1990

3. Improving Teaching Through the Assessment Process

Donald M. Medley
University of Virginia

3

Improving Teaching Through the Assessment Process

Donald M. Medley
University of Virginia

INTRODUCTION

The first opportunity to use teacher evaluation to improve teaching arises when a student applies for admission into an undergraduate teacher preparation program. At this time it is the responsibility of the program faculty to determine whether each candidate possesses those abilities and other personal characteristics that every teacher needs, but cannot expect to acquire in such a program, and to deny admission to those who lack one or more of them. The second opportunity arises when the student has completed the program. At this time it is the responsibility of the state certification agency to find out whether each candidate has acquired the minimum professional knowledge and skill necessary for certification as competent to enter the teaching profession, and deny certification to those who have not. Additional opportunities arise after the teacher enters into practice and either comes up for tenure or becomes a candidate for merit pay. At either point it is the responsibility of the school administration to ascertain whether the teacher is performing well enough to receive tenure or merit pay and deny them to those who are not.

If the evaluation made at each of these times is valid and is followed by appropriate action, the overall quality of teaching in the schools is expected to improve because incompetent teachers would be systematically eliminated from the profession. In order for this theory to work, each incompetent teacher who is eliminated must be replaced by another teacher who is competent. Thus the success of this strategy depends on the assumption that an ample supply of competent teachers is available to replace those we eliminate, an assumption unlikely to prove true.

There is a second strategy for using teacher evaluation to improve teaching, the success of which does not depend on this rather dubious assumption. This alternative strategy is to increase the competence of the incompetent teachers we already have instead of replacing them. The success of this strategy depends, like that of the first, on the validity of the teacher evaluations used. Unless the procedures used to screen out incompetent teachers are valid, all that the first strategy can do is increase teacher turnover. Unless the evaluation procedures used to upgrade the competence of the teachers we have are valid, all that the second strategy can do is prolong the training some teachers receive.

There are two major questions that must be answered before either of these strategies can be applied with any success. The first of these questions is: What should we evaluate? The second question is: How shall we evaluate it? Only when the first question has been answered is it possible to answer the second. Past efforts to use teacher evaluation to improve teaching have failed, largely because they have tended to neglect the first question and concentrate on the second. Before we can answer either question we must make and preserve careful distinctions in the meanings of four terms too often used interchangeably. These terms are *teacher competence, teacher competency, teacher effectiveness,* and *teacher performance.*

Some Important Definitions

In defining these four terms I will use the simple model of the teacher evaluation process shown in Figure 3.1. The diagram presents a kind of inventory of the points in a teacher's professional life at which evaluations designed to improve teaching can be made. It shows five points at which teachers may be assessed on different bases, and four points at which other relevant variables—usually called "context" variables—may be assessed.

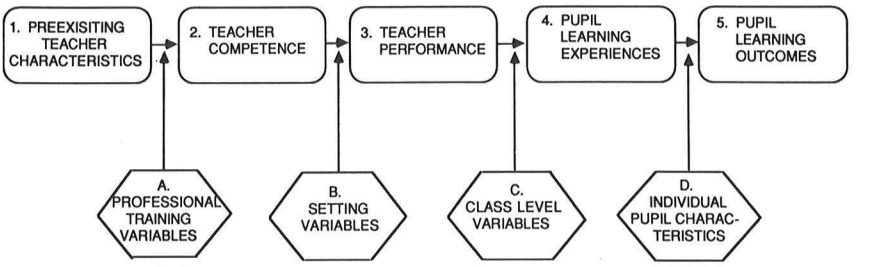

FIG. 3.1. Five Teacher Assessment Points

Preexisting teacher characteristics are assessed at Point 1, the earliest point at which any teacher evaluation is feasible. Evaluations of preexisting teacher characteristics may be used by teacher educators to improve teaching by using them to decide which candidates should be admitted into a preservice teacher preparation program and which should not.

There are a number of abilities and other characteristics that teachers need and are expected to acquire before beginning professional training. One example is the kind of academic ability the Scholastic Aptitude Test is used to measure. Another is the basic general knowledge, sometimes called "general literacy," that all high school graduates are expected to possess. It is generally agreed that any teacher of any grade or subject should be literate in this sense.

What is unique to these characteristics is that their development neither is nor should be part of a professional teacher education program. Hence students who lack such a characteristic when they begin their professional preparation will almost certainly not possess it when when they finish the program. If it is known that possession of the characteristic will be required for certification, then the time to evaluate it is before the teachers enter the program, not when they finish it.

Teacher Competence

Teacher competence is assessed directly at Point 2, usually as a basis for deciding whether the teacher should or should not be certified or licensed to teach. The state certification agency tries to improve teaching in the state by permitting only teachers with some minimum level of competence to become teachers.

Before we can evaluate competence validly and reliably enough to implement either of the two strategies for improving teaching,

we must have a precise definition of competence. This requires us first to *specify* exactly what we mean by the term, and then to *define* the knowledge, skills, and so forth, that a teacher must possess in order to be competent.

Specifying Competence. Competence is specified by identifying the teaching tasks or functions that a competent teacher must be able to perform. For some purposes we may need to specify competence rather narrowly, for instance, when specifying competence to administer and interpret individual intelligence tests or competence in using a particular method of teaching reading. For other purposes we may need to define competence more broadly, for example, as in competence to teach kindergarten, competence to teach high school mathematics, or competence to teach pupils with severe emotional handicaps.

Defining Competence. Once competence is specified, the second important step is to identify the knowledge, skills, and other qualities a teacher must possess in order to perform the functions specified. Only then can we say we have defined competence precisely enough to be able to evaluate it objectively, validly, and reliably.

Teacher Competency

Despite some negative connotations that it has acquired over the years, I shall use the term *teacher competency* to refer to any single item of knowledge, skill, or any other specific characteristic we have identified as one that a competent teacher is expected to possess. We can then say that *teacher competence* in performing a function is defined as *the possession of a specific set of teacher competencies* relevant to the performance of that function.

Teacher Effectiveness

Whether or not a competent teacher will be effective on the job depends in part on whether the set of competencies that make up the definition is sufficient to guarantee effectiveness. At the present state of knowledge about the nature of effective teaching this is most unlikely. At best, a definition of competence can and should incorporate all we know plus, perhaps, our best guesses about what we do not know, that will help a teacher perform the specified function.

The least that we can expect a teacher preparation program

faculty to do is to equip each graduate with this knowledge—in other words, the faculty should transmit to the teacher what they see as the relevant wisdom of the profession. And the least that we can expect of a valid evaluation of competence is a measure of how much of this wisdom each teacher has acquired.

Teacher Performance

Teacher performance is assessed at Point 3, usually as a basis for one or another administrative decision about teacher utilization. Teachers are hired, tenured, recognized as master teachers on the basis of evaluations of their performance on the job. School administrators can improve the quality of teaching by screening out teachers who fail to perform the specified function successfully and replacing them with teachers who do perform it successfully.

Teacher performance is defined, not in terms of competence nor in terms of what the teacher is able to do, but in terms of what the teacher actually does on the job. Unlike competence, which is evaluated on the basis of teacher behavior in a test situation, performance must be evaluated on the basis of the behavior of the teacher while doing the job he or she was hired to do. Evaluations of teacher performance are therefore based on observations of the processes and procedures the teacher uses in teaching, observations made during one or more visits to the teacher's classroom (not on the results the teacher obtains).

Assumptions. Valid performance assessment is possible only if two assumptions are true. One is the assumption that the teacher behavior observed is a representative sample of the teacher's behavior when he or she is not being observed. The other is that it is possible to specify rules of procedure that a teacher should follow.

The first of these assumptions is almost certainly unjustified and unjustifiable. The request many evaluators make that while the teacher is being evaluated he or she should act as though no observer were present is a request that the teacher is likely to ignore, and indeed has a perfect right to ignore. The right to do one's best when one's performance is being evaluated for employment, tenure, or promotion may well be a basic human right.

The second assumption is also questionable. It depends on the doubtful proposition that there is one way to teach in a given situation which is best for all teachers; and on the even more doubtful proposition that someone who has just walked in the door is a better judge of what a teacher should be doing at any given

moment than the teacher who has been there since the beginning of the school year. Both are inconsistent with the assumption that the nature of teaching requires a teacher to function as a professional problem solver.

Performance or Competence? In view of these limitations, it may seem odd that the vast majority of evaluations of practicing teachers are evaluations of this type, and that almost all decisions about teacher personnel are based on performance evaluations. The only explanation I can suggest is that what these evaluators are really trying to evaluate is teacher competence. It is much more difficult to infer teacher competence from teacher performance than it looks; and even if it were not, a teacher's competence is not the appropriate basis for the kinds of decisions that are based on these evaluations. It is not the teacher who is able to do the best job but the one who does the best job who should be hired and retained. The race goes not to the swiftest but to the first to reach the finish line.

Pupil learning experiences are assessed at Point 4. This term will be used to refer to any in-school pupil activity intended to result in pupil learning. Doing a workbook assignment is one example of a pupil learning experience; watching and discussing an instructional film is another. Listening to the teacher is a third, and perhaps the most popular of all. We all know that learning results from activity of the learner. Making sure that pupils engage in productive activities, that is, providing them with learning experiences appropriate to the goals of education is what schools and teachers are for.

Evaluations based on observations of pupil behaviors during visits to a teacher's classroom may provide a desirable alternative basis for the decisions about teacher utilization usually based on performance evaluations. What would be more logical than to evaluate a teacher's performance on the basis of the amount and quality of the learning experiences her pupils have in her classroom, that is, on the use she makes of the time pupils spend under her care?

Assumptions. Two assumptions must be true for evaluations of pupil learning experiences based on classroom observations to be valid. The first is that the pupil activities observed during a visit are representative of those that occur in that same classroom when the observer is absent. The second assumption is that it is possible to define the kinds of learning experiences the pupils in a certain class should be having, regardless of who their teacher is.

Let us compare these two assumptions with the parallel assumptions that underlie performance evaluations made at Point 3. The assumption that observed pupil behaviors are representative of "normal" pupil behaviors is somewhat more likely to be true than the assumption that observed teacher behaviors are representative of "normal" teacher behavior. For one thing, the pupils are not being evaluated, so their right to do their best is not involved.

The second assumption required at Point 4 is also more justifiable than the second assumption required at Point 3. If we take the point of view that the school system employs th teacher to provide pupils with appropriate learning experiences, it seems reasonable for the school system to define the kinds of learning experiences that are appropriate. Doing so does not mean that the school system must prescribe how the teacher should go about performing this function, as is the case when performance is assessed directly at Point 3. Assessment at Point 4 leaves teachers free to function as professionals and use whatever processes and procedures they think best.

Although some teacher-rating scales contain items that refer to related pupil behaviors (such as level of attention), I know of no instance in which the learning experiences of teachers' pupils have been the explicit and sole basis of evaluations of teacher performance made to support personnel decisions. The best example I know of the use of pupil-learning experiences as a basis for evaluating teachers occurred in a research project (Cf. Berliner, 1979; Denham & Lieberman, 1980).

Pupil-learning outcomes are assessed at Point 5. This term refers to changes in pupil status with respect to educational goals that take place during the period of time a teacher has the pupil in her class. The ultimate purpose of efforts to improve teaching is, of course, to increase pupil learning outcomes.

The amount and quality of learning outcomes in a teacher's classroom depend on a great many important factors. Teachers have a considerable amount of control over some of these, including their own competence and performance while teaching. But teachers have relatively little control over other factors, such as the support available from the school and community, the makeup of the class, and the characteristics of the individual pupils in the class.

Contextual Factors

So far we have discussed only those factors over which the teacher has considerable control, those which are or could be foci of efforts

to evaluate teachers. Let us now turn our attention briefly to those factors over which the teacher has relatively little control, represented in Figure 3.1 at Points A through D. Variables of these types are usually called "contextual factors."

A. Professional Training Variables. Type-A factors are characteristics of teacher training that affect teacher competence directly and affect teacher performance, pupil learning experiences and pupil learning outcomes only indirectly. Changing training variables can increase pupil learning outcomes by increasing teacher competence, although a lot of things can go wrong between Point A and Point 5.

B. Setting-Variables. Type-B factors are characteristics of the setting, that is, of the community, the school system, and the individual school in which the teacher is employed. Changes in setting variables, in, let us say, the administrative and supervisory support a teacher receives, can increase pupil learning outcomes by improving teacher performance.

C. Class-Level Variables. Type-C factors are characteristics of the pupils in a teacher's class as a group. Changes in the makeup of a class, in the mix of abilities, ethnic groups, mainstreamed pupils, and so forth, can, by changing the nature of this group, alter the learning experiences a pupil has in it, and increase (or decrease) pupil learning outcomes.

D. Individual Pupil Characteristics. Type-D factors are characteristics of the individual pupil that determine what and how much a pupil learns from a given learning experience. They include such things as aptitude for learning and motivation to learn.

Teacher Effectiveness

The term *teacher effectiveness* refers to the portion of what a pupil learns that is attributable to the performance of his teacher. It is so difficult and expensive to obtain valid measures of teacher effectiveness that they are useless for all practical purposes except research, especially studies of the validity of other ways of evaluating teachers. The technical problems that must be solved in order to obtain valid direct measures of teacher effectiveness are formidable. It is necessary, first to identify and then to measure all of

the important factors that affect pupil learning outcomes and then tease out and evaluate the effect of the teacher by statistical means. No less formidable are the difficulties to be overcome in obtaining defensible measures of pupil progress toward the important goals of education.

Even if direct measurements of teacher effectiveness were easy enough to obtain so that they could be used for routine teacher evaluations they would be of limited use in our efforts to improve teaching by either of the two major strategies defined earlier. The information such measurements contain about which teacher should be eliminated comes too late to be of any use. The time to eliminate an incompetent teacher is *before,* not after the teacher has taught long enough to become a candidate for permanent tenure. Nor do direct measures of teacher effectiveness contain any diagnostic information, any clue as to what the ineffective teacher needs to do in order to become more effective.

Needed Research. The principal use of direct measures of teacher effectiveness is in the research we so badly need to improve evaluations of teachers at Points 1 through 4. First of all we need *research in classroom learning,* that is, research correlating pupil learning experiences with pupil learning outcomes, adjusting for important individual pupil characteristics. Such research should tell school administrators what kinds of learning experiences maximize pupil learning outcomes, so they can evaluate a teacher on the basis of the amount of such learning experiences the teacher provides.

Next we need *research in teaching,* that is, research correlating teachers' performance and the learning experiences pupils have in their classrooms, adjusting for important class characteristics. Such research should tell supervisors how teachers should behave in order to provide pupils with the kind of learning experiences that research in classroom learning indicates they should have, so they can diagnose and prescribe ways in which teachers can improve their performance.

Next we need *research in teacher competence,* research correlating teacher competencies and teacher performance, adjusting for important setting variables. Such research should help teacher educators and state certification agencies to improve their definitions of competence and, therefore, improve the performance and increase the effectiveness of the teachers they train and certify.

Finally, we need *admissions research,* research correlating preexisting teacher characteristics with measures of teacher compe-

tence obtained at the end of training, adjusting for important training variables. Such research should tell admissions officers what characteristics to require students to possess in order to maximize the number who will acquire the competencies identified by research in teacher competence as ones every graduate should possess.

Focus of This Chapter

While the educators, certification agencies, and teacher educators of the country are waiting for the findings of all of this research, they have no choice but to continue to try to improve teaching by evaluating teachers as well as they can. The most highly visible efforts to improve teaching by using teacher evaluation are of course those being made by the large-scale teacher-evaluation programs so many states are operating. Most of these programs base their evaluations on conventional paper-and-pencil tests or on expert ratings of teacher performance. There is no evidence that scores on either type of instrument have any appreciable validity as measures of teacher competence, performance, or effectiveness. It is therefore highly improbable that any of these programs is effective in improving teaching.

It is the thesis of this chapter that, although the knowledge of the nature of teacher competence presently available is far from complete, it is sufficient to enable us to develop much more valid and reliable instruments for evaluating teacher competence—that can be administered at little or no greater cost in time or money than the virtually worthless ones in present use.

The first critical step we must take in order to develop such instruments is to define competence explicitly enough so that it can be measured. In order to do this we need, first, to *specify* competence in terms of what a competent teacher is supposed to be able to do. Only then will it be possible to *define* competence, to identify exactly what knowledge, abilities, and so forth, a competent teacher must possess.

Before a state licensure or certification officer (or anyone else) can design a valid system for evaluating teachers he or she must specify the kind of teachers wanted, that is, the teaching functions they should be qualified to perform. Will they be expected to function as elementary teachers, physics teachers, special education teachers? Then what must be decided next is precisely what competencies, what knowledge, skills, and so forth, a teacher should

possess in order to be declared competent to perform these functions.

At this point the state licensure or certification official should be able to turn to the research for guidance; but in the present state of the art of teaching, not enough is known about the relationship between competence and effective teaching to make it possible to arrive at an authoritative answer, a definition of teacher competence on which there is any general consensus. This fact does not reduce the need for the certification official to be precise in defining competence; if anything, it makes the need for precision more important. If the teachers certified as competent fail to perform satisfactorily it is important to be able to tell why, and revise the definition of competence accordingly.

The rest of this chapter focuses on these problems; on specifying, defining, and evaluating teacher competence.

A FRAMEWORK FOR DEFINING TEACHER COMPETENCE

What I propose to do next is present a kind of model definition or framework for a definition of teacher competence that will facilitate the related task of developing valid, objective, and practicable procedures for evaluating teacher competence. An inspection of almost any definition of teacher competence published in the past reveals a failure to distinguish between the task of *specifying* the functions a competent teacher must perform and that of *defining* the competencies needed to perform those functions. (For an excellent example see Johnson, Okey, Capie, Ellett, & Adams, 1978.) As we have seen, such a specification is a necessary first step in the process of defining teacher competence; but by itself such a specification is of little help in the construction of an evaluation instrument. The framework I present includes both a specification and a definition.

Because the model I propose to describe needs to be applicable to a definition of almost any kind of teacher competence, the function specified must be generic, must be one that any and every teacher is expected to perform. Does such a teaching function exist, and if it does, what is it? I suggest that any profession is defined by some one generic function that all members of that profession must perform; and that competence in that profession must be defined in terms of the knowledge, skills, and abilities needed to perform this generic function.

The Generic Function of the Teacher

I first became aware of the generic function of the teacher when I read the report to the American Association of Colleges for Teacher Education of its Bicentennial Commission (Howsam, Corrigan, Denemark & Nash, 1976). According to this report, the function of teachers in this society is the same as that of any other professional, which is to bring professional knowledge to bear on certain problems the society faces. The report notes that as civilizations advance and encounter more and more complex problems, they turn for solutions more and more often to persons with special competence to deal with such problems.

The people they turn to are members of what are called *learned professions.* These professions are called "learned" because practitioners of each one of them possess specialized knowledge and skill relevant to the solution of a certain class of difficult problems. The role society expects teachers to fulfill, like that of practitioners of other learned professions, is to apply specialized professional knowledge and skill to the solution of problems of a certain type.

Just as society expects physicians to apply the accumulated wisdom or "mystique" of the medical profession to the solution of health problems of their patients, so it expects teachers to apply the accumulated wisdom or mystique of the teaching profession to the solution of learning problems of their pupils. There is no doubt about the need for such knowledge, although there is some question in the public mind whether enough of it exists to make the average teacher any better able to cope with teaching problems than anyone else.

Three Types of Teaching Problems

Which way is the best way to evaluate a teacher's ability to solve a teaching problem depends very much on the nature of the problem. It is therefore useful to group the different kinds of problems teachers must solve according to how a teacher's ability to solve them is most validly—and easily—evaluated. We use the following three categories.

Category 1: Interactive Teaching Problems include teaching problems that arise in the classroom when pupils are present and interacting with the teacher—participating in a discussion, listening to a teacher presentation, working individually under the

teacher's supervision, or having learning experiences of some other kind under the teacher's guidance.

Category 2: Preactive Teaching Problems include teaching problems that arise when no pupils are present, while the teacher plans instruction, diagnoses pupil needs, evaluates test papers, or performs some other teaching task that does not involve interacting with pupils.

Category 3: Reflective Teaching Problems include problems teachers recognize while reflecting on or reviewing their own past performance with a view to improving future performances.

The first two categories were originally identified by Jackson (1966). Jackson pointed out that the abilities a teacher needs to make the almost instantaneous decisions required when teacher and pupils are interacting are very different from those needed to make the deliberate decisions made while reviewing past interactive sessions or planning future ones. The third category came to our attention in the work of Cruickshank and Applegate (1981), who have developed procedures for preparing teachers to solve problems of a third type. One of the characteristics of a learned profession is that the process of professional education continues throughout the practitioner's career, that the true professional never ceases to reflect on past performances with a view to improving future ones.

Teacher Competence and Teacher Performance

Let us turn now to the often-neglected step of defining the knowledge and the skills a teacher needs in order be competent to perform the generic teaching function, which is to solve teaching problems.

The problem-solving process can be conceptualized in different ways for different purposes. Because of the purpose this conceptualization is to serve, I have chosen to break up the process into four steps, each of which calls for different competencies, best evaluated by different methods.

The four types of competencies are referred to as: *perceptual skills, professional judgment, professional knowledge,* and *perfor-*

mance skills. The relationships of each of these types of competencies to performance are shown in Figure 3.2.

The four types of competencies are shown at the left of the diagram, with arrows from each leading to diamonds containing question marks, representing "branch points."

Let us agree that a teaching problem arises whenever a pupil does something that he or she should not do, or when something happens to him that should not happen. One pupil copying another's work during a test might be one simple example; a misspelled word on a test paper may be another.

Type 1: Perceptual Skills

It is obvious that a teacher cannot solve a teaching problem unless he or she is aware of the occurrence of the event that gives rise to the problem; the teacher must see the pupil copy or realize that the word is misspelled before he or she can deal with either of the problems just mentioned.

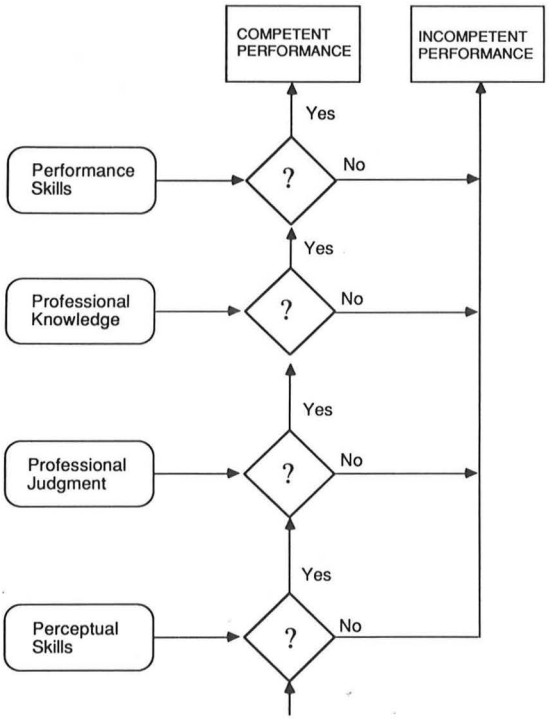

FIG. 3.2. Teacher Competence and Teacher Performance

The competencies a teacher needs in order to be aware of what is happening to the pupils will be called *perceptual skills.* Kounin (1970) has enriched the language of teaching by introducing the term *withitness* in referring to the aspect of this competency relevant to interactive teaching, and Berliner (1986) studied differences in what "expert" teachers and novice teachers see when they view the same classrooms. Pupils speak of teachers with high perceptual skills of this type as having eyes in the back of their heads. Possession of this skill enables some teachers to nip certain situations in the bud—to move a pupil to another seat before he or she even thinks of misbehaving.

Smith (1969) identified a somewhat different kind of perceptual skill relevant to interactive teaching—a skill needed to recognize abstract pedagogical concepts when they occur in th "real world," and he also invented "protocol materials" to be used to help teachers develop this skill. Knowledge of reinforcement theory is of little use unless you can recognize when a pupil is being reinforced.

Perceptual skills probably play a no less important role in the solution of preactive teaching problems; the ability to recognize arithmetic errors and misspelled words or to read pupils' handwriting may be examples.

Note that, in the figure, an arrow runs from perceptual skills to a branch point and that two arrows come out of it. If a teacher fails to see a problem behavior, that is, lacks the relevant perceptual skill, we follow the "no" arrow which leads us to "incompetent performance." This means that if the teacher fails to apply the competency, he or she fails to perform the function, that is, to solve the problem.

If the teacher does see the behavior (does apply the competency) we follow the "yes" arrow to the next branch point.

Type 2: Professional Judgment

Competencies of this type involve recognizing the behavior as problem behavior, as something that needs to be changed or corrected. During interactive teaching the most obvious examples have to do with the limits the teacher sets on pupil behavior; for example, how much noise, how much moving about, and so on, the teacher permits. Professional judgment in such matters is a major factor in classroom management. Professional judgment also has to do with teacher expectations, with the kind of pupil response or performance the teacher finds acceptable or praiseworthy from

which of her pupils. Professional judgment in preactive teaching also has a lot to do with expectations or standards of pupil performance.

If the teacher is aware of problem behavior but does not recognize that it is problem behavior, we follow the "no" arrow out of the second branch point in the diagram to "incompetent performance." If she does recognize the existence of the problem we follow the "yes" arrow to the next branch point.

Type 3: Professional Knowledge

If the teacher recognizes a problem, the next type of competency needed is knowledge of various possible responses to the problem and their probable consequences. Part of this knowledge may be regarded as "foundational," knowledge presumably acquired in professional courses in psychology, sociology, human growth and development, and so forth, and part of it comes from courses in methods or strategies of teaching; in either case, it must be functional in the sense that the teacher can relate it to the problem behavior he or she faces.

Unless the teacher applies professional knowledge and comes up with a response that solves the problem, we follow the "no" arrow out of the third branch point to "incompetent performance." Otherwise we follow the "yes" arrow to the next (and last) branch point.

Type 4: Performance Skills

Once the teacher has identified a solution to the problem, he or she needs only to implement the solution to solve the problem, as we find by following the "yes" arrow out of the fourth branch point, which leads to "competent performance." If the teacher is unable to implement the solution, we follow the "no" arrow to "incompetent performance."

Note that these four types of competencies are related sequentially; that is, that no opportunity to apply any one competency arises unless all preceding competencies have been applied successfully. Note also that successful performance is possible only if all four types of competencies are successfully applied.

Implications for Improving Teaching. This simple analysis should make it clear why, if we are interested in improving teaching, it is better to evaluate teacher competence than teacher performance. Because teacher performance is defined in terms of success

in solving teaching problems, all we find out when we evaluate teacher performance is whether or not the teacher solves the problem. This may be useful if the teacher succeeds; but if he or she does not solve the problem, we have no clue as to how or why he or she failed, no indication as to how we can help the teacher improve future performance.

When we evaluate teacher competence instead of performance we still find out whether or not the teacher succeeds in solving the problem; but if the teacher fails we learn a lot more. We learn which of the competencies the teacher needs to acquire in order to solve the problem, and have a clear indication of how to improve the teacher's future performance.

A Competency Matrix

If we combine the three types of problems and the four types of competencies just described, we generate 12 different kinds of competencies. In general, all of the competencies in the same cell may be assessed in the same way, or ways that are quite similar; and competencies in different cells are usually best assessed in different ways. The 12 cells form the skeletal map of the domain of teacher competence shown in Figure 3.3.

	PERCEPTUAL SKILLS	PROFESSIONAL JUDGMENT	PROFESSIONAL KNOWLEDGE	PERFORMANCE SKILLS
INTERACTIVE TEACHING PROBLEMS				
PREACTIVE TEACHING PROBLEMS				
REFLECTIVE TEACHING PROBLEMS				

FIG. 3.3. A Matrix of Teacher Competencies

I call the map "skeletal" because it contains no actual competencies, only empty cells. The matrix was originally designed for use with a set of competencies defined beforehand. The idea was first to assign each competency to one of the 12 cells and then to construct the evaluation instrument or instruments. Experience indicates that the matrix can also prove useful in the process of defining competence. Suppose, for example, that you wanted to define competence to teach one of the primary grades.

Following the structure in Figure 3.3, you might begin by specifying the functions such a teacher would be competent to perform. You would almost certainly specify these functions in greater detail than the matrix shows. You might subdivide interactive teaching problems into those related to classroom management, those related to the delivery of instruction, those related to evaluation, and so forth. Or you might subdivide them into problems that arise in introducing a new activity or lesson, presenting or developing new material, reviewing and summarizing, conducting guided practice, making an assignment, and ending a lesson or activity.

Next you would analyze the process of solving teaching problems (as shown in Figure 3.2) as it applied to problems of each of these kinds, defining in detail the competencies of each type that you considered most important to the performance of each function.

Suppose, for example, that one subdivision of interactive teaching problems you had specified contained problems related to "classroom management." You might consider what kinds of warning signs the teacher should be especially sensitive to (Type 1); what limits the teacher should set on pupil conduct (Type 2); what professional knowledge would be most useful (Type 3); and what techniques or strategies for dealing with pupils the teacher should master (Type 4). Or if a subdivision under "delivering instruction" had to do with teaching reading, a similar analysis might focus on what the teacher should listen and watch for while a pupil reads aloud, what kinds of errors the teacher should or should not interrupt the pupil to correct, and so on.

It should be apparent how much the completion of such a map of the particular domain of competence you wish to measure would simplify the task of constructing instrumentation to measure the precise competence you set out to evaluate.

We have seen that the process of constructing an instrument for measuring teacher competence involves three steps. The first two, specifying the functions a competent teacher must perform and

identifying the competencies needed to perform them, are by far the most difficult. They have already been discussed. The third step, constructing test exercises that require the use of each of these competencies and assembling them into one or more instruments, is discussed next.

EVALUATING SELECTED TEACHER COMPETENCIES

I deal with this third step by presenting three examples drawn from attempts to evaluate specific competencies for various purposes in which I have been involved. For the sake of brevity I discuss examples related to just one of the three types of teaching problems, those that arise during interactive teaching. Some examples of exercises related to preactive teaching problems have been published elsewhere. (McNergney, Medley, Aylesworth, & Innes, 1983.) Because I have had no experience in evaluating competencies related to postactive teaching problems I do not discuss them here.

Measurement-Based Teacher Evaluation

All three attempts used a general approach to teacher evaluation called *measurement based teacher evaluation,* which was designed to free teacher evaluation from any dependence on the expertise of the person who does the evaluation. The much-debated question, "Who should evaluate the teacher?" disappears when the evaluation is measurement based (Medley, Coker, & Soar, 1984.)

Measurement-based teacher evaluation was designed to emulate the familiar multiple-choice, paper-and-pencil test which, despite its many limitations, represents the most technically advanced methodology yet developed for assessing human characteristics from human performance. From a study of such tests we conclude that there are three essential conditions for objective measurement of human performance, as follows:

1. *All candidates being assessed must perform the same tasks or equivalent tasks.* In the case of a paper-and-pencil test, the tasks set for all candidates are the same: They must all answer same set of test items or questions.
2. *An accurate, quantifiable record of each candidate's perfor-*

mance of the tasks must be obtained. In the case of a paper-and-pencil test, the candidate records his own performance by marking an answer sheet that is machine readable.

3. *There must be a procedure for quantifying (or scoring) the performance that can be carried out by a clerk or a computer.* In the case of a paper-and-pencil test, a computer reads and scores the marks on the machine-readable answer sheet.

When these conditions are met, the validity and reliability of the measurements obtained ultimately depend on the degree to which successful performance of the tasks depends on the ability or other characteristic being measured. Given an appropriate set of tasks, the validity and reliability of the measurements obtained depends on how the performance records are quantified or scored.

Powerful analytical procedures have been developed for using empirical data to maximize test validity by refining the tasks (e.g., item analysis) and by refining the scoring procedures (e.g., scaling techniques). These procedures are fully applicable to the refinement of measurement-based teacher evaluation instruments.

Assessing Functional Professional Knowledge

My first example was a response to a request from the developer of a set of inservice teacher training packages, each of which was designed to increase teachers' professional knowledge of techniques for dealing with one type of interactive teaching problems. The developer asked us to construct an instrument that would measure whether teachers who had completed a package were more likely to apply the professional knowledge it contained in solving interactive teaching problems than teachers who had not completed that package. What was needed was what we call "a measure of functional professional knowledge," that is, a measure of the ability to apply professional knowledge to the solution of teaching problems—in this case, interactive teaching problems.

Multiple-choice tests have been widely used to measure knowledge of all kinds, including professional knowledge. But the professional knowledge these tests measure does not seem to be of any use to the teacher in solving interactional teaching problems. If it were of any use, a teacher's scores on the tests would correlate with his or her classroom performance. But repeated efforts to establish correlations between scores on tests of this type and mea-

sures of classroom performance have failed, even when the tests used were the best available (cf. Quirk, Witten, & Weinberg, 1973).

The reason becomes clear when we compare the tasks a student must perform to get a high test score with the ones a teacher must perform to succeed in the classroom.

Solutions. The items on a test as well constructed as the National Teacher Examinations are designed to measure the students ability to apply professional knowledge to realistic teaching problems. But every problem a student encounters on such a test has one and only one correct solution, a solution which a panel of experts all agree is the correct solution to that problem. The student's task is to decide which of four or five alternatives is correct. (If there is any doubt about which response to an item is the correct one, the item is discarded.) But when a teacher encounters what may look like a similar problem in the classroom, she is not given four or five alternative responses, one and only one of which is clearly correct. The teacher must think up his or her own alternatives and has no idea how many of them will be correct, if any. Some of the problems that come up have more than one solution, all equally acceptable. Some have none.

Strategies. When a student takes a paper-and-pencil test all of the problems are presented at one time in a neatly printed booklet, and the student is free to attack the problems in any order he or she chooses, to spend as much time as needed on each one, to take extra time to ponder difficult problems, to skip some items and to change his or her mind about some. Interactive teaching problems must be dealt with when they come up; there is no time to ponder, no going back, and to postpone a response is to fail that problem.

Scope. When a student takes a test he or she knows that the solutions to all of the problems on the test will come from a single area of knowledge that the class has had a chance to study; thus the student can forget everything else he or she knows about any other area of knowledge. For example, a student taking a course in educational psychology will not need to apply any previously learned knowledge about the teaching of reading. But a teacher interacting with pupils needs instant access to any knowledge of any subject he or she may possess (or may not possess).

I could go on, but these examples make it quite clear that the skills a student needs to do well on a multiple-choice test have

little in common with those a teacher needs to do well during interactive teaching.

A Simulation Exercise. We therefore set about devising a simulation exercise which would require skills more like those the teacher needs. The simulation exercise we constructed confronted teachers with a series of interactive teaching problems similar to those an elementary school teacher might encounter in a normal day in the classroom. Each problem was presented in the form of a brief verbal vignette projected on a screen, with audio. Each vignette was followed by two or more suggested responses the teacher might make to it. The suggested responses to each problem were presented one at a time (in audio only), and the teachers had 5 seconds in which to decide whether or not each response suggested was one they might make in that situation, and to record their decisions by marking the appropriate spaces on machine-readable answer sheets.

The sequence of problems was designed to resemble the normal sequence of events in a classroom, beginning when the first pupils appear in the morning and ending when they board the school bus in the afternoon. For the sake of efficiency in measurement, most (but not all) of the suggested responses presented involved knowledge from one of the instructional packages, but responses reflecting knowledge from different packages were intermingled in a haphazard order.

The complete exercise consisted of 45 vignettes and required teachers to react to almost 200 suggested responses. A sample vignette and the suggested responses that accompanied it follow:

> Margaret and Grace are both docile, well-behaved children who are close friends and who have both been doing well in your class. One day while the children are taking a unit test you see the girls cheating (Grace is letting Margaret copy some of her work).
>
> What might you do?
>
> 101. Confiscate Margaret's paper and send her out of the room.
>
> 102. Walk over and stand near the two girls for the rest of the period.
>
> 103. Do nothing until the test is over; then tell both girls that you are giving them zeros.
>
> 104. Tear up both of their papers.
>
> 105. Move Margaret to a different part of the room.

One point was added to the teacher's score on a package for each response she marked that reflected knowledge of that package. Some suggested responses were inconsistent with the recommendations in a package; one point was subtracted for each of these responses the teacher marked. Those suggested responses that had nothing to do with the training packages (but were included because they are responses that teachers are likely to make) did not count.

Remember that the scores obtained were not intended to evaluate a teacher's overall ability to solve interactive teaching problems, only how well he or she was able to apply specified knowledge to the solution of these problems. In other words, scores were not intended to reflect a teacher's perceptual skill, professional judgment, or performance skills; only professional knowledge.

Measurement Properties. In addition to being inexpensive and easy to administer, this exercise meets all of the conditions for objective measurement of human performance just specified. First, all teachers perform the same tasks; second, they record their own performances, and third, the records they make can be read and scored by a computer. Therefore, as we have noted, the validity and reliability of the measurements depend ultimately on the nature of the tasks that make up the exercise and how responses to them are scored.

I have already presented evidence of content validity in my description of the resemblance between the tasks that make up this exercise and those related to the use of professional knowledge that a teacher faces in the classroom. I do not have any empirical evidence of the validity of this exercise to report.

There is, however, some rather striking empirical evidence of the validity of an exercise constructed by Hayes (1988), which was closely similar to the one described here. The source of the professional knowledge measured was different; Hayes' instrument was designed to measure knowledge of 13 of the 14 BTAP competencies (see Table 3.1)—those relevant to the solution of interactive teaching problems.

Hayes administered her exercise to four intact groups. One group consisted of 46 experienced teachers; one consisted of 30 teacher education students doing their practice teaching; one consisted of 31 college students not preparing to teach; and one consisted of 30 adults who had had no college education.

Although none of these people were aware of the existence of the

TABLE 3.1
Competencies Measured in the Virginia Beginning Teacher Assistance Program

A.	Academic learning time
B.	Accountability
C.	Clarity
D.	Individual differences
E.	Evaluation
F.	Consistent rules
K.	Affective climate
L.	Learner self-concept
M.	Meaningfulness
P.	Planning
Q.	Questioning skill
R.	Reinforcement
S.	Close supervision
W.	Awareness

BTAP competencies, the experienced teachers, with a mean score of 199 points, scored significantly higher than the student teachers, whose mean score was 187. Both groups scored significantly higher than the other college students tested, whose mean score was 182, and the noncollege educated adults, whose mean score was 162.

Hayes' instrument is the first and only test of professional knowledge (or of any cognitive ability) I have seen on which teachers in service outperform teachers in training. These findings provide strong evidence of the potential validity of this kind of simulation. And they also provide evidence that some of what teachers learn from experience can also be learned from a study of the findings of research on teaching.

Assessing Multiple Competencies

The second example of the use of measurement-based teacher evaluation to improve teaching was developed for use in a preservice teacher education program. It was designed as a relatively inexpensive way of obtaining diagnostic information about students' progress in acquiring interactive teaching competencies. It yields separate measurements of competencies in three of the cells in the competency matrix: perceptual skills, professional judgment, and professional knowledge.

This is another simulation exercise, administered by projecting brief videotapes of classroom episodes on a screen. Each episode is

followed by a series of verbal statements about the episode projected on the same screen (with audio), sometimes one at a time, sometimes in groups. Each statement or group of statements remains visible for a predetermined period of time (usually a matter of seconds). The student's task in each case is to decide whether each statement is true or false and record his or her decision by marking the appropriate space on a machine-readable answer sheet. After the last statement about one episode disappears, another episode appears and the process is repeated. Here is a brief description of one such episode and the statements that follow it:

The film clip shows a teacher standing before a bulletin board picture which shows several people boarding a jumbo jet airplane, discussing the picture with a second-grade class.

83. Most of the students were having difficulty with the main concept the teacher was trying to get across.

84. The teacher should have made contact with the boy in the checked shirt. [R].

85. The learning environment would have been better if the teacher had maintained tighter control.

86. This teacher was using the inductive method.

87. If the teacher had stopped to call for quiet it would have taken even longer to get her main point across.

Each statement is designed to give a student an opportunity to demonstrate a competency of one of the three types being assessed. In most cases, statements relevant to all three competencies follow each episode.

Assessing Perceptual Skills. Statement 83 is intended to give the student who is performing the exercise a chance to demonstrate a perceptual skill of the type Kounin (1970) has called "withitness." Because the pupils are no longer visible when the statement appears, the student would have had to perceive whether or not the pupils were puzzled while the episode was still visible, without any specific prompting to do so.

Statement 86 was intended to assess a perceptual skill of the type described by Smith (1969), the ability to recognize an abstract pedagogical concept as it appears in the "real world" of the teacher. In order to know whether Statement 86 is true or false a student would need not only to know what inductive teaching is but be able to recognize it when he or she sees it.

Assessing Professional Judgment. Statement 84 was intended to give the student an opportunity to demonstrate the ability to apply professional judgment to an interactive teaching problem. The symbol [*R*] that appears at the end of the statement indicates that while the statement was visible on the screen the relevant portion of the episode (in this instance, the behavior of the pupil in question at the critical moment) was also visible. This is done to minimize the effect of the student's level of perceptual skill as a factor in his or her response to this statement. Otherwise a student whose professional judgment was excellent might fail this task because of a weakness in perceptual skill.

Statement 85 was also meant to assess professional judgment, specifically whether the student was able to assess accurately the level of control maintained by the teacher. It was not deemed necessary to replay any part of the episode in this instance.

Assessing Professional Knowledge. Statement 87 was intended to assess the student's ability to apply professional knowledge to the solution of an interactive teaching problem, in this case, knowledge of the probable consequences of a contemplated teacher behavior. Correct evaluation of this response requires the student to apply what Smith has called "clinical professional knowledge" (Smith, 1983).

Measurement Properties. This simulation exercise, like the first one described, fulfills all of the conditions necessary for objective measurement of human performance. All students perform the same tasks; they record their own performances on machine-readable answer sheets; and the records can be read and scored by a computer. The full range of procedures used to revise paper-and-pencil tests (item analysis, internal consistency analysis, factor analysis, etc.) are available for use in refining this instrument.

The validity and reliability of the scores, therefore, depends on the tasks the students are required to perform. In other words, they depend on what the user builds into the exercise. It should not be difficult for the instructors in a program to select episodes and frame statements that measure students' progress toward the objectives of each of their courses.

If all of the episodes and statements, representing all of the courses in the program, are assembled into one exercise and administered to all students regardless of where they are in the program, the experience will not only be an important learning experience in itself, but will enhance other experiences the students

have as well. Students will realize that these realistic teaching problems become easier to solve as they progress through the program, and will see the relevance of their course work to the problems they will encounter as teachers more clearly than students who do not have this experience (cf. Medley, 1988).

Discussions of these and other approaches to the assessment of competencies may be found in Brinkerhof (1978), MacDonald (1978), Medley (1984), Pottinger (1978), and Shearron (1978), as well as in the references cited elsewhere in this discussion.

Assessing Interactive Performance Skills

The third and last example I describe was intended to evaluate interactive performance skills. Interactive performance skills are generally regarded as the most difficult competencies to measure objectively, because they can be demonstrated—and therefore assessed—only while the teacher is interacting with pupils in the classroom. This aspect of teacher competence must therefore be inferred from teacher performance. This not only makes such evaluations relatively costly and cumbersome to obtain; but also makes it particularly difficult to satisfy the first two of the three conditions necessary for objective assessment, that is, to have all of the teachers who are to be evaluated perform identical or equivalent tasks, and to obtain accurate, quantifiable records of each teacher's performance.

For an example of the use of measurement-based teacher evaluation to infer teacher competence from teacher performance I use an evaluation system developed for use in teacher certification. Since July 1, 1985, any teacher who applies for a certificate to teach in the public schools of Virginia receives only a temporary, nonrenewable certificate which is good for 2 years. Before receiving a renewable certificate, candidates must actually demonstrate minimum competence to teach in their own classrooms during their 1st year of teaching.

The *Teacher Performance Record,* or *TPR,* is the instrument used to assess teacher competence in the Beginning Teacher Assistance Program ("BTAP"). The TPR is the best available example of the application of the measurement-based approach to the evaluation of interactive teacher performance skills; therefore, the following description is somewhat detailed but confined as closely as possible to the concerns of this chapter, which are principally methodological. Readers interested on a more complete account of the

program and its instrumentation should consult McNergney, Medley and Caldwell (1988) and Medley, Rosenblum and Vance (1989). Let me begin with a brief description of how the program operates.

The Beginning Teacher Assistance Program

Procedures. At the beginning of each teacher's 1st year as a paid, full-time teacher, each one hired in the state of Virginia receives a set of materials which list and describe what are known as the 14 "BTAP competencies" (McNergney, 1988). Three visits to each teacher, each made by a different trained observer, are scheduled during the early fall at a time convenient to the teacher and the observer. The teacher is asked to plan activities during these visits which will enable him or her to demonstrate the possession of each of the 14 BTAP competencies.

Before each visit, the teacher indicates what he or she plans to do during the visit, and describes pertinent characteristics of the class, by responding to an open-ended questionnaire. When the recorder arrives for the visit, he or she collects this document from the teacher and later codes this information onto an Opscan form for use in scoring the teacher's performance. The recorder then spends 30 to 45 minutes recording behaviors in the teacher's classroom and the visit ends.

Only experienced educators not currently employed full time are trained and employed as BTAP recorders. The role of the recorder is very different from that of a supervisor who evaluates teacher performance with a typical rating scale. The BTAP recorder is not expected to evaluate the teacher; the recorder's task is limited to that of making an objective, accurate record of the teacher's performance and sending it to Richmond where it will be read and scored by a computer.

If a teacher fails to demonstrate at least 12 of the 14 competencies during these three visits, three more visits (by different recorders) are scheduled during the next semester. In the meantime the teacher is encouraged to attend special workshops in each area of competence he or she failed to demonstrate, which are offered in every region of the state. If necessary, three more visits may be scheduled during the third semester.

The Teacher Performance Record. The instrument developed to measure the 14 BTAP competencies (the TPR) consists of two Opscan forms. One form, the one the recorder uses in the class-

room, is called the *Classroom Process Record,* or *CPR*; it consists essentially of a list of teacher behaviors the recorder is to look for during the visit. The second form is a list of items about the teacher's plans and the setting in which he or she will be observed. The recorder looks for these items in the teacher's answers to the questionnaire filled out before the visit, and indicates which of these items were found by marking the appropriate spaces. The complete record of one classroom visit includes one of these forms and seven CPR forms, properly completed.

During the first 3 minutes of a classroom visit the recorder marks certain teacher behaviors listed on the CPR that are responsive to pupil behaviors (e.g., teacher praises pupil's answer to a convergent question; teacher rebukes off-task pupil) as they occur. At the end of the 3-minute period, the recorder stops observing and marks other behaviors listed on the CPR that occurred during the period, most of which are teacher initiated (e.g., checks understanding; gives overview) and items that describe the situation during that period (e.g., recitation; small group activities).

Before any beginning teachers were observed, data were collected with the TPR in a representative sample of 662 classrooms of practicing teachers throughout the state. These data were used in developing, refining, and standardizing scoring keys for the 14 competencies.

Defining Beginning Teacher Competence

A specification of competence for the beginning teacher is no different than that of any other teacher since, from their first day on the job, beginning teachers are expected to perform the same functions as any other teachers. The difference lies in which of the competencies relevant to the performance of these functions a teacher who has just completed a preservice preparation program offered by a college or university may reasonably be expected to possess.

What colleges and universities are best equipped to do is to communicate knowledge to students; in the case of a professional school or program, this knowledge should mainly consist of functional professional knowledge or, as it is often called, knowledge of "best practice of the profession." Although most professional teacher education programs also try to help students develop performance skills, few such programs, if any, have the facilities necessary to be more than minimally effective in this area.

Considerations such as these suggest that the highest, and most

important, level of competence that it is reasonable to expect beginning teachers to possess is functional professional knowledge—professional knowledge that the teacher is able to apply to the solution of teaching problems. The graduate of a professional teacher education program may be expected to know and be able to apply the "best practice of the profession."

There is no consensus in the teaching profession about what this knowledge is, about what is the "best practice of the profession." All we can say at present about what knowledge such a consensus will contain, when and if it is reached, is that it will include knowledge whose relevance has already been established by sound empirical research. We therefore decided to define the competence of the beginning teacher as *the ability to apply the findings of research on teaching to the solution of teaching problems in their own classrooms.*

A first approximation to this knowledge was determined by reviewing the relevant research, as summarized in a number of published critical summaries. (principally Brophy & Good, 1986; Good, 1979).

Indicators of Competence. From our reading of this literature we put together 70 relatively homogeneous clusters of teacher behavior which we called "indicators of competence." These were the behaviors we would expect to observe either more or less frequently in the performance of teachers who were not only familiar with these research findings but able to apply them in their own teaching. Presence of positive indicators and absence of negative indicators would be taken as evidence of *functional professional knowledge* of competencies of the type we wished to assess. Following are four examples (all positive indicators):

C1. Preparing outlines, reviews, and summaries, beforehand
C2. Beginning the lesson or unit with a statement of purposes
C3. Making interrelationships among parts of the lesson clear to learners
C4. Ending the lesson or unit with a summary or review

Competencies. The next step was to group indicators that seemed to the project team to go together into 14 larger clusters of behaviors which we called competencies, shown in Table 3.1. (The four indicators just listed defined *Competency C, Clarity* in the table.)

Although this set of competencies incorporates much of the findings of the research, it was not intended to be, nor should it be

regarded as, definitive. It contains some but not all of a body of professional knowledge that every beginning teacher ought to learn, and learn to apply, in preservice training.

Operational Definitions. The operational definition of each of the 14 comptencies, the basis for a scoring key to be used in deriving a measurement of the competency from a TPR record, takes the form of a list of classroom *events* identifiable in a TPR record, each of which exemplifies one of the indicators of that competency. Before defining what I mean by an event let me define three kinds of items that a TPR record contains.

Items. A TPR record shows three kinds of items relevant to a teacher's performance:

- *Teacher behavior items* are things a teacher does, like reprimanding a pupil, asking a question which requires a pupil to recall a specific fact, or checking pupil understanding.
- *Situational items* describe transitory aspects of the situation in which the performance occurred, such as whether a discussion was going on or whether the class was broken up into small groups.
- *Setting items* describe stable aspects of the context in which the performance occurred, such as whether the class was a kindergarten class or a high school algebra class, or whether or not it contained mainstream pupils.

Events. An event is defined basically by the cooccurrence of two items, one behavioral and one situational. One event occurs when a teacher asks a recall question, one which requires a pupil to recall a specific fact, (recorded as a behavioral item) during a drill session (recorded as a situational item). A different event occurs when a teacher asks a recall question during a discussion period (recorded as a situational item). Although the teacher behavior is the same in both instances, the relationship of the two events to teacher competence may be very different because of the differences in the situations in which the behavior occurs. When (in what situation) a behavior occurs may be just as important as what behavior occurs. Although this is not always the case—the effect of some behaviors (like publicly rebuking a pupil) tends to be the same regardless of context—it is true in enough cases that it seems critically important to make this distinction between classroom events and behavioral items.

It seems equally important to have observers record items instead of events. One good reason is that the number of items an observer must be trained to discriminate is much smaller. If we define five kinds of questions, four instructional strategies, and three patterns of classroom organization, the computer will be able to distinguish $5 \times 4 \times 3 = 60$ different events; but the recorder needs to learn to recognize only $5 + 4 + 3 = 12$ items.

Another reason is that items tend to be much easier to discriminate than events, because fewer cues are needed. And a third is that it seems to be easier to record behaviors objectively than events.

Adjusting for Differences in Settings. Setting items could also have been used in defining events, but it seemed more efficient to use the information they contained in a different way. In BTAP they were used to compensate for nonequivalence of tasks due to differences in the settings in which different teachers were evaluated.

First, each setting item was treated as a two-level variable reflecting presence or absence of the condition defined by the item. For example, one setting item was marked if the teacher was observed teaching high school; otherwise it was left blank. Another was marked if the teacher's class contained one or more mainstream pupils, otherwise it was left blank.

Next the raw score of each teacher on each item scored on any of the keys was determined. The raw score on an event initiated by the teacher is its total frequency over all three records. The raw score on an event defined in terms of the teacher's response to a pupil initiation is its frequency relative to the number of opportunities to respond provided by appropriate pupil initiations.

The raw scores on each event in turn were then correlated with all of the setting items in one multiple regression equation, using the scores of the 662 teachers in the norm sample. If the raw scores on an event were not correlated with the presence or absence of any setting item, they were standardized (converted to standard scores) in the whole sample of 662 teachers. A teacher's standard score on such an item indicates whether that event is more or less likely to occur in that teacher's class than in the average teacher's class (and how much more or less likely).

If the raw scores on an event were correlated with any setting variable or combination of such variables, the sample of 662 teachers was subdivided into two or more homogeneous subgroups, and scores on the item were standardized separately in each of the

subgroups. In such a case the teacher's standard score on the item indicates whether that event is more or less likely to occur in that teacher's class than in the class of the average teacher *in the same setting* (and how much more or less likely).

For example, how often a teacher uses public praise was found to be correlated with whether or not the teacher was observed teaching one of the "primary" grades. (i.e., kindergarten or one of the first three grades). Primary teachers praised pupils publicly significantly more often than teachers of other grades. The sample of 662 teachers was therefore divided into two groups, one containing only primary teachers, and one group containing all other teachers. The frequency of this event was then standardized separately in each group. Now when a primary teacher's TPR record is scored, the frequency of this item is converted to a standard score in the primary group so that that teacher's score on that event is compared with those of other primary teachers only. And when the record of any other teacher, is scored, the score is converted to a standard score in the group of other teachers so that that teacher's score is not compared with those of primary teachers.

This justifies the assertion that, in any instance in which a teacher's performance is affected significantly by the setting in which it is observed, each teacher's performance is compared only with the performances of other teachers in the same setting. It also makes it unnecessary to set up separate norms for teachers of different grades, subjects, and so forth.

Competency Keys. A temporary scoring key was constructed for each of the 14 competencies by first identifying a set of events that reflected the indicators that defined that competency, and summing the standard scores (with positive or negative weights as appropriate) in each record. Each temporary key went through a number of revisions to maximize its internal consistency, estimated by coefficient alpha. The current versions of the 14 keys have internal consistency coefficients ranging from 0.62 to 0.86 with a median value of 0.71.

Setting Passing Scores. Passing scores were based on estimates of the percent of teachers who were employed in Virginia at the time when the norm data were collected that lacked each of the 14 competencies. The estimates were obtained by sending descriptions of the competencies to a sample of school principals and asking them to estimate these percents. (Most of them were in the vicinity of 10%.) We then set the pass score for each competency at

the corresponding percentile in the distribution of scores on that competency in the norm sample. In order to earn a passing score on a competency, then, a beginning teacher had to perform at least as well as an experienced teacher regarded by her principal as possessing that competency. In order to qualify for a renewable certificate of competence, a beginning teacher must earn a passing score on 12 of the 14 competencies.

Meeting the Conditions for Objective Assessment

Let us consider the degree to which these assessments fulfill our three conditions for objective assessment of human performance.

1. Equivalence of Tasks. Nominally, the tasks set for all teachers are the same: to demonstrate as many of the 14 competencies as they can. But the nature and difficulty of the task each teacher faces depends in part on the setting in which the task must be demonstrated, and especially on the kind of pupils in the class. Three stops were taken to compensate for such variations in difficulty.

The first step was inherent in the way competence was defined, competence was defined as functional professional knowledge of certain research findings, that is, as the ability to apply these findings to teaching problems. If, for example, a teacher responds to disruptive pupils in the way the research recommends, he or she is demonstrating competence, even if the disruptive behavior continues or worsens. (If anybody's competence is called into question in such a case, surely it is that of the researcher!) This greatly reduces the effect on task difficulty of differences in the ways different classes respond to the same teacher behavior.

The second step was to use relative instead of absolute frequencies in scoring events defined in terms of teachers' responses to pupil initiations. For example, suppose that one research finding was that effective teachers incorporate unsolicited pupil comments into discussions more often than ineffective teachers do. Because this can only be done if a pupil makes such a comment, the difficulty of the item depends on how common such comments are in the teacher's class. Instead of merely counting how often this happens, then, we also count the number of unsolicited pupil comments, the number of opportunities a teacher has to incorporate such comments, and use the proportion of opportunities in which the event occurs.

The third step designed to reduce differences in task difficulties was the adjustment for measurable differences in setting variables already described.

2. Quantifiable Performance Records. A machine-readable record of each performance is made by a disinterested observer trained to observe and accurately record such performances. The accuracy of the record (and ultimately the validity of scores based upon it) depends only on the recorder's skill in recognizing and recording the items listed on the schedule, not on the recorder's expertise as a judge of teacher performance.

3. Machine Scoring. Records made by BTAP recorders appear to the computer exactly like test answer sheets.

I suggest that the TPR meets the conditions for objective measurement of human performance well enough so that the validity and reliability of any score on the instrument depend almost entirely on the items contained in the instrument and on how they are scored.

Because the events scored on the TPR represent only a crude first attempt at a sample of the events that distinguish effective teaching from ineffective teaching, the validities of the 14 scores derived from it must also be limited. Much of this limitation could be removed by revisions in the instrument itself that are perfectly feasible. The approach shows considerable promise, more than any available alternative.

IMPACT OF THE PROGRAM

The Beginning Teacher Assistance Program program is explicitly intended to improve teaching through the use of teacher evaluation. In doing so it proposes to use two major strategies. The principal strategy is to identify teachers otherwise qualified for teaching certificates who lack one or more competencies essential for satisfactory performance and offer them assistance in remedying these deficiencies. As its name implies, the program was conceived of primarily as an assistance program. The second strategy for improving teaching is to screen out, by denying renewable teaching certificates to, those teachers unable to remedy the deficiencies.

By the end of the 1986–1987 school year, the competence of almost 2900 teachers had been assessed at least once; one cohort of 669 teachers had had three opportunities to be assessed; and hun-

dreds of teachers had been offered the opportunity to improve their competence (with or without assistance from BTAP) and be assessed again. From these data we can get some idea of the impact of the program.

Impact on Teacher Education Programs. Perhaps the most important effect of the program is the impact it has had on the way teachers are prepared in the state. Since its implementation teacher education students are being made much more aware of the existence of research on teaching and of some of its findings than ever before, as well as of the importance of learning to apply these findings in their own classrooms. It is unfortunate that so many instructors in the teacher education programs of the state seem to have decided to respond to the program by coaching students to pass BTAP rather than by helping them understand the research and master the functional professional knowledge behind the evaluations. But as their students practice demonstrating the indicators of competence they cannot help becoming aware of and even trying out teaching strategies and tactics they might not otherwise encounter, and becoming aware of the research base for them.

We can get some idea of what has happened from the fact that the first 669 beginning teachers evaluated in the fall of 1985 scored, on the average, 4.4 T-score points higher than the 662 experienced teachers in the norm sample. This happened even though this first group of beginners had no clearer idea in advance of what the instrumentation would measure than the teachers in the norm sample. And yet only 56% of this first group qualified for permanent certification by demonstrating possession of 10 of the 14 competencies.

Since the fall of 1986, teachers have been required to demonstrate possession of not 10 but 12 of the 14 competencies in order to qualify. Despite this increase in difficulty, 69% of the group first assessed in the fall of 1986 qualified on their first attempt. This 13% increase over the 1985 cohort clearly indicates that something has changed in the way teachers are trained in the Commonwealth of Virginia.

Impact on Teacher Competence. A more direct way of gauging the impact of the program on teaching in the state is by examining what happens to teachers who do not qualify for renewable certificates on their first attempt. In order to qualify on their second attempt, such teachers must learn to demonstrate at least one, and usually more than one, of the competencies they failed to demonstrate the first time.

No fewer than 88% of approximately 300 teachers who failed to qualify in the fall of 1985 increased their competence enough to qualify in the spring of 1986. It has been suggested that this first group may not have taken BTAP very seriously until they learned from the press that more than half of them had failed to qualify. If this was true, part of this dramatic improvement in competence may be spurious.

Subsequent experience does not support this idea. A second group was first assessed in the spring of 1986, just after the news broke. About 100 of them failed to qualify, but 96% of them improved enough to qualify in their second attempt (in the fall of 1986). And 88% of the 400 who failed to qualify on their first attempt in the fall of 1986 also improved enough to qualify on their second attempt.

This strongly suggests that, although most teacher education students now take the evaluation seriously, many of them are not acquiring enough competencies during their preservice preparation to qualify for certification without further preparation. This is further confirmed by the fact that 95% of the only group that has completed the program (the group first assessed in 1985) eventually succeeded in demonstrating 12 competencies and qualifying for renewable certificates.

Since the program began operation, the number of graduates of the teacher education programs of the state able to demonstrate 12 of the 14 BTAP competencies has steadily increased. Most of those graduates who do not demonstrate 12 competencies manage to acquire the additional competencies they need after graduation. Thus although the program is not denying renewable certification to many candidates, it does seem to be improving teaching in the state.

SUMMARY AND CONCLUSIONS

Failure to make important distinctions between three aspects of teaching has frustrated most past efforts to use teacher evaluation to improve teaching. These aspects are: teacher effectiveness (defined as the impact of a teacher's performance on her pupils), teacher performance (defined as the deployment of a teacher's competencies on the job) and teacher competence (defined as the possession of repertoire of competencies—knowledges, skills, etc.—relevant to effective performance of a specified teaching function).

Valid evaluation of each aspect requires different procedures, and each has a different role to play in the improvement of teach-

ing. Valid evaluation of teacher effectiveness must be based on pupil performance; valid evaluations of teacher performance must be based on the teacher's own performance on the job; valid evaluations of teacher competence must be based on the teacher's performance under test conditions.

Valid evaluations could be used to improve teaching by identifying and eliminating ineffective teachers and replacing them with more effective ones. But valid evaluations of teacher effectiveness are almost impossible to obtain, partly because it is so difficult to isolate the effect of teacher performance on pupils from the many other powerful factors that also affect it, and partly because of a lack of instruments that measure most of the important outcomes of education.

Valid evaluations of teacher performance could be used to improve teaching by identifying substandard performers and either reassigning or replacing them. But valid evaluations are difficult if not impossible to obtain because they require a better understanding of the teaching-learning process than is currently available.

Valid evaluations of teacher competence can be used to improve teaching by identifying incompetent teachers and either replacing them with competent teachers or by helping them to become competent (by pinpointing causes of incompetence and providing remedial treatment). Valid evaluation of teacher competence is feasible by the use of existing knowledge of the nature of teacher competence and available assessment procedures.

The process of developing valid, reliable and objective procedures for evaluating teacher competence involves three steps: (a) specification of the teaching function the competent teacher is expected to perform, (b) definition of the competencies (knowledges and skills) a teacher needs in order to perform this function, and (c) development of an instrument consisting of tasks designed to elicit demonstrations of these competencies.

Most of this chapter is devoted to a description of procedures for performing the three steps and a presentation of examples of procedures that have been used to evaluate teacher competence.

REFERENCES

Berliner, D. C. (1979). Tempus educare. In P. L. Peterson & H. J. Walberg (Eds.). *Research on teaching: Concepts, findings and implications* (pp. 120–135). Berkeley, CA: McCutchan.

Berliner, D. C. (1986). In pursuit of the expert pedagogue. *Educational Researcher, 15*(7), 5–13.

Brinkerhof, R. O. (1978). Competency assessment: A Perspective and an approach. *Journal of Teacher Education, 29*(2), 21–24.

Brophy, J. E., & Good, T. L. (1986). Teacher behavior and student achievement. In M. C. Wittrock (Ed.). *Handbook of research on teaching* (3rd ed., pp. 328–375). New York: Macmillan.

Cruickshank, D. R., & Applegate, J. M. (1981). Reflective teaching as a strategy for teacher growth. *Educational Leadership. 38,* 553–54.

Denham, C., & Lieberman, A. (Eds.). (1980). *Time to Learn.* Washington, DC: National Institute of Education.

Good, T. L. (1979). Teacher effectiveness in the elementary school. *Journal of Teacher Education, 30*(2), 52–64.

Hayes, L. J. (1988). *A simulation test of teacher competence.* Unpublished doctoral dissertation, Charlottesville, VA: University of Virginia.

Howsam, R. B., Corrigan, D. C., Denemark, G. W., & Nash, R. J. (1976). *Educating a profession.* Washington: American Association of Colleges for Teacher Education.

Jackson, P. W. (1966). *The way teaching is* (pp. 7–27). Washington, DC. Association for Supervision and Curriculum Development and the Center for the Study of Instruction of the National Education Association.

Johnson, C. E., Okey, J. R., Capie, W., Ellett, C., & Adams, P. T. (1978). *Identifying and verifying generic teacher competencies.* Athens, GA: College of Education, University of Georgia.

Kounin, J. S. (1970). *Discipline and group management in classrooms.* New York: Holt, Rinehart & Winston.

MacDonald, F. J. (1978). Evaluating pre-service teachers' competence. *Journal of Teacher Education, 29*(2), 9–13.

McNergney, R. F. (Ed.). (1988). *Guide to teaching.* Boston: Allyn & Bacon.

McNergney, R. F., Medley, D. M., Aylesworth, M. S., & Innes, A. H. (1983). Assessing teachers' planning abilities. *Journal of Educational Research, 77,* 108–111.

McNergney, R. F., Medley, D. M., & Caldwell, M. S. (1988). Making and implementing policy on teacher licensure. *Journal of Teacher Education, 39*(3), 38–44.

Medley, D. M. (1984). Teacher competency testing and the teacher educator. In L. J. Katz & J. G. Raths (Eds.), *Advances in teacher education* (Vol. 1, pp. 59–94). Norwood, NJ: ABLEX.

Medley, D. M. (1988). An outcomes-based teacher preparation program. In W. J. Gephart & J. B. Ayres (Eds.), *Teacher education evaluation* (pp. 58–83). Boston: Kluwer Academic.

Medley, D. M., Coker, H., & Soar, R. S. (1984). *Measurement-based evaluation of teacher performance: An empirical approach.* New York: Longman.

Medley, D. M., Rosenblum, E. P., & Vance, N. C. (1989). Assessing the functional knowledge of participants in the Virginia Beginning Teacher Assistance Program. *Elementary School Journal, 89,* 496–510.

Pottinger, P. S. (1978). Designing instruments to measure competence. *Journal of Teacher Education, 29*(2), 28–32.

Quirk, T. J., Witten, B. J., & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the National Teacher Examinations. *Review of Educational Research, 43,* 89–113.

Shearron, G. F. (1978). Designing and improving instruments for measuring competence. *Journal of Teacher Education, 29*(2), 18–20.

Smith, B. O. (1969). *Teachers for the real world.* Washington, DC: The American Association of Colleges for Teacher Education.

Smith, B. O. (1983). Closing: Teacher Education in Transition. In D. C. Smith (Ed.), *Essential knowledge for beginning educators* (pp. 140–145). Washington, DC: American Association of Colleges for Teacher Education, ERIC Clearinghouse on Teacher Education.