1990

2. Teacher-Performance Assessments: A New Kind of Teacher Examination

Edward H. Haertel
Stanford University, haertel@stanford.edu

2

Teacher-Performance Assessments: A New Kind of Teacher Examination

Edward H. Haertel
Stanford University

During the last week of July 1987, 20 fourth- and fifth-grade teachers spent four days at a simulated assessment center in an elementary school. Each teacher completed 10 performance exercises on the teaching of equivalent fractions. The following week, 20 high school teachers of United States history spent four days completing a like number of exercises on the American revolution and the formation of the new government. These field tests were the culmination of over a year's work by the Teacher Assessment Project (TAP), sponsored by the Carnegie Corporation of New York, under the direction of Professor Lee S. Shulman at Stanford University.

The TAP prototype exercises represent a fundamentally new kind of teacher examination, based on structured observations of teachers' performance in situations designed to elicit the same kinds of knowledge and skills as they use in teaching, lesson planning, textbook selection, or related activities. Used in conjunction with more conventional examination formats and additional kinds of evidence (e.g., academic training or documentation of on-the-job performance), exercises based on some of these prototypes are expected to play an important role in the certification process

being developed by the recently created National Board for Professional Teaching Standards. Some of the TAP prototypes directly simulate activities that are part of teaching or preparing to teach. Others, like discussing the performance of another teacher viewed on videotape, are more remote from the day-to-day work of teaching. All of the TAP exercises are designed to elicit forms of knowledge and analysis that may be critical for expert teaching.

For purposes of the TAP's research, prototypes were developed around two specific topics: the teaching of fractions, in particular the equivalence of fractions, at the upper elementary level, and the teaching of the American Revolution and the formation of the new government in a high school course on American history. Different exercises in each of these two content areas require from 45 minutes to 3 hours to complete and call on teachers to plan a lesson, critique a videotape of another teacher presenting a lesson, discuss the use of specific instructional materials, analyze and critique a textbook, or teach a lesson of their own choosing to a group of six students. Teacher examinees respond to particular student questions and comment on student homework problems (in mathematics) or brief essays (in history). Another exercise in elementary mathematics requires teachers to discuss the relationships among a set of possible topics from a unit on fractions, to select an appropriate sequence in which to teach those topics, and to explain their selection. They also demonstrate or describe methods of using specific household articles for teaching the equivalence of fractions, discuss the advantages and disadvantages of teaching students to use different methods for solving fraction problems, and describe their classroom routines for checking mathematics homework. High school history teachers engage in a group-planning exercise, in which three or four teachers work together to plan a unit on a specified topic. In another exercise, each teacher designs packets of instructional materials for a particular form of cooperative group-learning activity. The 1987 field tests of these exercises yielded a rich and extensive data base. In two weeks, they generated roughly 200 videotapes, 400 audiotapes, several thousand pages of observer notes, and hundreds of pages of notes and other writing by the 40 participating teachers themselves.

POLICY CONTEXT

Teachers in the United States take tests of various kinds, for various purposes. Their classroom performance as student teachers is

observed and critiqued by supervising teachers and higher education faculty members. For state licensure, most complete objective paper-and-pencil tests like the National Teacher Examinations (NTE), or basic literacy and numeracy examinations like the California Basic Educational Skills Tests (CBEST). Increasing numbers of states are also turning to structured classroom observations, using instruments like the Florida Performance Measurement System (FPMS) or the Teacher Performance Assessment Instruments (TPAI) as requirements for obtaining a clear credential (Sandufer, 1986). Later, at the point of the school district's tenure decision and at intervals thereafter, teachers may be evaluated on the basis of brief, informal classroom observations by the principal (Bridges, 1986).

In an effort to improve teacher preparation, to help teachers through the first, difficult years in the classroom, and to encourage more beginning teachers to remain in the profession, some states are also planning or implementing teaching residencies following teacher education. These may offer opportunities for new forms of assessment by a designated mentor teacher at the local site, which have yet to be explored. In addition to written tests and classroom observations, evaluation of teachers on the basis of their students' test performance are becoming increasingly common, as discussed by Berk in chapter 8.

The structured performance assessments being developed by the TAP are prototypes for a new type of teacher examination, distinct from all the forms of teacher testing just described. Exercises based on these prototypes may be included in a voluntary examination for practicing teachers, developed and administered by the teaching profession itself. In time, this form of exercise may also find application in assessments of teacher education students, in teacher licensure, and perhaps in the implementation of career ladder, merit pay, or mentor teacher programs.

Assessments for Teacher Certification Versus Licensure

In their role of protecting the public from harm, state governments issue licenses to practice various professions, including teaching. Licensure tests are often required to assure a minimum level of safe and effective practice. In contrast to licensure, the term *certification* generally refers to a form of recognition controlled by organizations representing practicing professionals, for example,

the National Board of Medical Examiners. Certification attests to some level of mature and expert practice. Following this usage, the "teaching certificates" issued by most states to beginning teachers would be called "teaching licenses."

One key recommendation of the 1986 report by the Carnegie Task Force on Teaching as a Profession, *A Nation Prepared: Teachers for the 21st Century*, was the creation of a National Board for Professional Teaching Standards, "to establish standards for high levels of competence in the teaching profession, to assess the qualifications of those seeking board certification, and to grant certificates to those who meet the standards" (Task Force on Teaching as a Profession, 1986, p. 62). Board certification would not occur until a teacher had at least several years of classroom-teaching experience, and would be entirely distinct from the state licensure required for beginning teachers. Teacher-certification tests will be used for a different purpose and with more experienced examinees than licensure tests. For these reasons, it is appropriate that the certification process require forms and levels of expertise well beyond those expected of beginning teachers. Structured performance assessments may help to address this broader range of knowledge and skills (Shulman, 1987a).

The TAP is not creating a national teacher examination. Although the Carnegie Corporation of New York has sponsored both the TAP and the creation of the National Board, they are independent of one another. The exercises created by the TAP will serve as a library of prototype performance assessments to assist the Board in developing its teacher-certification tests, but will also be generally available to interested researchers and test developers.

There are no testing applications envisioned for which exclusive reliance on structured performance assessments appears desirable, but such exercises may be used in conjunction with other requirements and forms of examinations to improve teacher licensure tests, as well as in certification testing. The State of Connecticut is at the forefront in developing such exercises as part of its Beginning Educator Support and Training (BEST) program (Pecheone, Baron, Forgione, & Abeles, 1988). Together with California, Connecticut has also taken the lead in forming the New Interstate Teacher Assessment and Support Consortium, which is intended among other functions to share information on performance-based teacher assessments and to coordinate similar development efforts among participating states.

Structured Performance Assessments and Teacher Education

The form and content of high-stakes tests can significantly influence the instructional programs that help examinees prepare for them (Fredericksen, 1984). The use of structured performance assessments in certification tests is expected to have a positive influence on teacher education programs, because these exercises employ tasks directly relevant to teaching. If these exercises come to play an important role in licensure examinations, their influence on teacher-education programs may be even more pronounced.

Preparation for structured performance assessments would involve practice in planning lessons, critiquing textbooks, answering student questions, and actual teaching, as well as discussions of the reasons for approaching these tasks in one way or another. Some activities of these kinds are already present in many teacher-education programs. An increase in this kind of activity would arguably improve teacher education.

DEVELOPMENT OF EXERCISE PROTOTYPES BY THE TEACHER-ASSESSMENT PROJECT

Like other professions, expertise in teaching requires mastery of a distinctive knowledge base (Shulman, 1987b). Indeed, one of the hallmarks of a profession is the possession of specialized knowledge and skill acquired through formal training and usually apprenticeship. No one who has not been trained as a lawyer would have much chance of passing a state bar examination, and persons who have not graduated from accredited medical schools are not even permitted to sit for the National Medical Board Examinations (Lareau, 1985). A certification test for teachers should likewise assess a distinctive knowledge base.

Areas of Knowledge Assessed by the TAP Exercise Prototypes

Teachers must have mastered the subject matter they are to teach, and must be familiar with general principles of sound pedagogy, but in addition, they must develop specific expertise in the teaching of a particular subject matter (Shulman, 1987b). This *pedagogi-*

cal content knowledge includes a repertoire of effective instructional activities, knowledge of common student misconceptions and stumbling blocks, metaphors and analogies that can help students to grasp new ideas, information about available curriculum materials and their appropriateness in different situations, and other matters.

A fifth-grade teacher must understand fractions, for example, differently from a mathematician. The teacher must not only understand how to work with fractions, but must possess a store of analogies, instructional activities, and alternative explanations and solution procedures for various kinds of fraction problems. The teacher must also know when and how to use all this information for instruction. A third-grade teacher must not only know how to read, but must also know how to organize the component skills of reading to impart them to learners. A high school physics teacher must not only understand kinematics, but must also know how to make real for students the connections between mathematical symbols and the real-world objects they can represent, and how to prescribe instructional activities that can force students to confront their naive, or "Aristotelian" ideas about motion (McCloskey, 1983).

The TAP exercises were developed around specific subject matters and topics of instruction in order to permit the examination of this pedagogical content knowledge. In one exercise, for example, teachers are interviewed about algorithms like cross multiplying to determine whether two fractions are equal. (The fractions a/b and c/d are equal if and only if $ad = bc$.) Among other questions, teachers are asked whether they would teach cross multiplication as a method for checking the equivalence of fractions, whether this method could be used to explicate underlying mathematical principles, what other methods they would teach, when they would use each method, and what difficulties the teaching of the cross multiplication method might create for students in their subsequent mathematics instruction. The kind of knowledge required to answer these questions or to justify the answers is distinct from a knowledge of the underlying mathematics, but may be critical for effective teaching of fractions.

Structured performance assessments can also tap a teacher's knowledge of curriculum materials. One exercise developed by the TAP requires a teacher to critique a United States history textbook and to evaluate the soundness of the history presented, the quality of the writing, the book's appeal to students, and its appropri-

ateness for different kinds of students, among other factors. In addition to a general critique, the examinee must respond to short answer questions about the quality of specific sections of the text.

In addition to pedagogical content knowledge and curriculum knowledge, some of the exercises piloted by the TAP attempted an examination of teacher performance skills and collegial interaction. Skill in performance refers to the teacher's ability to perform in front of a class of students—at a minimum to be articulate and moderately engaging. It was assessed primarily in an exercise in each field test that required teachers to present a lesson of their own choosing, planned in advance, to a group of six students. Collegial interaction refers to the teacher's ability to interact effectively with colleagues. It was assessed primarily in a group planning exercise, in which three or four teachers worked together to plan a unit on American history.

An additional exercise that was piloted but not included in the field test examined the teacher's interpersonal skill in managing a classroom disruption. All of the exercises also tested to some degree the teacher's knowledge of subject matter, and the set of exercises as a whole required adequate communication skills, including listening and speaking as well as reading and writing.

It would be rash to claim that all of these different aspects of teacher knowledge and skill were thoroughly or even adequately examined in the exercises piloted, or even that all of them could in principle be adequately examined using structured performance assessments. The assessment center may be an inappropriate context for the measurement of some of these skills, especially collegial interaction, performance skills, and interpersonal skills in managing classroom disruptions. Nonetheless, these exercises may have the potential to significantly extend the range of different kinds of teacher knowledge and skill that can be measured.

Performance Exercises for Teacher Certification

Structured performance assessments for teachers are still in their infancy. Although models exist in the performance center approaches developed for personnel evaluation in industry, these are generally designed to assess more or less generic managerial and organizational skills. Even the exercises used in performance centers for the selection of principals, operated by the National Asso-

ciation of Secondary School Principles (Hersey, 1986; Landholm, 1986), require relatively little specialized knowledge of school organization or pedagogy (Aburto & Haertel, 1986).

Design of the summer 1987 field tests began with the identification of many more potential exercises than were ultimately included, and with an initial conception of the types of knowledge to be assessed that was somewhat broader than the final conception. From these preliminary ideas, a set of exercises was chosen for development to represent a range of different teaching situations (qualities of schools, communities, and learners), response modes (demonstration, verbal responses, written products of different kinds), and varieties of activities (teaching, preparing to teach, collegial interaction). Considerations of fairness to teacher examinees from different ethnic backgrounds dictated that exercises not depend on detailed knowledge or experience in highly specialized instructional settings, although there was a tension between this concern and the desire to provide examinees with as much context and background as possible for each exercise. For each exercise, points of vulnerability were identified, and an attempt was made to avoid including several exercises that shared a common weakness. Finally, each potential exercise was examined for its representativeness of some larger class of exercises that would be more or less parallel to the prototype. Ideas that appeared to defy replication were not pursued.

Each exercise chosen for development went through a process leading from an initial sketch to a preliminary script, pilot-test materials, pilot by the author of the exercise, supervised pilot by another examiner, preparation of training materials for field test examiners, and finally, inclusion in the field test. Group and individual reviews of each exercise were required at specified points in this process (Wilson, 1988).

The TAP was assisted in this work by exemplary teachers in elementary mathematics and in high school United States history. Some of these were teacher collaborators who served as paid consultants to the project, but a larger number participated in a teacher advisory panel or contributed in other ways. The teacher collaborators were observed in their classrooms teaching the focal content of each assessment; responded to a series of structured interviews about their own background and experience, their pedagogical methods, details of their short-range and long-range instructional planning, their methods of student assessment, and other matters; assisted in developing stimulus materials for the

exercises; and served in the field test as examiners. Together with the teacher advisory panel, they also served as subjects for exercise pilots and participated in extensive discussions of specific exercises, which led to numerous improvements.

The classroom observations and interviews involving the teacher collaborators built upon earlier and concurrent studies on the knowledge base of teaching. These were referred to as the *wisdom of practice* studies in the TAP, and helped especially to define the pedagogical content knowledge to be assessed.

Scoring Performance Exercises

In the TAP field tests, dramatic differences were evident among the performances of different teachers, and especially between beginning and highly experienced teachers, but it is one thing to recognize the variability of teachers' performances and another to derive reliable and valid measurements from them. By design, nearly all of the questions posed in the various exercises have several correct answers. In scoring, it is necessary to recognize the validity of alternative responses while maintaining distinctions among different degrees of response quality.

Scoring and interpretation of exercise performance were of concern from the beginning of exercise development, but work on scoring began in earnest after the 1987 field test. In the following months, preliminary scoring schemes for nearly all of the exercises were revised and elaborated, with as many as three successive scoring systems developed and applied for some exercises. The study of scoring culminated during the summer of 1988, when teachers from outside the project were hired to score the exercises using the final scoring systems developed. This final scoring is providing information about interrater reliability and about the strengths and weaknesses of alternative scoring methods. Preliminary information about scoring is available in interim reports (e.g., Haertel, 1988; Shulman, Haertel, & Bird, 1988). Only a brief sketch of the scoring can be presented here.

A set of five "scoring dimensions" has been developed to guide and organize scoring efforts. The most important of these are "content-specific pedagogy" and "subject matter knowledge," together with "professional responsibility," "class organization and management," and "pedagogy, sensitivity, and responsiveness to students." Each exercise is scored only for those dimensions it can

inform. Scores are assigned on a six-point scale, from AAA (Distinguished) through C (Questionable). The same six-point scale is used for all dimensions and for all exercises. Most scores range from AA (Commendable) down to B+ (Adequate) or B (Limited). The rating of A (Satisfactory) is considered borderline with respect to a certification decision, although of course the cutting score, the scale, the scoring systems, and even the entire approach taken may be changed by the National Board. Where evidence with respect to a dimension is thin, the rating is enclosed in brackets to distinguish clearly between the strength of the evidence available and the quality of the performance. The ratings are supplemented by brief narrative comments as required to call attention to unanticipated or atypical aspects of performance that might have a bearing on a certification decision.

The dimensions have been useful in organizing the work of scoring development, but their convergent and discriminant validity and their ultimate role in scoring have yet to be determined. It is possible that scores across dimensions will turn out to be highly correlated, in which case the separate scores might be of limited value. It is also possible that scores for the same dimension across exercises will not correlate as highly as scores across dimensions for the same exercise, which might also call the use of separate scores into question.

Two general approaches have been taken to scoring the different exercises. One strategy is "holistic," relying on descriptions of B+ and of AA performances with respect to each dimension scored. An examinee's performance is reviewed and summarized in a standard format specifying particular elements of the performance that should be noted. The performance summary is then compared with the two descriptions for each dimension. If it closely matches one or the other description, the corresponding rating is assigned. If it is between the B+ and the AA descriptions, a rating of A may be used; a performance that surpasses AA can be assigned the AAA rating; and so forth. Alternatively, descriptions were sometimes prepared for all six levels to better define the entire scale.

An alternative to holistic scoring is to identify discrete, scorable elements of the performance. These are presented in a checklist for each dimension, which is used to record those elements present in a given protocol. In some scoring schemes, this identification of elements is augmented with simple ratings for each element, or brief comments. The scorable elements for each dimension are then combined following a more or less explicit rule. Initially,

rather than inventing some arbitrary rule, scorers are encouraged to deliberate about each protocol and arrive at a judgment about the preponderance of evidence. Later, after sufficient experience with this kind of system, it may be possible to formulate an explicit rule that captures the sense of these deliberations and makes this step of the scoring procedure more objective.

NEED FOR FURTHER RESEARCH

Educational researchers have used a variety of paradigms to study processes of teaching and learning (Shulman, 1986a; 1986b), and impressive progress has been made, but the knowledge base of teaching has yet to be codified as clearly and completely as that of many other professions. Many existing teacher examinations and observational systems are justified on the basis of findings from process–product research studies, which may show no more than that the teacher behavior to be chosen as "correct" was found to correlate with some learning outcome, in some particular time and place, with some particular teachers and learners. Moreover, research on the kind of content-specific pedagogical knowledge that was the focus of the TAP's structured performance assessments has been especially meager (Shulman, 1986b). This presents an obvious difficulty for developing, scoring, and interpreting tests of teachers' distinctive expertise. There is no case law, or textbook, or published research literature that sets forth generally accepted and empirically grounded answers for every question asked in the TAP exercise prototypes.

One possible conclusion would be that development of structured performance assessments designed to measure content-specific pedagogical knowledge cannot proceed until there is substantial professional agreement on questions of how particular topics in the curriculum should be taught, but I believe that is unduly pessimistic. The work of developing scoring schemes and warrants for asserting the superiority of some answers over others has proceeded concurrently with the development of the exercises, and progress has been impressive. Many decisions about the acceptability of specific answers must be regarded as provisional, and further research will clearly be needed before exercises of this kind are used to reach significant decisions. But for virtually every question asked in any of the prototype exercises, some answers are clearly acceptable, and others are clearly deficient.

Grounds for Judging Answers in Performance Exercises

The answers to performance exercises include responses to direct questions; as well as demonstrations like responding to student questions or presenting a prepared lesson; products like lesson plans; and other scorable responses. It is useful to distinguish two related issues in using these answers to reach decisions about examinees. First is the problem of evaluating answers for correctness or quality. Second is whether these particular, scorable responses ought to be counted in reaching the decision at hand. The first of these issues is logically prior to the second. If there is no basis for distinguishing better from worse answers to a question, it obviously has no place in an examination.

Correctness of Exercise Responses

Granting that there is no one right way to prepare a lesson plan, critique a textbook, or answer a student's question, there are, nonetheless, some clear criteria by which answers can be judged. First is factual correctness. Content-specific pedagogical knowledge is bound up with subject-matter knowledge, and the content conveyed by a teacher's lesson should be consistent with the generally accepted views of subject matter specialists. Correctness and precision are sometimes matters of degree, of course, and it may be proper to teach school children some generalizations to which experts would take exception, but the principle stands that the content of teachers' instruction should be accurate. The TAP exercise prototypes also evaluated teachers' knowledge of curriculum. In constructing their responses, examinees were often expected to draw on their knowledge of how typical elementary mathematics textbooks are organized, for example, or the kinds of manipulatives and other instructional materials typically used to teach basic concepts about fractions. In the next round of TAP exercise development, focusing on literacy in the early grades, teachers will be expected to be broadly familiar with children's literature, and to be able to suggest appropriate readings for different pedagogical contexts and goals. (It bears repeating that all of the TAP prototypes were designed to assess more than subject-matter knowledge or knowledge of curriculum materials per se. If the goal were no more than measurement of information, less costly forms of assessment could be used.)

Second, teachers' scorable responses should comport with ac-

cepted general pedagogical principles. Lessons should have some discernible purpose and structure, explanations should be clear, vocabulary should be appropriate to the level of the children addressed. Instruction should proceed systematically, unless there is some definite and probably explicit rationale for proceeding otherwise. If a teacher is asked to present a lesson prepared in advance to a group of six well behaved children, then she or he should in some way monitor the engagement and understanding of all six of them, and not entirely ignore those who fail to raise their hands.

Third, where there are generally accepted answers to questions of pedagogical content knowledge, these provide a standard against which to judge an examinee's answer. In one exercise, teachers are shown a method for checking whether two fractions are equivalent, and asked, among other questions, what other methods they might teach children for checking the equivalence of fractions. The expected answers include algorithms for reducing both fractions to lowest terms, or converting them to decimals. Granting that a teacher might produce some unexpected answer that was neither clearly correct nor clearly incorrect, all of the answers to that question that were in fact obtained during the tryout of the exercise could be scored without difficulty.

It must be acknowledged that generally accepted answers are not necessarily correct. A consensus of teachers, even expert teachers, might represent no more than conventional wisdom, some mixture of truth and folklore. But the best available knowledge, even if imperfect, is appropriately assessed in a certification test. The limitations of a professional consensus are less of a problem for exercise development than the difficulty of *determining* a professional consensus about the proper instructional treatment of particular curriculum topics. The work of teachers is largely private and individual. A masterful lesson might be captured on film or videotape, but this is rarely done. Even within a single school, teachers often fail to discuss their instructional practices with one another. Journals for teachers in particular school subjects offer sensible and promising instructional ideas, but these tend to be fragmentary and to lack broad empirical support. Moreover, the correct answers to pedagogical questions depend on a host of contextual factors. Teachers must tailor their instruction to the needs of different learners, and the TAP staff has discussed at length how to score an examinee's statement that "it works for me," or "this is what I have found with my kids." (The decision has been that such a statement alone is insufficient justification for a questionable answer.)

The involvement of many practicing teachers in developing the TAP exercise prototypes has already been described. In addition, the TAP was guided by "expert panels" in each content area. These included practicing teachers highly regarded by their peers, nationally known teacher educators specializing in the content area, and university-based scholars in the cognate discipline. Each panel was co-chaired by a university faculty member and a classroom teacher. The expert panels reviewed ideas for exercises, critiqued the exercises at several points in their development, and discussed scoring criteria and the levels of performance that should be expected.

The use of these different sources of information provides some assurance that the scoring schemes developed by the TAP would find a degree of support among experienced and successful teachers, and provides a model that may be followed in further exercise development. However, the existing research base and the involvement of a handful of experts and teacher collaborators are not enough to justify expansive claims that the "knowledge base of teaching" has been discovered.

Continued research on teacher testing can accelerate knowledge growth in teaching. Commentary on structured performance exercises and discussion of the merits of different responses can help to bring forth an expert consensus on the solution of the pedagogical problems these exercises pose. Together with other initiatives toward the professionalization of teaching, teacher certification can also help to change attitudes and professional norms that have impeded the sharing and testing of new instructional practices, and can encourage more attention to pedagogical content knowledge in teacher-education programs. The development of an empirical and consensual knowledge base of teaching can and should proceed concurrently with research and development on teacher assessment.

Determining What Should Be Covered by a Teacher-Certification Test

Even after agreement is reached on the scoring of responses to structured performance exercises, the inclusion of these exercises on a teacher certification test remains to be justified. This is part of test validity, and is properly addressed under the conventional rubrics of content-related, criterion-related, and construct-related validity (American Educational Research Association, American Psychological Association, & National Council on Measurement in

Education [AERA, APA, & NCME], 1985). Some of the groundwork for studies of content and construct validity has already been carried out by the Stanford TAP, although of course all aspects of validity will merit careful reexamination once a set of operational examinations has been prepared by the National Board.

With respect to content validity, the structured performance exercises in a certification examination should each be manifestly relevant to the work of teaching. Taken as a whole, the set of exercises used should be representative of some definable domain. It should be clear what areas of knowledge and skill the exercises are designed to cover, and balanced coverage of those areas should be provided.

Criterion-related validity may be difficult to assess, because appropriate criteria against which to validate performance exercises may not exist and may be difficult to construct. Better performance on structured performance exercises ought to imply some capability for better performance in the classroom, but defining and quantifying better classroom performance will be a challenge. A premature insistence on criterion-related validity evidence using inadequate criteria could be unwise. It may be that criterion-related validity will best be addressed at the level of the entire teacher-certification process, with other lines of argument used to support the inclusion of particular kinds of exercises in the overall examination procedure. The design of sound criterion-related validation studies will not be possible until the newly formed National Board has had a chance to consider in much greater detail the form of their examinations, and the nature and level of the status that board certification is meant to confer.

Construct validity is close to the heart of the argument for new forms of teacher examinations. The fundamental justification for the cost of developing and administering these assessments is their potential to measure forms of knowledge and skill critical to expert teaching but difficult or impossible to assess using other forms of examinations. Continuing research is essential to better define pedagogical content knowledge, to establish the capability of structured performance exercises to assess it, and to show that the knowledge and skills measured by these exercises really do matter in the work of teaching.

Future research must also attend to plausible rival hypotheses about what the exercises measure, including hypotheses about cultural, gender, or other forms of bias. A thorough investigation of assessments relying on interview responses must address the possibility that verbal fluency or glibness can exert an undue influ-

ence on scores, for example. Subtle effects of interactions between the gender and culture of the examiner and the examinee must also be considered. These issues have been discussed repeatedly as the TAP prototypes have been developed and scored, but conscientious exercise development is no more than a prelude to the empirical studies that will be required.

Reliability, Validity, Efficiency

In designing a certification test for teachers, as in many measurement problems, there is a tension among the three goals of reliability, validity, and efficiency. Reliability refers to the replicability or reproducibility of judgments—across occasions, across raters or judges, and sometimes across forms or versions of a test or other assessment. Validity refers to a number of concerns that bear on the appropriateness of score-based inferences. It encompasses content validity—the extent to which the assessment is representative of the knowledge and skills required of teachers; criterion-related validity—the extent to which scores on the assessment are useful in distinguishing examinees capable of different degrees of proficiency in the classroom; and construct validity—the extent to which claims for the measurement of a distinctive knowledge base of teaching and for the distinctiveness of the dimensions used in rating can be supported. Efficiency refers to the costs of an assessment, in preparing the examination, in examinee time, in examiner time, and in scoring. It is roughly the case that if any one of these three constraints were relaxed, the other two could be satisfied. (Only roughly because an instrument's reliability places a statistical limit upon its criterion-related validity.) Given unlimited time and resources, a complex, rich observation of actual classroom practice over a long period of time, conducted by a panel of carefully trained observers, could probably provide an assessment of high reliability and indubitable validity. Shortening an assessment of this kind (improving efficiency) would compromise reliability, whereas substituting more efficient forms of assessment to satisfy the ends of reliability and efficiency together would likely reduce validity by employing assessment tasks that were less like actual teaching.

Closely related to the tension among reliability, validity, and efficiency is the goal of objectivity in measurement. Different forms of assessments vary in the amount of judgment required in scoring the performance of each individual examinee. At one ex-

treme, an objective multiple-choice test requires virtually no judgment at all on the part of the examiner. At the other extreme, an unstructured, holistic rating of performance following a brief classroom visit allows enormous latitude for the observer. Other things being equal, a more objective measurement is likely to assure more equitable evaluation of all examinees, and is likely to be more reliable than an instrument that calls for more judgment on the part of the observer. However, the single minded pursuit of objectivity can lead to a sacrifice of validity.

Standard Setting

The immediate purpose of teacher-certification testing is to arrive at pass–fail decisions about individual examinees. Setting standards for reaching these decisions will be a complex and difficult task, which properly devolves upon the newly created National Board. The standards established for these first teacher-certification tests will express to the public and the profession the meaning of board certification. If board certification is to contribute maximally to the professionalization of teaching, it must represent a significant level of expertise and attainment, but at the same time, the standard must not be perceived as unrealistic or unattainable. The proportion of candidates who succeed will strongly influence attitudes toward the certification program, as well as both the supply of and demand for board certified teachers.

Teacher performance exercises will provide only one of several different kinds of evidence that are expected to play a part in the certification process, and the board will have to decide what level of performance to require in each area. It will also have to decide whether strengths in one area will be allowed to offset weaknesses in another, or whether separate standards will have to be met for each component of the certification process. Clearly, standard setting for teacher performance exercises cannot be divorced from the purposes and context of the entire certification procedure, but it may be helpful to comment in general terms on a possible approach.

As with present teacher-licensure tests, judgmental methods are likely to play a large part in standard setting. These are methods relying on direct examination of tests by panels representing relevant constituencies (teachers, administrators, the public, etc.). When these methods are applied to multiple-choice tests, panelists are asked to make a large number of narrow judgments about

items, sometimes rating their difficulty or importance, or deciding which distractors a minimally competent examinee should be able to eliminate. These small judgments by many panelists are then combined by some arithmetic procedure to yield an overall passing score for the examination. Judgmental methods used for multiple-choice tests include the Angoff method (attributed by Angoff to Tucker), the Nedelsky method, and the Ebel method (Berk, 1986).

Although these methods enjoy some support in the psychometric community (e.g., Berk, 1986; IOX Assessment Associates, 1983a; 1983b), they have also been strongly criticized (e.g., Glass, 1978; Shepard, 1980). Authors taking exception to these methods have questioned the logical basis for assuming that panelists are able to make accurate judgments of the kind required. In the context of teacher licensure, if one argues that the items directly ask about things that classroom teachers need to know, then it follows that classroom teachers may be in a position to say which or how many items prospective teachers should be able to answer. But if the items are conceived as no more than indicators of knowledge or skills that teachers need, then the judgment task called for seems to depend on the panelists' knowing both the minimum level of the underlying knowledge or skill needed for acceptable teaching performance, and the regression of item performance on the underlying skill. In practice, panelists often seem uncomfortable with their ability to make the judgments called for (Shepard, 1980).

The judgmental standard-setting methods used with multiple-choice tests would be unsuitable for use with teacher-performance exercises, for at least two reasons. First, performance exercise protocols do not provide any natural breakdown into discrete, scorable units corresponding to objective test items. This is by design and may be intrinsic to whatever value these exercises have in eliciting distinctive areas of knowledge and skill. Second, no clean distinction can be drawn between scoring and interpretation for these exercises. On a multiple-choice test, scoring is an objective, mechanical procedure, logically and operationally distinct from the interpretive step of arriving at a pass–fail decision by comparison to a cutting score. On performance exercises, the methods of scoring envisioned so far all call for judgments and interpretations as part of the initial quantification of the examinee's responses.

New or modified judgmental standard-setting methods appear highly promising for performance exercises, because the kinds of tasks set and responses elicited are more like actual teaching. A

teacher panelist might not know how to judge the probability that a minimally competent teacher would know the answer to a multiple-choice question, but might be far more comfortable reading through examples of textbook critiques, lesson plans, or packets of instructional materials, and deciding which are of sufficient quality to warrant certification. After panelists reviewed, discussed, and rated selected protocols, these could be used to construct a rating scale to which other protocols could then be compared. Alternatively, statistical methods for "policy capturing" might be used to determine those quantifiable features of protocols that distinguished those judged acceptable versus unacceptable, leading to an objective formula for scoring and rating future protocols from the same exercise.

THE FUTURE OF TEACHER PERFORMANCE EXERCISES

Much work remains to be done before teacher-performance exercises are ready for operational use. Work on methods of scoring is ongoing, and reliability and validity are only now being examined. In addition to research and development on the prototypes themselves, the structure of the larger certification process will require further clarification, as will the organization and logistics of test administration. That being said, the task is well begun, and results to date are very encouraging. As the National Board, the State of Connecticut, and other organizations and states proceed with the development of these tests, a clearer picture of their strengths, limitations, and range of potential applications will emerge.

Performance exercises may contribute to the *definition* of teacher expertise, as well as contributing to its *assessment.* Standards for exemplary teaching practice must reflect a consensus of mature teaching professionals, but there have been few major forums for the deliberations necessary to arrive at such a consensus. The activities of developing and scoring performance exercises and designing a National Board examination are providing significant opportunities for reflection and discussion about what board certification ought to represent, and performance exercises can provide concrete cases to focus such discussions. The work of the TAP has already raised a number of issues that the National Board may need to address: Should certification imply that in addition to making sound pedagogical decisions, a teacher is able to explain the rationale for those decisions? To what extent should certifica-

tion attest to a teacher's specialized knowledge about teaching in different sociocultural settings? How should significant controversies or philosophical differences among teachers be resolved? Teachers who disagree fundamentally whether elementary mathematics instruction should give priority to teaching algorithmic skills or mathematics as a problem-solving process may approach some performance exercises in entirely different ways. The National Board must not espouse some narrow orthodoxy, but neither can it be entirely catholic in its conception of teaching excellence.

The National Board is creating a conception of exemplary classroom teaching, and its performance exercises will embody that conception. A few years from now, it may be possible to show empirically that performance exercises can distinguish between degrees of classroom expertise so defined, but for the present, an emphasis on criterion-related validity evidence would be premature. As stated earlier, appropriate criteria against which to validate these exercises may not yet exist. Just as thermometers were first designed to reflect rough and ready notions of hot and cold, so teacher performance exercises must first succeed in representing rough and ready notions of good pedagogical practice. Just as thermometers came in time to be *definitive* of temperature, so structured performance exercises may help to define teaching expertise.

REFERENCES

Aburto, S., & Haertel, E. H. (1986). *Teacher Assessment Project Study Group on Alternative Assessment Methods: Executive summary of the Assessment Technologies Conference* (Report No. 5). Stanford, CA: Stanford University Teacher Assessment Project.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137–172.

Bridges, E. M. (1986). *The incompetent teacher.* Philadelphia, PA: Falmer Press.

Fredericksen, N. (1984). The real test bias. *American Psychologist, 39,* 193–202.

Glass, G. V. (1978). Standards and criteria. *Journal of Educational Measurement, 15,* 237–261.

Haertel, E. H. (1988, April). *Quantifying the wisdom of practice.* Paper presented at the meeting of the American Educational Research Association, San Francisco.

Hersey, P. W. (1986). Selecting and developing educational leaders. *School Administrator, 43*(3), 16–17.

IOX Assessment Associates. (1983a). *Appraising the legal defensibility of the National Teacher Examinations for the state of Kentucky.* Los Angeles, CA: Author.

IOX Assessment Associates. (1983b). *Appraising the Preprofessional Skills Test for the state of Texas. Report number one: Test suitability and performance standards.* Los Angeles, CA: Author.

Landholm, L. J. (1986). Observations of participants: Center helps one's monitoring of strengths, weaknesses, *NASSP Bulletin, 70*(486), 24–25.

Lareau, A. (1985). *A comparison of professional examinations in six fields: Implications for the teaching profession.* Stanford, CA: Knowledge Growth in Teaching, School of Education, Stanford University.

McCloskey, M. (1983). Naive theories of motion. In D. Gentner & A. L. Stevens (Eds.), *Mental models.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Pecheone, R. L., Baron, J. B., Forgione, P. D., Jr., & Abeles, S. (1988). A comprehensive approach to teacher assessment: Examples from math and science. In A. B. Champagne (Ed.), *This year in school science 1988: Science teaching—Making the system work* (pp. 191–214). Washington, DC: American Association for the Advancement of Science.

Sandufer, J. T. (1986). State assessment trends. *AACTE Briefs, 7*(6), 12–14.

Shepard, L. A. (1980). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447–467.

Shulman, L. S. (1986a). Paradigms and research programs in the study of teaching: A contemporary perspective. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed., pp. 3–36). New York: Macmillan.

Shulman, L. S. (1986b). Those who understand: Knowledge growth in teaching. *Educational Researcher, 15*(2), 4–14.

Shulman, L. S. (1987a). Assessment for teaching: An initiative for the profession. *Phi Delta Kappan, 69,* 38–44.

Shulman, L. S. (1987b). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review, 57,* 1–22.

Shulman, L. S., Haertel, E. H., & Bird, T. (1988, April). *Toward alternative assessments of teaching: A report of work in progress.* Stanford, CA: Teacher Assessment Project, Stanford University.

Task Force on Teaching as a Profession. (1986, May). *A nation prepared: Teachers for the 21st century.* Hyattsville, MD: Carnegie Forum on Education and the Economy.

Wilson, S. M. (1988, April). *Planning the pedagogical decathlon.* Paper presented at the meeting of the American Educational Research Association, New Orleans.