DOCUMENT RESUME

ED 440 955 SP 039 192

AUTHOR	Stansbury, Kendyll
TITLE	What Is Required for Performance Assessment of Teaching?
INSTITUTION	WestEd, San Francisco, CA.
SPONS AGENCY	Office of Educational Research and Improvement (ED), Washington, DC.
PUB DATE	1998-01-00
NOTE	20p.
CONTRACT	RJ96006901
AVAILABLE FROM	WestEd, 730 Harrison Street, San Francisco, CA 94107-1241. Tel: 415-565-3000; Web site: http://wested.org.
PUB TYPE	Reports - Descriptive (141)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	Elementary Secondary Education; Evaluation Methods; Knowledge Base for Teaching; *Performance Based Assessment; Teacher Competencies; *Teacher Evaluation; Teachers
IDENTIFIERS	Capacity Building; Program Legitimacy; Teacher Knowledge

ABSTRACT

The last decade has been marked by experimentation with various methods of assessing teaching knowledge, skills, and abilities that are alternatives to the multiple-choice tests and loosely-structured observations commonly used. While there exists a large body of literature on multiple choice tests as a methodology, less technical information is available on desirable performance-based alternatives. This report looks at the state-of-the-art of performance assessment of teachers to: (1) identify important elements in the process; and (2) summarize lessons learned from the experiences of those who have developed such assessments. It is aimed at educators and policymakers who are exploring the use of performance tasks in teacher assessment systems, focusing on similarities and differences between the assessment of beginning and experienced teachers. The report examines the key stages of development of a reliable and valid teacher assessment system including: defining the purpose of the assessment; identifying components of the assessment system; building capacity; and building legitimacy. The report concludes with a summary of the state-of-the-art of teacher performance assessment and identifies issues still on the frontiers of development. (SM)

SP

What is Required for Performance Assessment of Teaching?

by Dr. Kendyll Stansbury

January 1998

WestEd

Improving Education through Research, Development and Service

What is Required for Performance Assessment of Teaching?

by Dr. Kendyll Stansbury

January 1998

This document is supported by federal funds from the U.S. Department of Education, Office of Educational Research and Improvement, contract number RJ96006901. Its contents do not necessarily reflect the views or policies of the Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the United States Government.

The last decade has been marked by experimentation with various methods of assessing teaching knowledge, skills, and abilities that are alternatives to the multiple-choice tests and loosely structured observations in common use. The National Board for Professional Teaching Standards (NBPTS) has now certified hundreds of accomplished teachers based, in part, on the results of these assessments. Individual states such as Connecticut, Florida, Kentucky, and New York are either experimenting with or using performance assessments for licensure. California funds a variety of formative assessments cuch as portfolios and classroom observations to guide individualized beginning teacher support in its Beginning Teacher Support and Assessment (BTSA) programs. The Interstate New Teacher Assessment and Support Project (INTASC), sponsored by the Council of Chief State School Officers, is piloting beginning teacher portfolios for licensure decisions in eleven states.

Teacher assessment produces a judgment about teaching as it is examined against a set of standards. That judgment may be to issue or deny a license, to decide whether to continue a teacher's employment in a school or district, to identify areas of need for support to meet minimal standards, or to suggest areas for future growth. A performance assessment looks directly at what teachers do or produce, either under simulated conditions or by instruction of students in the teacher's own classroom. Once they are constructed, multiple-choice or short written response assessments require little knowledge of teaching to score. In contrast, interpretation of these complex performance assessments of teaching requires specialized knowledge of teaching as a craft. For this reason, the shift to performance assessment is seen by some (Darling-Hammond, 1986) as a shift from a bureaucratic to a professional view of teaching.

While multiple-choice and other selected response tests excel at assessing the breadth of a teacher's knowledge, they are less powerful at assessing the depth of that knowledge and the ability to apply it (Stansbury and Long, 1992). Performance assessments do the latter at the expense of the assessment of breadth of knowledge. For this reason, the certification systems that have adopted performance assessments continue to rely on selected response tests in the battery of instruments that form their total assessment system. There is a large body of literature on multiple-choice tests as a methodology (e.g., Linn, 1988); less is known about the newer performance assessments.

This report looks at the current state-of-the-art of performance assessment of teachers to 1) identify important elements in the process and 2) summarize lessons learned from the experiences of those who have developed such assessments. It is aimed at educators who are exploring the use of performance assessments in a teacher assessment system, concentrating on assessments used to assess beginning and experienced teachers. Although performance assessments are common in pre-service education,they are not addressed in this review, partly to narrow the scope of the review and partly because a previous examination (Izu et al., 1992) found little rigorous examination of reliability and validity for preservice assessments. It focuses on the following components of the development and use of performance assessments in an assessment system:

- defining the purpose of the assessment

- components of an assessment system
- building capacity
- building legitimacy

It concludes with a summary of the state-of-the-art of teacher performance and identifies issues still on the frontiers of development.

Defining the Purpose of the Assessment

Before choosing instruments and constructing a teacher assessment system, it is necessary to be clear about the purpose for which the assessment is to be used. If the purpose involves high stakes, than the technical quality of every aspect of the assessment must meet legally defensible standards. If the consequences are less severe, e.g., identifying areas for professional growth, then the process need not be so rigorous. However, even in the case of low stake consequences, the technical quality of an assessment affects its usefulness and perceptions of its legitimacy.

There are three purposes for which performance assessment are used to assess teachers:
- licensure or certification
- hiring, continued employment, and career ladder advancement
- professional development

Licensure or Certification

Performance assessments are used for the initial licensure of beginning teachers by various states. Florida, North Carolina, South Carolina, and Connecticut use observations. In the Florida Performance Measurement System, observers record frequency counts of effective and ineffective behaviors which have been identified on the basis of positive or negative correlations with student outcomes. In North Carolina, South Carolina, and Connecticut, observers script lesson events and use their professional judgment to sort the evidence and evaluate it against pre-specified criteria. Kentucky is field testing a set of on demand and portfolio tasks that are completed during student teaching and the first year of teaching. New York asks candidates to submit a thirty minute unedited videotape showing classroom instruction in both whole group and non-whole group settings.

A consortium of eleven states participate in INTASC development and piloting of subject-specific portfolios. Portfolios in secondary mathematics and secondary English/language arts are being pilot tested; standards are being developed in secondary science and elementary education. The portfolios are scored by experienced teachers and teacher educators in the certification area according to a rubric. The state of Connecticut is using similar portfolios for licensure in the areas of secondary mathematics and secondary science and is developing and piloting portfolios in the areas of secondary English/language arts, secondary social studies, special education, and elementary education.

Educational Testing Service has a performance assessment called Praxis III that consists of an observation plus written responses to questions and brief interviews. California uses Praxis III as part of the Direct Application Program to license teachers in private schools with five or more years of successful teaching experience who have not completed a teacher preparation program. Ohio is piloting Praxis III for the licensure of beginning teachers.

NBPTS offers certification of accomplished teaching in the areas of Early Childhood Generalist, Middle Childhood Generalist, Early Adolescent Generalist, Early Adolescence/English-language arts, Early Adolescence through Young Adulthood/Art, and Adolescence and Young Adulthood/Mathematics. They are developing or field testing assessments in a number of other certification areas. These assessments are tied to a specific student developmental level, defined by an age span, and, except for the generalist certificates, cover a content area as well. Performance assessments are a major part of the battery of instruments used to produce certification decisions. For example, in the Early Adolescence through Young Adulthood/Art assessment, teachers complete a portfolio illustrating and commenting on examples of their teaching and professional activities. They then attend a one-day assessment center where they complete a series of on demand tasks.

Hiring and Continued Employment

Interviews and observations are commonly used by districts in hiring and continued employment decisions. However, a small scale study of teacher evaluation practices in California (Izu et al., 1992) found that the technical quality of these assessments is generally weak.

Professional Development[1]

Teacher induction programs are beginning to use observations and portfolios as means of assisting beginning teachers to identify needed areas of growth. Sometimes these assessments are conducted in connection with a licensure process. In induction programs, mentors or other support providers are often responsible for conducting the assessments. Even when others conduct the assessment, mentors often assist the beginning teachers in interpreting the results and trying to make the indicated changes in their practice. Educational Testing Service has modified Praxis III into an instrument called Pathwise for use in the formative assessment of teachers. Many school districts and teacher preparation institutions use Pathwise. In addition, California and Ohio are piloting programs which use Pathwise-trained mentors to collaborate with beginning teachers in designing an individualized support and development plan based on initial Pathwise results. At the end of the year, the beginning teachers are assessed again.

In some programs, these assessments are not conducted by others. Another use of portfolios is for teachers to evaluate their own teaching practice against a set of standards to identify areas for improvement and/or further growth. Two models of this type of assessment are the WestEd

[1] Portfolios and videotapes of teaching also offer professional development opportunities for teachers to examine and improve their practice or to experiment with and evaluate new teaching techniques. Although valuable activities, these are not instances of assessment if they are not examined against standards.

Beginning Teacher Support and Assessment (BTSA) portfolio and the Connecticut State Department of Education support seminars offering formative connected to a portfolio assessment for licensure.

The WestEd BTSA portfolio process trains support providers to assist a beginning teacher in reflecting on their teaching through constructing a portfolio. The portfolio includes three to six entries corresponding to different domains of the *California Standards for the Teaching Profession* (California Department of Education/Commission on Teacher Credentialing, 1996) and an entry on classroom-based research in one teaching domain. Support providers work with beginning teachers to select and reflect on evidence for each entry. Each entry, as well as the entire portfolio, concludes with a reflection on professional growth during the induction year that is informed by a domain-specific rubric.

The WestEd BTSA program is a locally-based generic teacher assessment model where districts or consortia make local adaptations to the general portfolio training. In contrast, the Connecticut State Department of Education is a subject-specific model using standardized assessment. Certification-specific support seminars are delivered regionally by two teachers who are experts in the area. The seminars focus on content-specific teaching practices. They differ between certification areas, but follow the same general framework. Each seminar begins by modeling criteria by which to evaluate teaching through comparing two samples of beginning teacher work, discussing a research article, or participating in and listening to a critique of a model lesson. The beginning teachers then meet in small groups to share artifacts from their own classroom such as a description of their students, a lesson plan, or a videotape of their teaching. They then critique their own and each other's work against the criteria presented earlier. The experienced teachers monitor group discussions, offering guidance as needed.

These two models address the use of teacher evaluation for the professional development of individual teachers. However, an alternative model might center on a group of teachers to come together to develop or to adopt standards for either teaching or for students. They then work together to examine both their individual and, if a school faculty, collective work against these standards and to identify both areas of strength and areas for improvement. This model ties together teacher assessment and student assessment in service of improving both instruction and student outcomes. The Western Assessment Collaborative at WestEd, which focuses on whole school change, uses such a model to focus on student assessment, but it could be adapted to address teacher assessment as well.

The specific purposes -- licensure/certification, hiring/continued employment, or professional development -- that a teacher assessment is intended to accomplish must be kept in mind as the components of an assessment system are developed.

Components of the Assessment System

The development and implementation of an assessment system requires attention to the

following components:

- standards
- assessment instrument(s)
- scoring framework
- exploration of technical quality

Standards

Teaching is a multi-faceted job requiring diverse knowledge, skills, and abilities in areas such as general workplace skills (e.g., coming to school on time), supervising students (e.g., enforcing safety rules during the loading of buses), and pedagogy (e.g., designing tasks and facilitating discourse so that students learn important knowledge and skills.) A single performance assessment can only focus upon a limited number of areas. The focus of an assessment is set forth in *content standards*, which define critical aspects of what a teacher should know and be able to do and set priorities among different aspects of teaching. The content standards communicate to teachers or teacher candidates the aspects of teaching that will be assessed. Standards are stated in a variety of ways, usually in a statement with some additional elaboration that further illustrates the standard. For example, here are standards statements from various sources addressing the learning environment:

California Standards for the Teaching Profession, **California Department of Education/Commission on Teacher Credentialing**

Teachers create a physical environment that supports positive social interactions and engages ALL students in purposeful learning activities. Teachers design and maintain safe learning environments in which students are treated fairly and respectfully and assume responsibility for themselves and one another. Teachers use instructional time effectively and implement procedures and routines that encourage students to participate in decision-making and to work independently and collaboratively. Expectations for student behavior are clearly established, understood, and consistently maintained.

Model Standards for Beginning Teacher Licensing and Development: A Resource for State Dialogue, **Interstate New Teacher Assessment and Support Consortium**

Teachers use an understanding of individual and group motivation and behavior to create a learning environment that encourages positive social interaction, active engagement in learning, and self-motivation.

Professional Standards for Teaching Mathematics, **National Council of Teachers of Mathematics**

The teacher of mathematics should create a learning environment that fosters the development of each student's mathematical power by--

- providing and structuring the time necessary to explore sound mathematics and grapple with significant ideas and problems;
- using the physical space and materials in ways that facilitate students' learning of mathematics;
- providing a context that encourages the development of mathematical skills and proficiency;
- respecting and valuing students' ideas, ways of thinking, and mathematical dispositions;

and by consistently expecting and encouraging students to --

- work independently or collaboratively to make sense of mathematics;
- take intellectual risks by raising questions and formulating conjectures:
- display a sense of mathematics competence by validating and supporting ideas with mathematical argument.

- *Early Adolescence through Young Adulthood/Art: Standards for National Board Certification*, **National Board for Professional Teaching Standards**

  Accomplished teachers establish environments where individuals, art content and inquiry are held in high regard and where students can actively learn and create.

As the different standards illustrate, groups differ in the specificity of teaching addressed in the standards. The standards excerpts above range from the generic to the very specific. The *California Standards for the Teaching Profession* apply to all teachers regardless of what they teach. The INTASC principles apply to beginning teachers, although the first principle addressing content knowledge, is elaborated separately for each certification area. The *Professional Standards for the Teaching of Mathematics* apply to all teachers of mathematics, regardless of grade level. Finally, the National Board for Professional Teaching Standards example not only applies to teachers of a single certification (usually content) area, art, but also to teachers of a specific age group, early adolescence through young adulthood (ages 11 to 18+).

While the content standards help determine *what* an assessment focuses on, another set of standards, *performance standards*, are needed to guide judgments of *how well* particular teaching performances linked to specific assessment tasks reflect the standards. Performance standards are not self-evident. They require preparation and practice with previously rated exemplars, or benchmarks, in order to apply them accurately. The performance standards become the scoring criteria for the assessment. Assessment instruments differ in their ability to capture different aspects of teaching. Therefore, the performance standards are not only instrument-specific, but are often task-specific so that they focus on key features of a specific task.

Assessment Instrument(s)

Since performance assessments provide more depth than breadth of information, it is important that they focus on aspects of teaching that are considered to be the most important to assess. Usually, financial considerations and time constraints limit the number of aspects of teaching to be assessed. Assessment instruments differ in their abilities to assess particular aspects of teaching. For example, observation assessments of a single lesson do not provide information on how a teacher sequences and weaves together instruction over time so that it has a cumulative impact. Once the focus of the assessment is identified, a choice of suitable assessment instruments can be made.

A variety of types of instruments capture a teacher's performance and thinking. They include classroom observations or videotapes that portray interactions between and among teachers and students, interviews that focus on a teacher's knowledge and thinking, structured tasks that simulate teaching events, and portfolios which may include entries developed using other assessment methods such as observations or videotapes.

Classroom Observations/Videotapes. Classroom observations and videotapes are similar in that they both examine teaching in a teacher's own classroom. Classroom observations

appear to be the dominant form of performance assessment of teaching for student teaching evaluations and for continued employment decisions. As noted previously, several states use observation or videotapes in licensure decisions. Videotapes are also an essential part of the INTASC and NBPTS portfolios.

Observations and videotapes have high degrees of authenticity because they portray applications of teaching decisions with a teacher's own students. They require some contextual information from the teacher, e.g., instructional objectives or student characteristics, in order to accurately interpret what is seen. The difficulty in maintaining the technical quality of videotapes by amateurs, especially the audibility of students and the teacher, makes it necessary to allow for multiple chances to produce a videotape of acceptable quality. Scoring of both observations and videotapes takes considerable training for raters to ensure reliability, because the lessons and teaching techniques being observed can vary considerably.

Interviews. Interviews appear to be commonly used in hiring teachers, sometimes at initial stages of consideration and more often as part of the final selection process. An analysis of practices in a small number of districts (Izu et al., 1992) found that standard rating forms were fairly common, but common questions were used less often. When interviews were conducted by a group, the evaluation criteria were more clearly defined, compared to when interviews were conducted by individual principals.

Both the state of Connecticut and NBPTS experimented with interviews in connection with structured tasks as a method of assessing teaching, and both abandoned them in favor of other methods. Interviews are difficult to evaluate, and may reward more verbally facile teachers. Contradictions between a teacher's description of their teaching and videotapes or observations of their interaction with students have been noted (Izu et al., 1992; personal communications with Connecticut assessment developers, 1995). Any further development of interviews as a teacher evaluation method will need to address the predictability of actual teaching behaviors from interview data.

Structured tasks. Structured tasks are sometimes used in hiring in the form of demonstration lessons or sample lesson plans to narrow the final field of candidates (Izu et al., 1992.) However, a more common use is in licensure or certification decisions. The state of Kentucky is experimenting with the use of structured tasks in licensure and they are a part of the NBPTS assessment center exercises. Because these structured tasks are administered in an on demand setting and because a large bank of tasks has not been built up, assessment developers are understandably reluctant to provide descriptions of them.

One study of prototypic structured tasks in development found that beginning teachers tended to have difficulty responding to hypothetical students and teaching practices that differed from their own (Stansbury and Long, 1992). However, structured tasks appeared to be a new assessment format for these beginning teachers, so it is difficult to tell whether the difficulty lay in the teachers' developmental level, their unfamiliarity with the methodology, or both.

Portfolios. Portfolios are growing in use in teacher preparation and are being developed for use in licensure and for teacher development. Common entries include lesson plans, student work with teacher responses, videotapes of teaching and learning events, and teacher reflections or commentaries which explain and evaluate teaching decisions. Portfolios must be carefully structured toward some purpose or they risk being little more than scrapbooks. The inclusion of student work and/or videotapes plus teacher reflections increases the degree of authenticity of a portfolio by reflecting both teacher decisions and the consequences of those decisions.

A special interest group of the American Educational Research Association devoted to teacher portfolios, Portfolios in Reflective Teaching and Teacher Education, has recently been established. Some California induction projects, sponsored by the state, use portfolios as a means of professional development and assessment for beginning teachers. INTASC is experimenting with integrated portfolios in secondary mathematics and secondary English, with other fields scheduled to begin development in the future. Portfolios are also a major component of the process for certification by the NBPTS in all areas.

Scoring Framework

Classroom observations/videotapes, interviews, structured tasks, and portfolios vary in their abilities to assess specific teaching abilities. The assessment purpose and focus not only help determine the appropriate assessment instrument, but also inform the development of the scoring framework.

The scoring framework for a performance assessments sets forth the criteria by which the performance is to be judged. It is what moves a task from being an interesting performance to an assessment of teaching. Scoring frameworks must apply across different teaching contexts and various teaching styles. Content standards that inform scoring usually embody a philosophy of teaching and learning, but scoring frameworks must also accommodate variations within that philosophy. (See Wilson and Wineburg, 1993 for an illustration of the complex issues involved in developing scoring frameworks that value different approaches to teaching.)

Scoring frameworks can either be holistic, analytic, or a hybrid. Holistic scoring seems to be the dominant form of scoring high stakes performance assessments of teaching. In holistic scoring, the assessors examine a performance against a rubric or set of rubrics that describes different levels of accomplishment. Rubrics are not self-evident, but must be accompanied by examples of performances that illustrating the different levels. Analytic scoring, in contrast, breaks a performance into pieces and rates each piece separately. Sometimes the ratings are then aggregated into a total score.

The experience of high stakes assessment has been that considerable time needs to be invested in bringing potential assessors to the levels of understanding needed to score reliably. The more scoring relies on professional judgments for interpretation of the performance as opposed to counting frequencies or noting the presence or absence of specific characteristics, the more preparation is needed. In addition, multiple assessors and a process for adjudicating differences are needed to ensure reliable judgments.

For professional development, the performance may be assessed against an ideal standard to determine the extent to which the performance does or does not reflect the standard rather than being assigned a comparative value. Teachers can then identify what they should keep on doing as well as possible areas of improvement. For example, a teacher using the *Professional Standards for the Teaching of Mathematics* to examine his/her teaching may consider the extent to which the tasks assigned to students embody the characteristics of mathematical tasks as defined in the standards. It should be noted that just as rubrics are accompanied by performance samples, the widely-praised *Professional Standards for the Teaching of Mathematics* illustrates each standard through several concrete examples.

Exploration of Technical Quality

The assessment instrument and accompanying scoring framework are judged through exploration of technical quality. This process is vital for high stakes assessments that must withstand potential legal challenges. However, they are equally important for teacher assessments with lower stakes, since the usefulness of an assessment is directly affected by its rigor. Three issues of technical quality will be described here: validity, reliability, and bias.

A teacher assessment is valid when interpretations of its outcomes correspond to levels of teaching quality. However, there are no universally agreed upon criteria for good teaching and no widely accepted method of measuring any criterion. This makes the determination of validity of teacher assessments challenging.

The traditional strategy of establishing validity for high stakes performance assessments, borrowed from that of selected response tests, has been to solicit reviews from groups of experienced educators, either though mail surveys soliciting responses to standards or task structures or in face-to-face meetings reviewing assessment materials. Review during development is invaluable, as it helps to identify potential problems in the assessment structure and materials. Reviews upon completion of the assessment are done by a wider group of educators who have not been previously involved in the assessment development.

Reliability is also an issue for assessments. Performance assessments involve complex judgments. Sufficient time must be provided to familiarize assessors and other users with the assessment methodology and how the scoring framework is applied. Differing interpretations of the standards and the scoring framework must be surfaced and resolved so that different assessors are interpreting the same performance in similar ways. Moreover, if multiple forms of assessment tasks are used, they must be shown to produce similar results through an equating process.

High stakes assessments study reliability through having two assessors score a sample of assessments and comparing the two sets of scores. Other assessments may rely on more informal comparisons of the interpretations of performances to check on consistency in applying scoring criteria. Experience with student assessments and with high stakes classroom observation assessments of teaching suggests that performance assessments result in lower levels of

reliability than selected response assessments, though acceptable levels of reliability can be reached.

With respect to generalizability, no published analyses of the generalizabilty of teacher assessments were found. However, the generalizability of student performance assessments across tasks is low, compared to the generalizability across multiple-choice tests, and the generalizability across assessment formats is still lower (Baxter et al., 1992). Equating of scores across formats is also in an embryonic stage, though various procedures based on either examinee data or professional judgment have been proposed (Loyd et al., 1995/1996). It appears that no one has even tackled the isse of how to measure the comparability of parallel assessments addressing different subject areas or different subject and grade levels such as those used by Connecticut or the NBPTS.

As noted above, many scoring frameworks for teaching performance assessments rely heavily on the professional judgment of educators to interpret performances. Under such circumstances, it is extremely important to guard against extraneous sources of bias. However, there are preliminary indications that the same ethnic/racial and gender differences noted in selected response tests are appearing in performance assessments as well despite efforts during development to guard against bias. Both NBPTS and INTASC have initiated studies of potential sources of bias to see if modifications in either assessment tasks or scoring frameworks are needed. In addition, INTASC has convened a bias review panel incorporating diverse perspectives, including teacher preparation, second language acquisition, urban teacher preparation, linguistics, cultural anthropology, and mathematics pedagogy. The panel will focus on the materials developed for the mathematics assessment, including the INTASC principles, the mathematics standards and the mathematics handbook that instructs teachers how to prepare their portfolio.

Building Capacity

The most rigorously designed assessments can be compromised if those implementing them do not have the necessary understandings of teaching and learning. Educators with the most years of experience were trained in an era of a behaviorist paradigm when teaching meant telling students what to do, providing guided practice in doing it, and increasing student ability to do it independently. The more recent constructivist paradigm stresses students as active inquirers, requiring a teacher to monitor students as they work and help them learn to direct their own thinking in productive directions. Educators who believe in one paradigm will have difficulty in using performance assessments that are grounded in another. The two paradigms differ in many ways, including what is important to pay attention to in teaching and the types of evidence of successful teaching. This dissonance between paradigms is not easily overcome, especially in the relatively short time allotted for assessor training.

A certain level of assessor knowledge of content and inquiry processes is also necessary. For example, an observer of a teacher directing a discussion among students about fractions needs to

be able to tell whether the discussion is mathematically sound, the extent to which the representations used are facilitating or hindering the conversation, and whether the teacher is showing a pattern of capitalizing upon or failing to act upon significant student comments.

Assessors who judge observations and portfolios must be willing and able to see teachers who teach in ways that differ from their own and still recognize effective teaching principles being applied. Teachers differ in teaching style, teaching context, and in the strengths and needs of their particular students. Assessors must be able to see through these differences to apply the scoring criteria. This requires a sophisticated understanding of effective teaching principles. Most teachers have limited exposure to other teachers' practices, so they have little experience in seeing across these differences.

These conditions of isolation do not contribute to developing a critical eye toward teaching practices. Moreover, teachers who have not experienced collegial conversations about teaching and learning are often reluctant to be critical of other teachers' practices. The disposition to take a critical stance, based on evidence, toward teaching practice is necessary to be an assessor of performance assessments. Unfortunately, many teachers find this threatening, even when they agree with the judgments made. Much of the preparation of teachers to use performance assessments must foster the disposition to be critical.

Although some beginning teachers who have recently completed teacher preparation programs may be more familiar than assessors with the constructivist paradigm and with content and inquiry processes, they have this knowledge at a novice level. Development of their content and craft knowledge as well as their abililties to judge their own teaching is an important part of a successful assessment system as well.

We need to work toward building an assessment culture so that teachers are not reluctant to make and defend judgments about both their own teaching and that of others. Little (1993) argues that current patterns of staff development and organizational structures make this difficult to happen. Teachers need time and opportunities to examine teaching practices and the resulting learning with colleagues in order to develop their critical capacity as well as to appreciate alternative solutions. Although a performance assessment may help create a taste for these type of activities, fostering these capabilities may require a restructuring of the school day and organization as well as reprioritizing the use of time to provide and support opportunities for reflection.

Building Legitimacy

Technical quality studies provide a legal and psychometric basis for the legitimacy of assessments. However, performance assessments, in part because they are a new methodology and in part because of the higher costs associated with their labor intensiveness, may need a broad base of legitimacy in order to survive. Perceptions of the legitimacy of a teacher assessment by teachers and other stakeholders affect the level of resources devoted to an

assessment system, the amount of time that participants are willing to commit to it, and the degree of opposition encountered. In Georgia and Louisiana, teachers filed suit against classroom observation systems for licensure, and the systems were thrown out by the courts. In contrast, in Connecticut, where a highly successful and valued assessment system was already in operation, experienced teachers are asking when the professional development seminars offered to beginning teachers that incorporate elements of the portfolio assessment will be made available to them.

Performance assessments of teaching are labor intensive for both teachers and assessors, if done with rigor. Their continued use requires high levels of commitment from all participants. In licensure and certification assessments, experienced teachers are not generally compensated for their services at rates that match what similarly experienced teachers receive for consulting services. Principals, who assess teaching as a part of their job, need to figure out how to use teacher assessment to accomplish other responsibilities, to reallocate time from other responsibilities for performance assessments, or to figure out how to use teachers in this role. Teachers who assess their own work for professional development must find the time and collegial support to do so Since performance assessments make high demands of participants in terms of time and effort, perceptions of the legitimacy of the assessment are extremely important.

The review and critique of assessment materials associated with formal studies of validity provides an opportunity to surface controversial issues that may diminish support. When done early enough, decisions can be made whether to change the assessment to avoid the controversy or to develop strategies for defending a particular position. Review by a broad variety of stakeholders can illustrate how perceptions of the assessment differ among role groups.

To avoid being blindsided by unanticipated controversies, it is good to be proactive in reaching out to stakeholders to familiarize them with new assessment processes. For example, some California BTSA programs report inviting representatives of the teacher unions to the beginning teacher and support provider training so that they have an opportunity to understand the assessment and how it is to be used. INTASC and the NBPTS explore different perceptions by involving representatives of various stakeholder groups in policymaking or advisory groups. The state of Connecticut schedules presentations to a variety of groups explaining its new subject-specific portfolio assessments and how they are and are not similar to the generic observation assessment in current use. These groups state and national professional associations, Regional Educational Service Center (RESC) staff, and teachers and principals previously active in the state induction program. In soliciting feedback from these various groups, the elementary team summarized issues that emerged and any resulting modifications in a document that was distributed to participating teachers, administrators, and RESC staff. According to the team, informal feedback suggested that this resulted in general perceptions that concerns had been heard and attended to, even when the response was that the issue lay outside the scope of the work being undertaken at that time.

Another strategy for gaining legitimacy is to link the assessment to lines of research on effective teaching or to specific reform agendas. The INTASC licensure assessments are linked to existing student standards and teacher standards developed by professional associations such as

the National Council of Teachers of Mathematics, National Council of Teachers of English. In Connecticut, assessor training for the observation assessment has become a desired qualification for applicants for administrative positions. The National Board for Professional Teaching Standards has taken a different approach in appointing a Technical Advisory Group of distinguished psychometricians. This panel advises the assessment development teams on development and conducts a series of studies on the technical quality of the battery of assessments for each certification area.

NBPTS has also worked with states and districts to create incentives in the form of rewards for teachers to take and pass these assessments, signifying the value of certification by significant stakeholders. These incentives include waiving most state requirements for the highest level of licensure, one-time salary bonuses, a percentage salary increase for the life of the certificate (ten years), or payment of a portion or all of the testing fees.

State of the Art in Performance Assessment of Teaching

With the allocation of necessary resources and time, including professional development of assessors and teachers, performance assessments have been shown capable of producing reliable and valid evaluations of teaching and have gained some measure of legitimacy among key stakeholders. Classroom observations for licensure have a track record over a decade old. Portfolios are more recent, but the NBPTS has found them to be of sufficient technical quality to be used in certification decisions, and INTASC results have been promising enough to warrant continued development.

Issues that are still on the frontiers of development of performance assessment of teaching include:

- agreement on teaching standards
- continued work on the technical quality of assessments
- building teacher capacity
- using performance assessments at the district level

Agreement on teaching standards. Teacher assessments have yet to generate the same level of controversy as student standards. (Cf. the controversies over "outcome-based education" in Pennsylvania and Virginia that resulted in revision of state content standards for students or about the *Curriculum and Evaluation Standards* for mathematics students produced by NCTM.) However, this is not necessarily because there is wide agreement on effective teaching practices, but more likely that teacher standards have received less publicity than student standards. Before than can be widespread agreement on teaching standards, there must be some agreement on student content standards. Since teacher performance assessments can only focus on a limited number of aspects of teaching, desired student outcomes affect the areas of focus chosen. Over time, curricular tastes in the United States seem to oscillate between an emphasis on knowledge of facts and procedural knowledge and an emphasis on more process-related skills such as the ability to inquire portrayed in science and mathematics student standards or the ability to produce meaning by both constructing and interpreting a variety of texts portrayed in the English student

standards. A balanced curriculum that addresses both sets of outcomes appears to be difficult to achieve. The recent attack on whole language programs and NCTM standards and advocacy for more emphasis on phonics and computation signals a continued struggle over curricular emphases.

Technical quality of assessments. Rigorous performance assessments of teaching (or of student learning, for that matter) are still new enough that technical issues are still being explored. Selected response tests still play a role in the evaluation of some aspects of teaching competence such as subject matter knowledge, where breadth of knowledge is important to assess. Much the same measures developed for use with multiple-choice tests are being used to evaluate the technical quality of performance assessments, despite their differences. It may be that as performance assessment methodology continues to develop, analogous ways of conceptualizing and establishing validity, reliability or generalizability may be developed that capitalize on the strengths of performance assessments and suggest appropriate limitations on their use.

Methods have been developed to avoid bias in selected response tests and to review for potential sources of bias in completed test forms. As systematic (and well funded) explorations for potential sources of bias are completed by INTASC and NBPTS, we may better understand methods of guarding against bias in performance assessments.

Performance assessments remain expensive to implement. The verdict is not yet in on whether possible benefits from using them -- improved instructional and personnel decisions, better targeted professional development, more sophisticated understandings of effective teaching -- outweigh their costs.

Building teacher capacity. High stakes performance assessments solve the problem of teacher capacity by carefully screening assessor candidates. Assessment models that depend on local assessors (e.g., schools who wish to use performance assessments) do not have as much latitude to reject potential assessors. We need to know more about how to develop the dispositions to take a critical stance toward teaching practice as well as how to foster the development of an assessment culture where teachers are both willing and able to assess their own teaching and to offer constructive feedback to their colleagues in a nonthreatening way.

Part of this involves recognizing that being a teacher means being a lifelong learner. Teachers should not be embarrassed by gaps in their knowledge of content or content pedagogy, but identify these as a sign for a need for further professional development. They also need to see collegial conversations with their colleagues about teaching and learning as vehicles for professional development. What structures and norms facilitate these dispositions, and how do schools move toward acquiring them?

While formative assessment for professional development requires less rigor than summative assessment for licensure, some standard for agreement across assessors is necessary to ensure that assessments are providing accurate feedback to teachers to guide their development. What

degree of reliability is necessary and how reliability can be effectively measured with limited resources remains unclear.

Using teacher performance assessments at the district level. Districts already rely upon performance assessments in the form of classroom observations and interviews for decisions about hiring and continued employment, but the technical quality of the implementation is not high, and teachers don't generally value any information that they receive from them (Izu et al., 1992.) The question of improving teacher and administrator capacity to use performance assessments is one issue that needs to be explored. There are other sets of issues related to management and structure. Strategies for standardizing the use of performance assessments for employment decisions across administrators need to be developed. Moreover, these strategies must achieve standardization while at the same time honoring variance in teaching styles and creative adaptations to better meet the needs of students. Administrators and other assessors need to be prepared and supported to make reliable judgments. Information from performance assessments needs to be more meaningful for teachers. For rigorous performance assessments to take root, either current ways of ensuring consistency and reliability need to be streamlined or a case needs to be developed for the benefits of devoting extensive district resources to the teacher evaluation process. This is particularly true for small districts.

In the absence of career ladders, experienced teachers who demonstrate teaching expertise and leadership in district and state policy initiatives are evaluated using the same process and the same formal criteria as beginning teachers. Is this a good use of time and resources? What manageable alternatives are there? Are these alternatives appropriate for small as well as for large districts?

Models of the use of student work as a means to look at schoolwide teaching practices and to suggest needed improvements exist, such as the Western Assessment Collaborative at WestEd, the PACE Portfolio project directed by Dennie Palmer Wolf at Harvard. We need to understand better the factors that affect a school's ability to adopt and use these models. A concern with equity requires that instead of using these factors as entrance criteria for working with schools and districts, that either alternative models or methods of working to change these factors be developed as well. Otherwise the differences in student performance between schools will continue to widen.

Conclusion

Performance assessments offer structured opportunities to focus on the links between teaching and student learning. These occasions can help individual teachers learn how to examine teaching practices -- their own as well as others' -- and thereby grow professionally.

In recent years, models of rigorous performance assessments, particularly observations and portfolios, have become or will shortly become available to assess teaching. Work in progress holds out hope for further methodological improvements. The potential of performance assessments for use in high stakes assessment is clear. A number of problems, including teacher

and district capacity as well as the streamlining of methods of ensuring technical quality, remain to be solved before they are available for widespread district use.

Bibliography

Baxter, G., Shavelson, R., Goldman, S. and Pine, J. (1992). Ealuation of Procedure-Based Scoring for Hands-On Science Assessment. *Journal of Educational Measurement*, 29, 1-17.

Darling-Hammond, Linda (1986). A Proposal for Evaluation in the Teaching Profession. *Elementary School Journal*, 86, 533-551.

California Department of Education/Commission on Teacher Credentialing (1996). *California Standards for the Teaching Profession* Sacramento, CA: Authors.

Interstate New Teacher Support and Assessment Consortium (1991). *Model Standards for Beginning Teacher Licensing and Development: A Resource for State Dialogue*. Washington, D.C.: Council of Chief State School Officers.

Izu, JoAnn, Long, Claudia, Stansbury, Kendyll, and Tierney, Dennis (1992). *Assessment Component of the California New Teacher Project: Evaluation of Existing Teacher Assessment Practices*. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Little, Judith Warren (1993). Teachers' Professional Development in a Climate of Educational Reform. *Educational Evaluation and Policy Analysis*, 15, 2, 129-151.

Loyd, Brenda, Englehard, Jr., George, and Crocker, Linda (1995/1996). Achieving Form-to-Form Comparability: Fundamental Issues and Proposed Strategies for Equating Perfomrance Assessments fo Teachers. *Educational Assessment*, 3, 99-110.

National Board for Professional Teaching Standards (1994). *Early Adolescence through Young Adulthood/Art: Standards for National Board Certification*. Detroit, MI: Author.

National Council for the Teaching of Mathematics (1991). *Professional Standards for the Teaching of Mathematics*. Reston, VA: Author.

Stansbury, Kendyll and Long, Claudia (1992). *Assessment Component of the California New Teacher Project: Summary Evaluations of Innovative Assessment Methods*. San Francisco, CA: Far West Laboratory for Educational Research and Development.

Wilson, Suzanne and Wineburg, Samuel (1993). Wrinkles in Time and Place; Using Performance Assessments to Understand the Knowledge of History Teachers. *American Educational Research Journal*, 30,4, 729-769.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

NOTICE

REPRODUCTION BASIS

☐ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☒ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").