, thus lowering reliability coefficients.

In terms of validity, there are also fundamental differences between psychometric tests and sociometric tests. That is, in a psychometrically derived test we try to measure some trait by eliciting some related responses. In a sociometric test the behavior is actually sampled. In effect, the predictor is the same as the criterion, as long as we are not interested in drawing inferences from the behavior observed (Remmers, 1963).

Also, there are human bias factors in rating scales. These include such things as 1) opportunity bias due to time sampling problems, 2) experience bias, that is the behavior patterns may differ between those of an experienced teacher and a practice teacher, 3) criterion distortion which is error built into a rating scale by including several correlated behaviors, thus weighing the behavior disproportionately, and 4) rating biases due to various response sets (Brodgen and Taylor, 1950; Remmers, 1963).

The areas of attitude, personality, and affective domain competencies are much more comprehensive than this analysis can provide, but the concerns raised in this section are indicative of the problems involved in assessment of this competency domain. Two references which provide an inventory and analysis of existing instruments are as follows:

DeMarte, Patrick; Johnson, Donald; Molenkamp, Alice, "Report on the Affective Dimension in Teacher Education," Rochester Area Colleges, Rochester, New York, 1975.

Beatty, Walcott, Improving Educational Assessment and An Inventory of Measures of Affective Behavior, Association for Supervision and Curriculum Development, NEA, Washington, D.C., October 1969.

Product Assessment

Consequence objectives may be the most interesting and controversial of the competencies. These require the teacher trainee to produce changes in students, usually achievement gains. The focus of assessment in this situation is primarily on the students who are being instructed by the teacher. Two different areas of focus include student achievement and the activity a student engages in. An example of the latter is "students being attentive to class activities." Teacher competencies and assessment approaches to this area have been described by Hatfield (1974). He notes that competencies relating to students being attentive in class may include use of designated conference techniques, techniques for controlling disruptive behavior of students, and managing overall activities in the classroom. In evaluating these types of teacher competencies at the performance level, two approaches could be used: 1) to see if the teacher actually used the techniques, and 2) to see if the teacher, in fact, achieves the purposes of the technique. The teacher is evaluated not just for using the technique but on whether the student is actually confronted in a meaningful way and responds to that confrontation (Hatfield, 1974).

Problems related to evaluation of teacher performance described in 1) above are discussed in the section on performance assessment. In 2) the teacher is evaluated on the basis of whether the student "responds to that confrontation," or "if the student actually becomes attentive to the activities." If the determination of this is left to the judgment of the observer, the problems of observation techniques as previously discussed

must be considered. These include such factors as "halo effect" and other response sets, reliability of observers, and sampling concerns, among others. If an attempt is made to make the determination more objective, then there is a problem of establishing a criterion level; e.g., how many students must be attentive. It would appear that the more subjective approach utilizing professional judgment is the more viable approach at this time, but is nevertheless not appealing from a measurement point of view.

Medley, Soar, and Soar (1975) believe that assessing teacher competence on the basis of pupil behavior is not appropriate. Among their concerns is one of morality, that one human being's advancement is dependent on the behavior of another (the pupil), which is not and should not be entirely under his/her control.

Evaluating teacher performance utilizing student achievement (as measured by test scores) as the criterion of effectiveness has also received attention in competency-based education programs. Medley, Soar, and Soar (1975), however, contend that evaluation of teaching through evaluation of pupil outcomes is not a viable strategy. Several problems have been cited, and again reference can be made to Turner's criteria.

> Using changes in pupil behavior over a long period (Turner's Level 1) or shorter period (Turner's Level 2) as the measure of performance of a teacher candidate to make the "go-no go" decision on development of a competency poses several complexities in addition to those set forth above. They include the need to state the competency in terms of pupil behavior, assessment in terms of a change in behavior based on a minimum of two observations (before and after intervention by the teacher), observing and recording performance relevant to the teacher competency under consideration, and most problematic of all, accurately identifying the teacher's contribution to the change observed (Merwin, 1973, p. 13).

Airasian (1974) states that the data which must be gathered to evaluate teacher's effects upon student learning are not at all clear. Research indicates that a large portion of the variance in student ability and achievement is attributable to early environmental factors. Also,

> The attribution of causation aspect offers an even greater challenge if the competencies are written in terms of ability to bring about change in pupils, the process must involve separation of those changes attributable to the teacher's efforts from those that cannot be so attributed. Children's learnings are affected by interactions with other children, the

> extent to which their parents are interested and become involved in what they learn, what they see on TV, how the school is organized, the scheduling of their time by others, and a host of other factors. Since these factors will impinge on different pupils in different ways, one can hardly say that one teacher has demonstrated "competency" and another has not simply on the basis of changes in the performance of their two groups of pupils (Merwin, 1973, p. 14).

Okey and Humphreys add that little is known about how to adjust expectations of teacher success when they work with pupils that have different entering abilities, backgrounds, aptitudes, motivation, and learning rates. Differences in subject matter difficulty, instructional materials, and classroom settings may also have important effects on pupil achievement, and therefore, teacher consequence measures (Okey and Humphreys, 1974, p. 8).

According to Flanders (1974) one difficulty with measures of learning is the overemphasis on subject matter achievement. Flanders suggests that using a test of subject matter as the only criterion of learning is inadequate, because student learning includes much more. For example, staying in school and not dropping out, learning to like school and the process of learning, gradually learning how to be more self-directing and independent, learning how to make moral and ethical judgments, etc., may be more important measures of teaching than are scores on content tests. Also, given a focus of subject matter and a research design consisting of pretest, teaching-learning, and posttest, it was found that posttest achievement is much more strongly associated with pretest scores (at least ten times more) than it is with any measure of teaching. This is due to the pretest to posttest gain being mainly a function of ability, and therefore in any assessment of teaching, student ability would have to be controlled more thoroughly. Also, standardized achievement tests are designed to be insensitive to the influence of a particular teacher and reflect, instead, the total developmental background of the student. In summary, Flanders states that conclusions are not really about teaching effectiveness; instead, they are about student effectiveness (Flanders, 1974, p. 312).

At the other end of the spectrum are the measures of pupil outcomes, particularly the criteria used to assess these. Abramson (1971) in discussing product criteria as a measure of teaching performance, points out the problem of the ultimacy of the criteria. For example, does effective teaching reflect gain in immediate factual knowledge, or improved skills of an intermediate nature, or ability to apply these facts and skills, or the more comprehensive "success" in life types of skills (Abramson, 1971, p. 2).

Also, Soar (1973) argues that attempts to measure teacher competence through pupil gain in higher level objectives appears to be exceedingly difficult and probably impossible in many cases. McNeil and Popham (1973) cite technical problems in assessing learner growth such as concerns about the adequacy of measures for assessing a wide range of pupil attitudes and achievement at different educational levels and in diverse subject-matter areas, failure to account for instructional variables that the teacher does not control, and the unreliability in the results of teacher behavior, that is, inconsistent progress of pupils under the same teacher.

Further problems which are related to the analysis and interpretation of learning scores, according to Stake (1973) include: grade-equivalent scores, the "learning calendar," the unreliability of gain scores, and regression effects. Instructional specialists (Hively, Patterson, and Page, 1968), according to Stake, have questioned the appropriateness of grade equivalents or any other "norm referencing" for interpreting items. They object to defining performance primarily by indicating who else performs as well. That is, the items on all standardized tests have been selected on the basis of their ability to discriminate between the more and less sophisticated students rather than to distinguish whether or not a person has mastered his task, indicating successful attainment of the instructional objectives. Grade equivalents are too gross to measure individual short-term learning (Lennon, 1971; Stake, 1973).

In terms of the learning year, there is some basis for miscalculations. For example, winter is a time for most rapid academic advancement, summer the least. Also, there is a common belief that schooling should not aim at terminal performance, but rather at continuing performance in the weeks and months and years that follow.

Concern with the unreliability of gain scores can be viewed in the manner described by Quirk (1972), or by Stake (1973). Consider for example, using a typical standardized achievement test with two parallel forms, A and B; each having a reliability of +.84. Their correlation (that is, the correlation of parallel forms Test A with Test B) in his example was +.81. And, in using a standard formula (Thorndike and Hagen, 1969) the reliability of gain scores (A-B or B-A) would be +.16. Using the raw score and grade equivalent standard deviations from the test's technical manual, assuming 9.5 items and 2.7 years respectively, on the average, a student's raw score would be in error by 2.5 times, his grade equivalent score would be in error by .72 years, and his grade equivalent gain score would be in error by 1.01 years (Stake, 1973, p. 215).

Regression effects; i.e., initially low scores tend to move up toward the mean while initially high scores tend to drop rather than gain, have also caused some misinterpretation of the effects of instruction. Lord (1963) discussed this universal phenomenon and various ways to set up a proper correction for it.

In spite of these concerns, the evaluation of teacher performance utilizing student achievement as the criterion of effectiveness has received considerable attention in competency-based teacher education programs. One such attempt is the utilization of microteaching and the development of teaching tests. McDonald argues that by stipulating the objective, providing the teaching materials, and controlling the variability of the pupils, the degree of the teacher's skill may be assessed. Also, a teacher's skill can be assessed under a variety of different teaching conditions using microteaching sessions. However, this approach still has several limitations such as relatively short lessons and a small number of students used. Therefore, McDonald and others developed a mini-course format to use for more complex teaching situations. The results of his analyses of these teaching performances indicate that the microteaching performances are relatively poor predictors of the teaching performances in the mini-courses. He concludes, however, that the microteaching is more useful for assessing the degree to which a teacher has basic skills, whereas the mini-course is most useful in assessing how teachers integrate these skills into complex teaching performances.

In discussing student teaching and internship experiences, McDonald suggests that these can be used to assess daily performance under uncontrolled conditions. They are useful for providing information on what teachers are likely to do in contrast to what they are able to do. Also, on-the-job observation can be used to assess such factors as teaching style (McDonald, 1974, p. 24).

In student teaching, some writers (e.g., Okey and Humphreys, 1974) suggest applying consequence objectives via criterion referencing.

A number of concerns have been directly related to the teaching test approach. For example, teaching performance tests may have insufficient reliability to permit their effective use in teacher evaluation (Glass, 1972). Medley, Soar, and Soar point out that teaching tests can only measure how effective a teacher is in achieving short-term goals, which are the least important goals of education. Also, they point out, stability coefficients (which describe correlations between mean gain scores of two classes taught

by the same teacher) are around .3, certainly not acceptable. Milman (1973) suggests that with more reliable measures utilizing more items, collected on larger student groups, after longer instructional sessions, such teaching performance tests will be a more reliable indicator of teaching effectiveness.

In concluding this section, Airasian's comment appears appropriate.

> In sum, while it is always possible to evalute teaching competency by measuring student learning, the issues remaining to be settled before such evaluation can be undertaken in an intelligent manner, fair to both teachers and students, suggests that student learning measures not be used to evaluate individual teachers at present (Airasian, 1974, p. 19).

Experiences Assessment

Expressive objectives have no pre-determined outcomes, they require only the experiencing of certain activities. In this case it may be necessary only to evaluate whether or not one has indeed participated in the experience. A check list of necessary activities is one means of assessing whether or not the individual has participated appropriately. In those cases where observation of the activity does not occur, other kinds of evidence may be required, such as diaries, descriptions, or testimonials that the individual was present. Since this domain requires little data, it is the easiest to "assess" but also yields information of a less rigorous nature.

Summary

Competency-based teacher education has been defined in various ways but there is general agreement on at least two basic elements. The first essential characteristic is the specification of teacher competencies which form the basis of the entire program. The second is the design of assessment techniques directly related to the specified competencies.

Competencies have been written in a variety of ways and have been related to various domains or competency areas. The competency domains identified in the literature are knowledge, behaviors (performance), attitudes, consequences, and experiences. There also

seems to be a variety of viewpoints as to how competencies should be written. One approach is to write them as general statements of behavior with some broadly defined expected level of achievement. Another approach is to develop specific performance objectives derived from the competency statement. Competency statements may also be written as behavioral objectives. In each of the competency domains cited the form of the competency statement must be examined to determine appropriate assessment techniques.

There are a number of assessment factors in general which need to be considered in the evaluation of competencies. The nature of the standards, that is, criterion selection, is an essential aspect. Other concerns are comprehensiveness and fidelity of the assessment system, validity and reliability of data, and general utility of the process. In addition, Turner has provided six criterion levels for competency evaluation which provide a framework for identification of assessment areas.

Assessment of knowledge competencies generally can be accomplished through paper-and-pencil testing. Of all the assessment areas the knowledge domain is the most developed. Inherent in this process is criterion-referenced testing. A problem one then encounters is the application of traditional psychometric characteristics of tests. The setting of criterion levels also has many difficulties associated with it.

Teacher outputs were identified as possibly being a unique group of teacher competencies as opposed to being classified under the knowledge or performance categories. Outputs represent primarily observable dimensions of teacher productivity and serve as a bridge for connecting teacher behavior with learner outcomes.

Assessment of teaching behaviors or performances requires observation of the individual demonstrating the skill. This may be accomplished by rating scales or structured observation systems (systematic observation scales). It has been argued that teaching performance rather than pupil learning should be the focus of assessment because measuring teacher effectiveness by measuring change in pupils is probably only possible for simpler lower level objectives. Assessing teacher performance deals only with the lower levels of Turner's criteria. Problems encountered in this competency area relate to establishment of criterion levels, comparability of conditions, and observation errors.

Other elements to be considered in performance assessment are the nature of the content being taught, the background of the pupils being taught, and general effects on learning which may not be accounted for. An extremely important consideration is the context of performance assessment. How the competency is defined, the focus of investigation, and the criteria are all said to be a function of the context in which assessment

is to take place. It has also been stated that the identification of the context in which competencies are to be demonstrated becomes as critical as the identification of the competencies themselves. Some competencies may be demonstrated under simulated conditions while others require a classroom setting.

One aspect which received considerable attention is that of sampling, and thus the relationship of an individual's performance at a given point in time to his actual ability to demonstrate a competency should he so choose. The predictive relationship between performance and competence is affected by adequacies in sampling.

Perhaps the most widely used method of assessing teacher performance is subjective rating where an observer evaluates the candidate through observation and possibly through the use of some type of checklist. One method of assessing teacher performance that has received considerable attention is the use of systematic evaluation techniques. The importance of the specificity of the competency statement is evident in the use of systematic observation techniques. The more specific the competency statement, the lower the inference level in arriving at evaluation decisions.

Two important considerations in the use of systematic observation scales are validity and reliability. Content, concurrent, and construct validity are areas which must be accounted for. Reliability has been identified as the essential element in the use of a measurement device, and a variety of reliability perspectives have been described. Coefficients of observer agreement and analysis of variance have been used to determine reliability. Three aspects of reliability are the reliability coefficient, the coefficient of observer agreement, and the stability coefficient. Other problems for consideration are standardized conditions of observations, deleterious effects on teacher behavior, and fakeability under such conditions.

Simulation is one approach that has been suggested and tried in various means. Many extraneous variables are controlled in such situations but there is a concomitant loss of test fidelity, although this is much more realistic than paper-and-pencil testing.

The area of attitudes is difficult to assess and is usually not subject to formal evaluation in teacher education programs. A distinction has been made between personality characteristics and affective competences. Approaches to assessment of personality are primarily projective techniques. Instruments which have been utilized more frequently are the Minnesota Teacher Attitude Inventory, the California F Scale, and the MMPI. The fakeability of such tests has been noted as a potential source of error. Among the devices used for assessment for affective teaching competencies are systematic observation techniques, self response questionnaires, Q-sort techniques, the semantic differential, and rating scales.

A major concern implied throughout the paper is the need to examine the feasibility of assessment of a given domain prior to making a decision as to whether or not competencies should be written for that area and in what form. Although assessment in the attitude domain is faced with a variety of problems, it would be dangerous for a program to exclude competencies in this area because they cannot be readily assessed.

Consequence objectives require the teacher trainee to produce changes in the student, usually achievement gains, although the activity a student engages in is another possible criterion. In evaluating activities students engage in, a number of problems are encountered such as the causal relationship between teacher performance and student activites, observation problems such as halo effect, sampling concerns, and others.

Evaluating teacher performance utilizing student achievement also has a number of serious problems. Some research indicated that a large portion of the variance in student ability and achievement is attributable to early environmental factors. Other concerns are the ultimacy of the criteria, adequacy of measures for assessing pupil gains at different levels, and in different areas, reliability of gain scores, and regression effects. It has been concluded that student learning measures cannot be fairly used to evaluate individual teachers at present.

Expressive objectives do not have pre-determined outcomes, they require only the experiencing of certain activities. Instruments used in this domain include checklists, descriptive reports, anecdotal records, etc. Since this domain requires little data it is the easiest to "assess" but also yields information of a less rigorous nature.

Epilogue

In analyzing assessment problems related to teacher competencies, the author has attempted to synthesize the diverse opinions on a variety of assessment concerns found in the educational literature. There may be some areas of importance, however, which have been omitted or have not been given appropriate depth of treatment. It is also possible that conflicting or alternative viewpoints on certain aspects have not been presented. The author is interested in any information which would clarify or otherwise contribute to this paper, and would welcome readers to send their comments.

REFERENCES

Abramson, T. "Performance-Based Certification and Observation Techniques," Division of Teacher Education, the City University of New York, April 1971.

Airasian, P. W. "Performance-Based Teacher Education: Evaluation Issues," Performance Education-Assessment, Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 12-20.

American Psychological Association, et al. Joint Committee, Technical Recommendations for Psychological Tests and Diagnostic Techniques, Psychol. Bull., 1954, 51, Supp.

Angoff, W. H. "Scales, Norms, and Equivalent Scores," in Educational Measurement, (Second Edition), by R. L. Thorndike(Ed.), Washington: American Council on Education, 1971, pp. 508-600.

Baird, J. H., and York, D. B. "Performance Based Teacher Assessment," Brigham Young University, 1971.

Bogardus, E. S. "Measuring Social Distances," Journal of Applied Sociology, 1925, 9, 299-308.

Brogden, H. E., and Taylor, E. K. "The Theory and Classification of Criterion Bias", Educ. Psychol. Measmt., 1950, 10, pp. 159-186.

Brown, B. B., Mendenhall, W., and Beaver, R. "The Reliability of Observations of Teachers' Classroom Behavior," The Journal of Experimental Education, Vol. 36, No. 3, Spring 1968, pp.1-10.

Burns, R. W. "Achievement Testing in Competency-Based Education," Educational Technology, November 1972, pp. 39-42.

Carver, R. P. "Two Dimensions of Tests, Psychometric and Edumetric," American Psychologist, July 1974, pp. 512-518.

Cox, R. C. "Confusion Between Norm-Referenced and Criterion-Referenced Measurement," Phi Delta Kappan, Vol. LV, No. 5, January 1974, p. 319.

DeMarte, P.J.; Johnson, D.H.; Molenkamp, A.D., "Report on the Affective Dimension in Teacher Education," Rochester Area Colleges, 1975.

Dill, N. L. "A Theoretical Reformulation of the Concepts of Competence and Performance in Teacher Education," paper prepared for the annual meeting of the American Educational Research Association, Chicago, April 15-19, 1974.

Dodl, N., Director. The Florida Catalog of Teacher Competencies, (First Edition), Florida Department of Education, Division of Elementary and Secondary Education, Florida Educational Research and Development Program, Tallahassee, Florida, January 1, 1973.

Dressel, P. L., and Mayhew, L.B. General Education: Exploration in Evaluation, Washington, D.C.: American Council on Education, 1954.

Dzuiban, C.D., and Esler, W. K. Florida Technological University, "Criterion Referenced Tests: Some Advantages and Disadvantages," presented at the annual meeting of the American Educational Research Association, Chicago, Illinois, April 15-19, 1974.

Flanders, N.A. "The Changing Base of Performance-Based Teaching," Phi Delta Kappan, Vol. LV, No. 5, January 1974, pp. 312-315.

Getzels, J. W., and Jackson, P.W. "The Teacher's Personality and Characteristics," Handbook of Research on Teaching, N. L. Gage, Editor, Rand McNally and Company, AERA, 1963, pp. 506-582.

Ghiselli, E.E. Theory of Psychological Measurement (New York: McGraw-Hill, 1964).

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions," American Psychologist, Vol. 18, 1963, pp. 519-521.

Glass, G. V. "Statistical and Measurement Problems in Implementing the Stull Act," Stanford University Invitational Conference on the Stull Act, October 1972, Palo Alto, California.

Guilford, J. P. Psychometric Methods, (Second Edition), (New York: McGraw-Hill, 1954).

Haladyna, T. "An Investigation of Sub Scale and Test Reliability of Criterion Referenced Tests," Southern Illinois University at Carbondale, AERA paper, Chicago, April 15-19, 1974.

Hambleton, R. K. and Novick, M. R. "Toward an Integration of Theory and Method for Criterion-Referenced Tests," Journal of Educational Measurement, 1973, 10, pp. 159-170. (Also published as ACT Research Report, No. 53., Iowa City, Iowa: American College Testing Program, 1972).

Hatfield, R. C. "Evaluation of Teacher Competence Based on Pupil Behaviors," Competency-Based Teacher Education: A Potpourri of Perspectives, Association of Teacher Educators, Bulletin 38, 1974, pp. 40-44.

Hills, J. R. "Use of Measurement in Selection and Placement," in Thorndike, Educational Measurement, (Second Edition), edited by R. L. Thorndike, Washington, D.C.: American Council on Education, 1971, pp. 680-732.

Hively, W.; Patterson, H. L.; & Page, S. H. "A 'Universe Defined' System of Arithmetic Achievement Tests," Journal of Education Measurement, 5, No. 4, 1968: 275-290.

Howell, J. J. "Performance Evaluation in Relation to Teacher Education and Teacher Certification," Division of Teacher Education, the City University of New York, April 1971.

Kay, P.M., "Measurement Techniques: What We Have and What We Need." In Exploring Competency-Based Education, W. R. Houston, Editor, Berkeley, California: McCutchan, 1974.

Lennon, R. T. "Accountability and Performance Contracting," paper read at the Annual Meeting of the American Educational Research Association, February 1971, New York.

Livingston, S. A. "A Note on the Interpretation of the Criterion-Referenced Reliability Coefficient," Journal of Educational Measurement, 1974, 10, p. 311. (This is a reply to "Note on the Variances and Covariances of Three Error Types," Journal of Educational Measurement, 1973, 10, pp. 49-50, by Chester W. Harris.

Lord, F. M. "Elementary Models for Measuring Change," Problems in Measuring Change, edited by Chester W. Harris, pp. 21-38. Madison: University of Wisconsin Press, 1963,

McDonald, F. J. "The State of the Art in Performance Assessment of Teaching Competence," Performance Education-Assessment, Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 21-27.

McNeil, J. D., and Popham, W. J. "The Assessment of Teacher Competence," Second Handbook of Research on Teaching, American Educational Research Association, edited by Robert M. W. Travers, Rand McNally & Company, Chicago, 1973, pp. 218-244.

Medley, D. M. "Teacher Examination Services for the Seventies," Teacher Behavior Research Group, Educational Testing Service, presented at the conference on Performance Evaluation, New Jersey, December 11-12, 1969.

Medley, D. M., and Mitzel, H. E. "Measuring Classroom Behavior by Systematic Observation," Handbook of Research on Teaching, edited by N. L. Gage, a project of the American Educational Research Association, Rand McNally & Company, Chicago, 1963, pp. 247-328.

Medley, D.M.; Soar, R.S.; and Soar, R., Assessment and Research in Teacher Education: Focus on PBTE, PBTE Series: No. 17 AACTE, June 1975.

Merwin, J. Performance-Based Teacher Education: Some Measurement and Decision-Making Considerations, PBTE Series: No. 12, AACTE, June 1973.

Millman, J. "Performance-Based Indicators of Teaching Effectiveness, Has the Time Come?" presented at the annual meeting of the Northeastern Educational Research Association, November 12, 1974.

Millman, J. "Psychometric Characteristics of Performance Tests of Teaching Effectiveness," presented to AERA, New Orleans, February 1973.

Morse, K. R., Smith, C. A., & Thomas, G. P. Assessment of Competence, An Assessment Model for Performance-Based Training Models, Teaching Research, Oregon State System of Higher Education, Monmouth, Oregon, August 1972.

Okey, J. R. and Humphreys, D. W. "Measuring Teacher Competence," paper presented at the National Association for Research in Science Teaching annual meeting, Chicago, 1974.

Popham, W. J. "Alternative Teacher Assessment Strategies," Performance Education-Assessment, Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 34-38, (a).

Popham, W. J. "Applications of Teaching Performance Tests to Inservice and Preservice Teacher Education," Performance Education-Assessment, Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 39-50, (b).

Popham, W. J. "Identification and Assessment of Minimal Competencies for Objectives-Oriented Teacher Education Programs," Performance Education-Assessment, Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 51-57, (c).

Quirk, T. J. "Test Scoring Based on the Instructional Objective as the Basic Criterion Unit," Journal of Secondary Education, February 1970, Vol. 45, No. 2, pp. 61-65.

Quirk, T. J. "Performance Tests for Beginning Teachers: Why all the Fuss?" invited address delivered to the Second General Meeting of the New Jersey Performance Evaluation Project, North Brunswick, New Jersey, May 27, 1971.

Quirk, T. J. "Psychological Problems of Competency-Based Teacher Education Programs," final revised version, Educational Testing Service, an invited address delivered to the Workshop on Problems of Competency-Based Teacher Education, the Teacher Corps, State University of New York at Albany, May 13, 1972.

Quirk, T. J. "Some Measurement Issues in Competency-Based Teacher Education," Phi Delta Kappan, Vol. LV, No. 5, January 1974, pp. 316-319.

Raths, J. "Problems Associated with Describing Activities," Observational Methods in the Classroom, edited by Charles W. Beegle and Richard M. Brandt, 1973.

Remmers, H. H. "Rating Methods in Research on Teaching," Handbook of Research on Teaching, edited by N. C. Gage, Rand McNally, Chicago, 1963, pp. 329-378.

Rosenshine, B, and Furst, N. "Research on Teacher Performance Criteria," Research in Teacher Education, A Symposium edited by B. Othanel Smith, 1971.

Roth, R. A. Performance-Based Teacher Education: A Survey of the States, Michigan Department of Education, Teacher Preparation and Professional Development Services, September 1974.

Ryans, D. G. "An Information-System Approach to Theory of Instruction with Special Reference to the Teacher," paper presented at the annual meeting of the American Educational Research Association, February 13, 1965.

Sandefur, J. T. "An Illustrated Model for the Evaluation of Teacher Education Graduates," AACTE Commission on Standards, Washington, D.C., September 1970.

Schalock, H. D., Garrison, J.H., and Kersh, B.Y. "From Commitment to Practice in Assessing the Outcome of Teaching," Performance-Based Education-Assessment, the University of the State of New York, the State Education Department, Division of Teacher Education and Certification and Multi-State Consortium on Performance-Based Teacher Education, September 1974, pp. 58-90.

Scott, W. A. "Reliability of Content Analysis: the Case of Nominal Scale Coding," Public Opinion Quarterly, Vol. 19, Fall, 1965, pp. 321-325.

Sherwin, S.S. Performance-Based Teacher Education: Results of a Recent Survey, Educational Testing Service, Princeton, New Jersey, 1973.

Soar, R. S. "Methodological Problems in Predicting Teacher Effectiveness," The Journal of Experimental Education, Vol. 32, No. 3, Spring 1964, pp. 287-291.

Soar, R. S. "Accountability: Assessment Problems and Possibilities," Journal of Teacher Education, Vol. XXIV, No. 3, Fall 1973, pp. 205-212.

Stake, R. E. "Measuring What Learners Learn," Educational Research House(ed.), School Evaluation: The Politics and Process, Berkeley, California. McCutchan Publishing Corp., 1973.

Stern, G. G. "Measuring Noncognitive Variables in Research on Teaching," Handbook of Research on Teaching, edited by N. L. Gage, American Educational Research Association, Rand McNally & Company, Chicago 1963, pp. 398-447.

Thorndike, R. L. and Hagen, E. Measurement and Evaluation in Psychology and Education (Third Edition), New York: Wiley, 1969.

Turner, R. L. "Levels of Criteria," Performance-Based Teacher Education Vol. 1, No. 5, December 1972. Multi-State Consortium.

Weber, W. A. "Competency-Based Teacher Education: An Overview," (slide-tape materials), Westport, Connecticut: Videorecord Corporation of American, Inc., 1970.

Young, J.I., "Model of Competency-Based Evaluation," Brigham Young University, 1973.