Appearing in M. M. Kennedy (2010) Teacher Assessment and the Quest for Teacher Quality: A Handbook. San Francisco CA: Jossey Bass.

CHAPTER THREE

Assessment for Learning in Preservice Teacher Education

Performance-Based Assessments

Ruth Chung Wei, Raymond L. Pecheone

Stanford University

Teacher education programs have long used a customized set of curriculum-embedded assessments to support teacher candidate learning. High-stakes summative assessments have been left up to the states, which have usually tested basic skills, content knowledge, and, increasingly, pedagogical knowledge. However, recent changes in national and state accreditation processes have put program outcomes under the microscope, and the policy environment increasingly demands that teacher education programs provide evidence that their graduates have learned to teach. The quest for more valid licensing examinations has led some states and teacher education programs to look toward the use of performance-based assessments that measure teachers' competencies with more authentic instruments as the basis for licensure and professional development.[1] Preservice teacher credential programs across the country have independently created and implemented their own assessment systems that include performance-based approaches, focusing not only on teaching knowledge but on the application of this knowledge in practice. (For descriptions of a wide variety of assessment approaches used in teacher education programs, see Castle & Shaklee, 2006; Wise, Ehrenberg, & Leibbrand, 2008; Coggshall, Max, & Bassett, 2008.)

The strength of university-based approaches to assessment is that they are often used in formative ways to support candidate learning, while current licensure examinations are only summative in nature and do not generate detailed information about specific strengths and weaknesses in candidate performance

that can provide feedback to support candidate growth and program improvement. However, as accrediting agencies have moved toward outcome-based evidence of program effectiveness, the bar for meeting standards of reliability and validity has been raised. Even when local assessments are authentic, are thoughtfully implemented, and reflect program values, they may not have the psychometric properties that would allow policymakers to evaluate the validity or reliability of the information these assessments produce for the purpose of informing the licensure decision.

Several questions are raised by these seemingly competing (formative versus summative) approaches to teacher evaluation at the preservice level:

- Can curriculum-embedded performance tasks validly and reliably measure candidates' teaching knowledge and skill?
- Under what conditions can performance-based approaches support preservice teacher learning and professional development?
- Can a performance-based assessment be used to both inform summative licensure decisions for preservice teachers and support their professional learning and development?

To answer these questions, this chapter critiques the strengths and pitfalls of performance-based approaches to preservice teacher assessment, drawing on a review of research conducted to assess the technical quality and usefulness of these assessments for making high-stakes decisions and for supporting teacher learning. As part of this review, we highlight in greater detail a particular performance-based approach in preservice teacher assessment, the Performance Assessment for California Teachers (PACT), a project that has provided an innovative set of instruments to measure teaching effectiveness in a standardized and more reliable and valid way, and yet may also be used for formative purposes. The chapter describes this assessment system in some detail and summarizes the research documenting its validity and reliability and its formative function.

WHY PERFORMANCE-BASED APPROACHES?

Over the past two decades, teaching performance assessments (TPAs) have found wide appeal in the context of teacher education programs and teacher licensing for their innovative ways of assessing teacher knowledge and skill, as well as their formative impact on teacher learning and instructional practice. Spurred by the shift of the National Council for Accreditation of Teacher Education (NCATE) from process-oriented accreditation to accreditation based on systematic performance assessment of teacher candidates, teacher education programs have been forced to grapple with the question of how to measure

teacher competencies reliably and validly so that they can document the contributions of their programs to teacher quality.

Darling-Hammond, Wise, and Klein (1999) argue that more authentic assessments of teaching that simulate the complexities of teaching practice can improve the validity of licensing assessments and provide a valuable educational experience for teachers in the process of preparing for the assessment. Darling-Hammond and Snyder (2000) identify four characteristics of authentic assessments of teaching: (1) the assessments sample the actual knowledge, skills, and dispositions desired of teachers in real teaching and learning contexts; (2) the assessments integrate multiple facets of knowledge and skill used in teaching practice; (3) multiple sources of evidence are collected over time and in diverse contexts; and (4) assessment evidence is evaluated by individuals with relevant expertise against an agreed-on set of standards that matter for teaching performance. Darling-Hammond and Snyder highlight four assessment tools that meet these criteria: cases, exhibitions of performance, portfolios, and problem-based inquiries (or action research).

In their description of a variety of TPA formats, Long and Stansbury (1994) include portfolios that document a teacher's actual teaching experience over a specified period of time; semistructured interviews that ask teachers to answer a standardized set of questions designed to assess their knowledge and pedagogical skills in a particular content area, or to perform specific tasks such as designing a lesson unit, planning a lesson, and evaluating student performance; and semistructured simulation tasks that require teachers to respond in writing to a set of tasks in a classroom teaching scenario.

The performance assessments developed in the past two decades by the Interstate New Teacher Assessment and Support Consortium (INTASC) and the National Board for Professional Teaching Standards (NBPTS) represent the best-known national initiatives to develop alternative performance-based teacher assessments. These assessments are based on complex and holistic views of teaching and validated professional teaching standards, and represent authentic measurement tools that are context sensitive, longitudinal, and individualized (Darling-Hammond, 2001).

Researchers argue that in addition to being more authentic measures of teacher performance, innovative performance-based approaches to teacher assessment provide powerful professional development opportunities and stimulate teacher learning (Athanases, 1994; Anderson & DeMeulle, 1998; Darling-Hammond & Snyder, 2000; Davis & Honan, 1998; Haynes, 1995; Lyons, 1996, 1998a, 1998b, 1999; Rotberg, Futrell, & Lieberman, 1998; Tracz, Sienty, & Mata, 1994; Whitford, Ruscoe, & Fickel, 2000). However, not all TPAs are alike. They come in a variety of formats including case studies; tasks that ask teachers to analyze student work, evaluate textbooks, analyze a teaching video, or solve a teaching problem; lesson planning exercises; and portfolios.

There are a variety of tasks and activities that are called ''performance assessments'' and there is even variation in how portfolio assessments are defined. (See Chapter One, this volume.)

Because TPAs vary widely in format and content, it is important to distinguish among them and to evaluate their technical quality (validity, reliability), particularly if they are to be used for summative licensing decisions. In addition, it is important to investigate the conditions under which performance-based assessments can serve a formative purpose to support teacher learning and reflection.

CURRENT STATE OF THE FIELD

A 2006 survey[2] conducted by the American Association for State Colleges and Universities (Wineburg, 2006) found that the four most common means of collecting evidence about the effectiveness of teacher preparation programs were (1) observation systems that included faculty-developed rubrics and program standards; (2) surveys of cooperating teachers, schools principals, and program graduates, both during the program and beyond; (3) work samples and portfolios of teacher candidates; and (4) state certification tests (for example, Praxis I and II). The study also indicated that while state colleges and universities were expending enormous energy and resources to assess preservice teachers and compiling data on their programs, most of the assessments lacked evidence of validity and reliability.

In this review, we evaluate curriculum-embedded performance-based approaches to assessment that are in current use in preservice teacher preparation programs. While exit and follow-up surveys or aggregated results of state certification tests may have some value as measures of program or teacher quality, we do not believe they provide sufficient information with enough rich detail about the qualities of individual teachers. In addition, neither of these types of instruments provides information that is formative for both individual teachers and programs.

In our review of the literature, we identified four primary types of curriculum-embedded performance-based assessments that are in wide use across preservice teacher credentialing programs:

- Observation-based instruments and systems
- On-demand performance tasks
- Child case studies
- Portfolio assessments and teacher work sampling

While there may be other types of performance-based assessments being used in some programs, these four types represent genres that are relatively

well known, and the examples we use have been documented in research articles. This research review does not purport to account for every available research source or every example that merits review within each of these assessment genres.

In our analysis of the strengths and weaknesses of these different assessment approaches, we used three primary evaluative criteria:

1. How useful is the performance-based assessment for formative purposes?
   - Does the assessment promote teacher reflection and learning and lead to professional growth for individual teachers?
   - Does the assessment provide immediate and useful information for programs that can drive program faculty learning and inform revisions of program content and design?
2. How credible and defensible is the performance-based assessment for summative purposes?
   - Is there evidence that the assessment is based on valid constructs and aligned with validated teaching standards and professional expectations?
   - Is there evidence that the assessment can be scored reliably across scorers and sites and that there is a scoring moderation process for ensuring the fairness of summative decisions that depend on the scores?
3. How practical and feasible is the performance-based assessment (when there is information on this aspect of the assessment approach)?
   - What are the practical resources, such as time, human resources, and technology, required to implement the assessment?
   - What are the financial resources required to implement the assessment?

OBSERVATION-BASED ASSESSMENT

Observations of teacher candidates within the context of their student teaching placements is one method of performance-based assessment that is likely to be used in all credentialing programs. Observation-based ratings instruments have a long history and have evolved over time. Throughout their history, observation forms and checklists have focused on specific behaviors that reflected the dominant view of effective teaching (Arends, 2006b). In the past,

these instruments reflected a simplistic view of effective teaching as a set of discrete countable behaviors. In this age of accountability and state regulation, observation instruments and ratings systems often lift language straight out of the state teaching standards or national standards (for example, INTASC model standards), as though these standards statements represent valid constructs in themselves.

Observations of teaching, even if the sampling is infrequent and of short duration, are considered in teacher education to be indispensable to the evaluation of a candidate's readiness to teach. In some programs and circumstances, student teaching evaluations based on observations can trump all other evidence of candidates' proficiencies. For example, a candidate could be earning high marks in courses, but exhibit extremely negative behavior and lack of rapport with students. Such individuals are often counseled out of a program to minimize harm to students. While many teacher education programs place great faith in their observation protocols, the technical quality of the observation instruments and procedures in current use leaves much to be desired.

While research has found that assessors can be trained to reach a high level of interrater agreement on observation ratings instruments, most credential programs do not have the time or resources to provide sufficient training to supervisors and cooperating teachers to achieve an acceptable level of agreement. In addition, the quality of the scales used to score teaching performance varies. In some cases, teachers are scored as having ''met'' or ''not met'' the performance criteria, and in others, the scales may have four or five levels of performance, usually from novice to advanced or expert levels. The problem with many observation ratings instruments used to evaluate student teachers is that there are no descriptions of what performance would look like at these various levels. This means that evaluators (supervisors and cooperating teachers) are left to their professional judgment to decide what is considered proficient or passing performance. In more carefully constructed scales, rubrics describe in detail a developmental continuum of performance with clear indicators of expected performance at each score level (an example is the New Teacher Center's Continuum of Teacher Development). These descriptors not only bolster the ability of evaluators to score reliably, but also serve a formative purpose for student teachers by providing clear and concrete images of more advanced performance so that they have something to strive toward. Finally, while four- to five-point scales used to assess teaching performance appear to provide enough variation in scores so as to differentiate performance among beginning teachers, research and practice indicate little variance in the ratings candidates receive, with almost 100 percent of candidates who successfully complete a credential program receiving at least the minimum ratings required for passing student teaching.

Observation-Based Teacher Education Tools

An entire chapter could be written on the strengths and weaknesses of observation-based evaluation instruments; however, we focus on a few examples of instruments for which documented evidence about their technical quality could be gathered to evaluate their merit as both summative assessments of teaching performance and instruments that can serve formative purposes.

Washington Performance-Based Pedagogy Assessment. Washington State is currently using an observation-based teacher assessment at the preservice teacher education level. The Washington Association of Colleges of Teacher Education worked with the Office of Superintendent of Public Instruction to develop the Performance-Based Pedagogy Assessment (PPA), designed to be used during student teaching to assess candidates' ability to:

- Set clear learning targets and assessment approaches
- Use empirically grounded instructional techniques
- Engage traditionally marginalized students
- Effectively manage classroom activities and students

The PPA also places a new focus on evidence of student learning. Candidates are required to design and implement an assessment that provides evidence of student learning. The instrument requires at least two observations, and prior to the observations, candidates provide assessors with a description of their class and student characteristics, their lesson plans and planned assessments, and a rationale for the plans (Wasley & McDiarmid, 2004).

The PPA evaluates preservice teachers across ten dimensions of teaching; five are scored based on the written ''sources of evidence'' provided prior to the observation, and five are scored based on the observations and evidence of student learning presented by the student teacher. Each of the dimensions has between four and nine analytical scoring criteria on which candidates are scored as having ''met'' or ''not met'' the criteria or the evidence is ''not observed.'' (There are no level descriptors or indicators that specify what it means to have met or not met the criteria.) In order to pass the PPA, candidates must be scored as having met all fifty-seven criteria evaluated across ten dimensions, with evidence collected during two or more cycles of observation. (See Office of the Superintendent of Public Instruction, 2004, for specific directions to teacher candidates and the scoring rubrics.) Candidates are evaluated on the PPA by their university supervisor and the cooperating teacher.

Technical Quality. The validity of the Washington PPA rests on its alignment with validated standards for the teaching profession. The ten scoring dimensions of the rubrics are derived from the ten standards of the Washington Administrative Code (WAC): Effective Teaching Requirements for Teacher Preparation Program Approval. A validity study (Tisadondilok, 2006) examining the alignment of the Washington PPA against the INTASC model standards indicates that the INTASC standards are mostly aligned with the ten WAC standards, on which the PPA assessment rubrics are based. Tisadondilok also examined the construct and consequential validity of the PPA. She found that most faculty and supervisors at one university who had experience using the PPA instrument to evaluate teacher candidates' instruction felt confident that the PPA allowed student teachers to demonstrate their knowledge and skills. The percentage of faculty and supervisors expressing confidence in the construct validity of the PPA instrument and process ranged from 68 to 89 percent across the WAC standards. Interviews of these faculty and supervisors, however, indicated that they felt that there was also some construct irrelevance in the PPA, too many scoring criteria, and a redundancy in some of the criteria. In addition, faculty did not feel that the passing standard (meeting all fifty-seven of the criteria across the ten dimensions) was reasonable or fair.

There is no available information on the reliability of the Washington PPA instrument or the protocols for training supervisors and cooperating teachers to observe and score. However, given that the scale is a two-point scale (met or not met), and all candidates who are granted a license have to have met all of the scoring criteria, the ability to assess score reliability is severely threatened because there is little to no variation in scores. A two-point pass/fail scale also makes the rubrics less educative for teachers and program faculty. It provides neither detailed diagnostic information about candidates' performance nor detailed images of highly effective performance to guide candidate or program learning.

Impact on Candidate Learning and Program Improvement. In terms of consequential validity, Tisadondilok's (2006) interviews with faculty and supervisors at one university suggest that implementing the PPA resulted in a common understanding within and across universities in the state about what constitutes effective teaching, that the instrument and the WAC standards helped teacher candidates gain a clearer understanding of what is expected of their teaching performance, and that it may have prompted supervisors to pay greater attention to evidence of student engagement and learning, in contrast to the previous focus on student teacher behaviors. In addition, some supervisors reported that the PPA had made their observations and evaluation of student teachers more systematic and formal. But some supervisors expressed resentment about a substantially increased amount of paperwork during observations that led them

to pay less attention to the teaching itself. Finally, the faculty and supervisors were asked to discuss the impact of the PPA on teacher candidates. Most of the discussion focused on candidates who failed the PPA and subsequently were denied a license or were counseled out of the program and efforts to support teacher candidates who failed the first time by helping them to meet the scoring criteria on the PPA.

Absent from these discussions was evidence that the university was using results of the PPA to inform program review and revision or that faculty used the results to support teachers in improving their teaching skills (beyond helping them pass the assessment). Yet a portion of the PPA is inherently formative: candidates are required to submit their written materials (characteristics of their classes and students, a lesson plan, planned assessments, and a lesson plan rationale) to their supervisors and cooperating teachers prior to the observation to receive feedback. Candidates then revise their written materials and resubmit them to be scored on the PPA rubrics. In addition, because candidates may take as many tries as is necessary to pass the PPA and are evaluated on the PPA instrument by their cooperating teacher and university supervisor (rather than an external evaluator), results of the PPA evaluation may inform the mentoring received from their cooperating teachers and supervisors to respond to identified weaknesses in their instruction.

Teacher Work Sample Methodology: Observation Component. Another observation instrument that was developed by a teacher education program and is used in the context of student teaching is the observation component of the Teacher Work Sample Methodology (TWSM), developed originally by Del Schalock and colleagues at Western Oregon University. (See the section on structured portfolios later in this chapter for a detailed description and summary of the TWSM as a whole.) Rather than requiring teachers to record their instruction on videotape, the cooperating teacher and university supervisor are responsible for observing and evaluating teacher candidates' teaching during the two- or three-week unit documented in the TWSM. One version of the scale (Rating Teaching Strategy Decisions Made During TWS Implementation) can be found in Girod (2002).

Instruction is evaluated along two major dimensions: ''Establishing a classroom climate conducive to learning,'' which has eleven separate evaluative criteria, and ''Engaging pupils in planned learning activities,'' which has six separate evaluative criteria. Similar to the Washington PPA, the TWSM observation instrument is scored on pass/fail basis: met and not met (this evolved from met, partially met, and unknown). In a more recent publication (Girod & Girod, 2006), the lesson observation scale used to assess teachers' instruction of a lesson that falls within the TWSM unit seems to have been revised to be scored on a five-point scale (1 = not observed, 2 = emergent, 3 = developing,

4 = competent, 5 = proficient). However, there are no descriptors for these levels of performance that would indicate specific indicators of performance.

Technical Quality. In a journal publication documenting the technical quality of the TWSM, McConney, Schalock, and Schalock (1998) reported that the observation instrument is aligned with the proficiencies required by the Oregon Teacher Standards and Practices Commission. In addition, they reported that the agreement between cooperating teacher and supervisor ratings was between 81 and 98 percent across the evaluative criteria. These agreement figures are based on a three-point scale: met, partially met, and unknown. No interrater agreement data have been reported for the revised five-point scale.

Impact on Candidate Learning. There are no published research reports about the formative opportunities of the observation-based portion of the TWSM in isolation from the TWSM assessment experience as a whole. However, given that the observations are conducted by the cooperating teacher and supervisor, who have an ongoing opportunity to provide feedback to candidates, the instrument may support the mentoring process.

Praxis III—Educational Testing Service. One observation-based assessment that has been developed for teacher licensure but is not implemented by teacher education programs is the Educational Testing Service's (ETS) Praxis III. The evaluation of teaching in the Praxis III is based on a classroom observation of one lesson, documentation of lesson plans and teaching materials, and a semistructured interview. The summative decision is made after a teacher has completed two or more assessment cycles, completed by two or more assessors. Teacher performance is assessed on nineteen criteria across four domains:

- Organizing content knowledge for student learning
- Creating an environment for student learning
- Teaching for student learning
- Teacher professionalism

Teachers are assessed on analytical rubrics that are scored on a scale of 1 to 3.5, with descriptors or indicators of performance along three points (1, 2, 3). (See Dwyer, 1998, for an example of one of the score scales.)

Praxis III is now used in at least two states (Ohio, Arkansas) to convert a preliminary license to a standard five-year license. In the 2007 fiscal year, Ohio spent about $4.2 million for Praxis III administration, including hiring and training regional coordinators who train assessors (about thirteen hundred statewide) and assigning them to administer the assessment for each entry-year

teacher (between five thousand and six thousand each year in the state). Each assessor is paid four hundred dollars per assessment (Ohio Legislative Service Commission, 2007).

Technical Quality. A lengthy development process, including job analyses, research studies, and field trials, was used to develop the assessment (documented in Dwyer, 1994, and briefly described in Dwyer, 1998), to ensure its construct validity. Development began in 1987, and the assessment underwent many revisions before its first operational use in 1993. The careful and iterative process used to develop the Praxis series exams was necessitated by the use for high-stakes licensing decisions, which require that assessments meet the highest standards of technical and legal defensibility. Assessors undergo a five-day training and must pass an assessor proficiency test to ensure their ability to score accurately and produce a satisfactory record of evidence.

Results from the 1992 pilot test indicate that across the nineteen scoring criteria, pairs of assessor ratings were within one-half point of each other for an average 85 percent of candidates assessed (Dwyer, 1998). While this level of interrater agreement is satisfactory for making high-stakes decisions, the fact that it depends on five days of training and a test of calibration suggests that comparable levels of agreement may not be possible among supervisors and cooperating teachers in preservice teacher education without a comparable investment in time and resources to train these assessors.

Impact on Teacher Learning. Danielson and Dwyer (1995) suggest that the analytical rubrics used to score beginning teachers' performance based on observation could provide feedback to those being assessed and subsequently support teachers' professional learning, guide the design of individual professional plans, or, when aggregated for all teachers in a school or district, be used to inform the design and provision of professional development. However, there is little published evidence to date that supports formative learning outcomes for teachers being assessed on the Praxis III. In addition, in order to inform the design of professional development plans, individual teachers must voluntarily release their Praxis scores to districts, employers, or mentors (because licensing test scores are considered private information and only pass/fail outcomes are reported). There is little published evidence that the Praxis III scores have been used in this way for formative purposes.

Observation-Based Research Tools

A number of other observation-based instruments have been developed for research purposes and can be used with high levels of reliability.

Classroom Assessment Scoring System. The Classroom Assessment Scoring System (CLASS) is an observational instrument developed by Pianta and colleagues at the University of Virginia to assess classroom quality in preschool through third-grade classrooms (Pianta, 2003). The focus of this observation instrument is on the interactions between and among teachers and students, and CLASS's three evaluative dimensions (Emotional Support, Classroom Organization, and Instructional Support) are based on developmental theory and research. The three dimensions are derived from constructs that have been validated in child care and elementary school research studies, literature on effective teaching practices, focus groups, and extensive piloting (CLASS, n.d.).

Technical Quality. The CLASS has been used and validated in more than three thousand classrooms from preschool to fifth grade (CLASS, n.d.). The overall level of reliability (agreement within one point on a seven-point scale) was 87.1 percent. Training for using the observation instrument takes two days, training to use the instrument for professional development takes two or three days, and training for the certification of trainers takes five days. CLASS ratings of classroom quality have been linked to the development of children's academic performance and language and social skills at the end of preschool and gains in their performance during the preschool years (Howes et al., 2008; Mashburn et al., 2008). This evidence of predictive validity is remarkable because it has been historically difficult to conduct the kind of research that could provide this kind of validity evidence about a performance assessment instrument.

Impact on Teacher Learning. While the CLASS instrument has been used in research to assess the effects of teacher education by following graduates of a teacher education program into their first years of employment (see Teachers for a New Era grant proposal by La Paro & Pianta, 2003), it is unclear that the instrument has ever been used in the context of preservice teacher education to evaluate the teaching practice of student teachers or support their professional learning. However, the CLASS instrument has been used as a protocol for an Internet-based professional development program for in-service pre-K teachers called MyTeachingPartner (MTP).

The MTP program has two components: on-demand access to video examples of high-quality teacher-child interactions that are sampled to represent specific dimensions of the CLASS observation instrument, and a Web-based consultation service in which teachers make and upload to the Internet video recordings of their own teaching, which are analyzed for the purpose of providing targeted feedback to teachers along the CLASS dimensions of teachers' emotional, organizational, and instructional interactions with students. The content of the Web-mediated consultation includes identifying positive and negative examples of teachers' interactions with students, problem solving

to identify and implement alternative approaches, and a supportive relationship with a more expert consultant. In a controlled evaluation study, Pianta, Mashburn, Downer, Hamre, and Justice (2008) used the CLASS instrument to assess whether there were any differences in the classroom interactions and behaviors of teachers who had on-demand access to video exemplars as well as opportunities for online consultation, versus that of teachers who had access to the video exemplars only. They found that teachers who received the online consultation had significantly greater improvements in their classroom interactions with students than did teachers with video access only, especially in classrooms with higher proportions of students in poverty. While effect sizes were small and the ability to draw conclusions about causal relationships is limited, the findings support the idea that online coaching, focused on specific qualities of teaching through the lens of the CLASS instrument, can provide supportive learning experiences that lead to improved instruction.

Instructional Quality Assessment. Another classroom performance ratings instrument that was developed primarily for research purposes is the Instructional Quality Assessment (IQA), developed and piloted by the National Center for Research on Evaluation, Standards, and Student Testing (CRESST). (See Junker et al., 2006, for an overview of the IQA.) The IQA relies not only on observations of teaching practice but on structured interviews with teachers and students and a collection of teacher assignments. It assesses this body of evidence along four dimensions and twenty separate criteria: Academic Rigor, Clear Expectations, Self-Management of Learning, and Accountability. The design of the IQA was shaped by the Principles of Learning, a set of guidelines for instructional practice (Resnick & Hall, 2001; Institute for Learning, 2002) that integrates strong pedagogical knowledge with deep and rigorous subject matter knowledge (Junker et al., 2006).

Technical Quality. In addition to reflecting research-based principles of learning and effective teaching, the instrument has been piloted in several different contexts to assess the ability of external assessors to score reliably with a standard training protocol (encompassing two and a half days). Reliability estimates from a 2003 pilot test indicate an overall 96 percent level of agreement (exact matches or within one point) in ratings (the Spearman *R* statistic was less robust, at .58). While the level of interrater agreement on individual scoring criteria calls for some improvement in the design of the rubric and training protocol, the researchers conclude that the overall level of agreement is satisfactory, depending on the purpose of the evaluation.

In more recent research on middle school teachers and the quality of their instruction in reading comprehension and mathematics, the quality of teaching measured by the IQA was found to be a significant predictor of student

achievement on the Stanford Achievement Test (Matsumura et al., 2006; Matsumura, Garnier, Slater, & Boston, 2008). Matsumara and colleagues found that it took at least two observations and four samples of assignments to yield stable measures of teaching quality. In addition, the observation tool had high levels of reliability (86.4 percent exact agreement for English language arts and 81.8 percent in mathematics), while the assignment tool had moderately high levels of reliability (71.3 percent in English language arts and 76.3 percent in mathematics). This validity and reliability evidence builds a strong case for the usefulness of the IQA as a summative measure of teaching performance.

Impact on Candidate Learning. There is little published research to date that describes how the IQA has been used as a professional development tool. However, one of the strengths of the IQA is that each score level (1–4) of the rubrics has an associated descriptor intended to precisely and explicitly capture the expectations of performance at each level. (Copies of the rubrics can be found in Junker et al., 2006.) Not only does the design of the rubrics reduce the need for extensive rater training and a reliance on the professional knowledge and judgment of raters, but it also makes possible the opportunity to share results of the assessment with teachers in ways that could support their ongoing learning and development. In addition, by focusing on the technical core of teaching, that is, the interaction of teacher, student, and content, the IQA provides ''a vision of the elements of teaching that matter most,'' based on a ''coherent, research-based model of how high-quality instruction unfolds in classrooms'' (Stein & Matsumura, 2008, p. 197).

Discussion of Observation-Based Assessments

High-stakes decisions about student teachers are routinely based on the observations made by their supervisors and cooperating teachers. Yet the technical quality of the instruments used in preservice programs rarely meets the demands of a high-stakes instrument. In some ways, it appears that the usefulness of these instruments is tied to their ability to support collaborative examination of instruction, student teacher reflection, and formative discussions with supervisors and cooperating teachers that are grounded in a common set of teaching expectations or standards. The way that instruments are used in teacher education raises questions about how the observation data are used in combination with other information to make judgments about a prospective teacher's competence in the classroom. Given that the pass rates in student teaching are extraordinarily high (often approaching 100 percent), it appears that the assessment of student teaching using ratings instruments is treated more as a rite of passage than an objective evaluation of teaching.

If programs are committed to using observation-based evidence as the basis for passing or failing student teaching, then attention to the technical quality

of these instruments and the processes used to train and calibrate assessors is sorely needed. Observation-based ratings instruments that have been developed carefully to ensure content and construct validity and to reflect the research base on effective teaching (for example, Praxis III, CLASS, and ITQ) show that it is possible to develop credible and reliable evaluation instruments and processes, with a heavy investment in the development of a valid instrument and a modest investment in the training of assessors—about two days of training (for CLASS and ITQ). One question about reliability that arises in practice is whether it is possible for evaluators to score reliably when they know the student teacher personally (those who directly supervise and mentor them and are invested in their success), even if they have undergone extensive training and calibration. This question arises for any curriculum-embedded performance assessment that is scored internally by course instructors, supervisors, and cooperating teachers who know the candidates. We have found in our own research on the Performance Assessment for California Teachers that raters who know the candidate in some capacity score marginally but significantly higher (0.2 point higher on a four-point scale) than raters who do not know the candidate. This is a practical measurement problem that needs further exploration.

Although there is limited evidence that the technically robust instruments we describe can also support teacher learning and development, they have the potential to be harnessed by programs to support and mentor student teachers when used by university supervisors and cooperating teachers. The observation instruments that are more reliable and credible (Praxis III, CLASS, and ITQ) are also designed to provide more diagnostic and specific indicators of performance at each score level. This design feature is an important one to consider as a way to bolster the quality of an observation instrument, both to improve scoring reliability and improve its use as a formative tool for teacher and program learning. Developing high-quality scales of this type, which also have predictive validity, is extremely difficult, and most programs are unlikely to have the resources or capacity to engage in the necessary development and research work. Drawing on or adapting existing scales and training systems looks to be a smart investment. State governments, national organizations, and consortia of universities should also be involved in procuring the resources and harnessing external expertise to get smarter about doing this kind of development work.

ON-DEMAND PERFORMANCE TASKS

On-demand performance tasks are similar to traditional sit-down exams in that they present to the teacher candidates a standardized set of prompts, administered in a standardized way. What makes on-demand performance

tasks different from traditional exams is that they present candidates with problem-based scenarios or simulations that mimic authentic teaching situations and do not necessarily have one correct answer (Arends, 2006b). For example, teachers may be given a topic and a description of students in a class, and then they are prompted to describe their plans for a lesson and explain how their plans reflect the strengths and needs of students in the class. A wide range of responses could satisfactorily meet the evaluative criteria for the task. Other tasks may prompt teachers to evaluate a student work sample and describe what kind of feedback they would provide to the student or describe how they would handle a challenging classroom situation.

Well-designed on-demand performance tasks prompt candidates to explain their thinking and decision making so as to make their content knowledge, pedagogical knowledge, and pedagogical content knowledge transparent to the assessor (Arends, 2006b). The teacher licensing tests used in some states include on-demand performance tasks designed to measure teachers' pedagogical knowledge and skills. The content pedagogy Praxis exam administered by ETS, for example, is a one-hour on-demand performance task that can include case studies or other types of open-response essays. Although the Praxis series exams are not embedded in teacher education programs, they present an example of an on-demand performance task that has sufficient technical quality to inform high-stakes decisions: adequate levels of interrater agreement and validation of the assessments based on job analyses. Nevertheless, these types of tasks are clearly less authentic than performance tasks that involve real classrooms and students whom teachers know well that require teachers to implement their plans and evaluate work samples completed by their students.

Another well-known example of an on-demand performance task is the NBPTS Assessment Center tasks, which assess candidates' content knowledge for teaching. Another example is the INTASC on-demand assessment Test of Teacher Knowledge (TTK), which measures ''beginning teacher's professional knowledge in areas such as theories of teaching and learning, cognitive, social and physical development, diagnostic and evaluative assessments, language acquisition, the role of student background in the learning process, and other foundational knowledge and skills essential to the profession of teaching'' (INTASC, 2009a). INTASC has created two secure forms of the TTK, which is still under development, and field-testing has provided validity and reliability analyses. INTASC plans to continue developing items for the TTK and administer the test for states interested in using it.

Central Connecticut State University's Mid-Point Assessment Task

One example of an on-demand performance task that is implemented within a teacher education program is Central Connecticut State University's Mid-Point

Assessment Task (MAT), a two-hour essay exam administered in the semester prior to student teaching. While the assessment was intended originally as a gatekeeping device for admission to student teaching, it now serves primarily as a tool for advising students and for program evaluation (N. Hoffman, personal communication, April 29, 2009). The MAT requires that teacher candidates plan a lesson for a hypothetical group of students, describe how they would differentiate instruction for two focus students, and design assessments to measure student learning of the learning objectives. The information given to candidates about the teaching situation includes contextual information about the school and the unit of study, a lesson plan taught the previous day by a hypothetical teacher, information about class achievement of objectives in the previous day's lesson, and specific profiles of two focus students (one who struggles academically and the other a high-achieving student), along with samples of the focus students' work from the previous day's lesson. Candidates complete their reflective essays by responding to a series of questions presented on computer in a group setting (Arends, 2006b).

Technical Quality. Candidates' responses are scored using a rubric (with eight dimensions) aligned with the program's conceptual framework and modified INTASC standards. The tasks also mirror, to some extent, the tasks in the Connecticut BEST portfolio assessment (see more on the BEST portfolio below). All candidate responses are double-scored by faculty members and triple-scored if there is a conflict in the scores. Interrater reliability levels of 0.85 have been reported (Arends, 2006b). The program stopped using the tool for high-stakes purposes for two reasons: not all program faculty were able to score reliably, even with training, and a sufficient number of candidates who did not meet the passing standard but performed well in student teaching (false-negatives) caused concern about the validity and fairness of the assessment. In addition, some students commented on the need for access to resources outside the testing situation and a desire for unlimited time to plan for the next lesson (N. Hoffman, personal communication, April 30, 2009). This suggests there are some validity issues to consider when using on-demand tasks with formats that are less authentic to instructional practice.

Impact on Candidate Learning. When the MAT was piloted as a high-stakes assessment for entry into student teaching, candidates who received a failing score were offered special workshops to help prepare them for a second try. Candidates who failed the second time had to wait another semester before taking the assessment again (Arends, 2006b). Since the program has chosen to focus on using the tool primarily for advising students and as a program evaluation tool, the results of the assessment are now used to guide students in developing a diagnostic self-improvement plan in collaboration with a faculty

member. In addition, MAT scores for all students and the improvement plans for students who failed the task are sent to the student teaching supervisors to ensure that areas of weakness identified by the MAT are immediately addressed and remediated. Finally, the program has found that engaging all faculty in the process of scoring the MAT has supported a culture of using evidence of candidate performance for program revision and articulation. Score data from the MAT are also used to guide program improvement (N. Hoffman, personal communication, April 30, 2009).

Western Oregon University's Cook School District Simulation

On-demand performance tasks have the potential to become more innovative and authentic when computer technology is harnessed. The Cook School District Simulation was developed by Western Oregon University to simulate the experience of the Teacher Work Sample Methodology (TWSM; see details on this methodology in the section on structured portfolios below). The simulation tool has been piloted extensively with undergraduates, graduate students, and practicing teachers and is in use at three universities in their teacher education programs.

The simulation was designed to be used as a Web-based practice space for teacher candidates to practice the skills needed to make instructional decisions based on an analysis of student learning outcomes. A bank of two hundred students, whose characteristics and simulated academic behaviors were constructed based on teachers' descriptions of real students they had taught, are used to populate the class selected for the particular grade level and content area selected by the user. The user is directed to analyze the public profile of each student in the class to design instruction for learning objectives selected by the user. The user selects from a range of instructional strategies that (in the private profile of each student) are associated with varying levels of student engagement and student learning. In addition, users select from a range of assessment types that are also associated with varying levels of success for individual students. An algorithm uses both the public and private profiles of students, along with information about the selected learning objectives, instructional strategies, and assessment types, to produce the student outcomes (level of engagement and performance on the assessments). Feedback from formative assessments and the simulated student on-task behaviors may be used by users to make adjustments to instruction (Girod & Girod, 2006).

Girod (2008) suggests that the simulation can also be used to measure candidates' skills before and after the formative test in a number of areas:

- Analyzing the school and classroom context of teaching
- Identifying learning objectives appropriate to the school and classroom context

- Designing instruction toward meeting those selected learning objectives
- Adapting and differentiating instruction depending on the context
- Designing assessments to measure student learning
- Evaluating the effectiveness of instruction by analyzing student work
- Planning instruction based on the analysis of student work.

Impacts on Candidate Learning. Girod and Girod (2006) report that teacher candidates who practiced the skills required for the TWSM on the Cook School District simulation for six hours, in comparison with a control group, demonstrated significantly higher levels of skill in creating a classroom climate conducive to learning, adapting instruction to align with student needs; and using a broad array of instructional strategies in their student teaching experiences. (These comparisons were based on TWSM portfolio scores, self-ratings, and lesson observations.) In addition, interviews with users highlight other important impacts on candidate learning: clearer and deeper understanding of the importance of alignment; understanding the importance and challenges of individualizing instruction; understanding the role of assessment in supporting, scaffolding, and judging learning and progress; and understanding the importance of data-driven decision making, analysis, and systematic reflection (Girod, 2008). It seems likely that these impacts on candidate learning are associated with the use of the simulation as a professional learning tool (a practice space) rather than its use as an assessment. This characteristic distinguishes the Cook School District Simulation from other forms of on-demand performance tasks that are designed to serve an assessment function only.

Technical Quality. The technical quality of the simulation as an assessment instrument currently rests on the reliability and validity of the TWSM. Given the limited information about the technical quality of the Cook School District Simulation to date, it seems premature to use the simulation as a high-stakes assessment. Although the simulation leads candidates through tasks that are similar to that of the TWSM, there is a difference between authentic contexts and real contact with real students. The validity of the instrument changes when the evidence base on which candidates' performance is evaluated changes. The student outcomes produced by the simulation algorithms may be true to the students' profiles, but different teachers are differentially effective in producing student learning even when using the same instructional strategies and the same assessments to teach the same learning objectives. Thus, the ability of the simulation to assess the true instructional effectiveness of candidates is called into question.

Discussion of On-Demand Performance Tasks

Of the four types of performance-based teacher evaluation assessments reviewed in this chapter, on-demand tasks seem to be the least intrusive into the curriculum of teacher education, provided that they are designed to be aligned with existing curriculum and do not require extensive time beyond what is normally taught to prepare candidates for the assessment. On-demand formats are also the easiest to implement as they are completed within a given amount of time under standardized conditions. Depending on the nature of the scoring rubrics and procedures, scoring these assessments could be either more efficient with respect to training and scoring time or as demanding and time intensive as the scoring process used for the other three types of performance-based assessment. Thus, it is difficult to assess whether on-demand assessments are more or less labor intensive to implement.

Although the technical quality of some on-demand tasks is robust (for example, Praxis, NBPTS Assessment Center tests, INTASC's TTK), it is unclear how useful they are in terms of their ability to provide formative opportunities that support teacher learning. When on-demand tasks are used exclusively for high-stakes summative purposes such as licensing tests and there are no opportunities for candidates to receive a detailed evidentiary record of why they were successful or why the performance fell short, there does not appear to be much opportunity to learn from the assessment. While it is possible to build an on-demand task into the context of a teacher education program in such a way that there are opportunities for feedback, revision, and learning, as the Cook School District simulation is used to provide practice with the TWSM and Central Connecticut State University's MAT is used for candidate advising and program evaluation, the less authentic nature of on-demand tasks may limit their formative function. For teachers to truly learn about how to improve the execution of their plans, for example, they must have opportunities to implement their plans and learn from what worked and what did not. This is not an experience that a simulation or a reflective essay on a teaching scenario can easily duplicate.

CHILD CASE STUDIES

Child case studies have a long tradition in teacher education and are most often found in elementary-level credential programs, although some middle-level and secondary programs also assign adolescent case studies. In their survey of American Association for Colleges of Teacher Education (AACTE) programs, Salzman, Denner, and Harris (2002) found that 46 percent of respondents (representing 50 percent of teacher education program units) reported the use of case studies as one measure of candidate outcomes. Case studies are

narrative reports that are usually focused on building a child's developmental profile, including physical, social, emotional, and academic and cognitive development, through interviews and observations of the child in a variety of contexts. The ultimate purpose of building this profile is usually for the candidate to draw some implications about the most appropriate ways to work with the child or to design instruction or an intervention focused on meeting the child's educational needs.

Bank Street College Child Study

A well-known example of a teacher education program that requires all candidates to complete child case studies is the Bank Street College of Education, which prepares teachers for early childhood, elementary, and middle school levels. The child case study is an assignment embedded within a single course, The Study of Children in Diverse and Inclusive Settings Through Observation and Recording, and is usually completed during one of the two semesters in which they are enrolled in student teaching. This course has been offered since the college's inception in 1931. During the first half of the semester, candidates observe and take field notes on their observations (using running records, a form of continuous narrative that documents everything the child does as it is done) of a focus student in a variety of situations at school. During this period, candidates are required to submit weekly observation assignments and receive feedback from the instructor. During the second half of the semester, candidates use two methods for analyzing their field notes: an age-level study in which they analyze the child's development across a number of dimensions (social/emotional, physical, cognitive) and a study of patterns in the child's behavior and exceptions to those patterns. Results from these analyses are synthesized into a narrative report in which candidates make inferences about the reasons for the child's patterns in behavior, make links to theories of development learned in the program, consider the educational needs of the child, and pose strategies for working with the child effectively in an educational setting (Gropper, 2008).

Technical Quality. While it appears that a standard set of prompts is used for the case study assignment (with some tweaking of the assignment permitted across faculty teaching the course), there is little evidence that there are standard ways of evaluating or scoring the case studies, or that there is any effort to calibrate the course instructors who score these assignments. In addition, it is unclear whether there is a basis for establishing the validity of the assessment as a measure of teaching competency, although the process of observing and recording children's behavior is closely aligned with assessment practices used by the National Association for the Education of Young Children (Gropper, 2008).

Impact on Candidate Learning. In course evaluations, teacher candidates and graduates of the program have cited the course in which they complete this case study as being one of the most valuable courses they have taken at Bank Street, suggesting that this assignment provides a powerful formative assessment experience (Gropper, 2008):

> They learn to notice children's behavior in detail that eluded them before taking the course: they learn to think about the meaning of the behavior without judging it and to generate strategies for working more effectively with the child in the academic/cognitive, social/emotional, and physical realms. They learn to use observation notes for a range of assessment purposes which include curriculum planning, parent conferences, and referrals for evaluation [Gropper, 2008, p. 194].

Wheelock College's Focus Child Assessment Project

A variant of Bank Street College's child study used by Wheelock College emphasizes the collection and analysis of assessment data in addition to formal and informal observations and running records, with the purpose of designing plans for improving instruction. This case study is assigned in the undergraduate elementary program in Meeting the Diverse Learning Needs of Elementary Students, a course required of all undergraduate student teachers at the college. The assessment is completed over one semester and is composed of the following:

- Informal and formal observations with documentation
- A series of formal neurodevelopmental assessments
- A review and evaluation of the focus child's academic performance (class work, state test results)
- Use of assessment results to inform the development or adaptation of instructional plans for the focus child
- Student work samples, reflection on findings, and analysis and revision of instructional materials to support the child's learning
- A final report that documents all components of the project

Similar to the Bank Street College case study, there are weekly assignments and opportunities to receive feedback on the components of the final report (McKibbens, Shainker, & Banks-Santilli, 2008).

Technical Quality. The assessment criteria and ten scoring rubrics used to evaluate the case study are aligned with the Association for Childhood Education International standards, Wheelock College Education Standards, student learning content standards (Massachusetts), and the Massachusetts Professional Standards for Teachers. The construct validity of the child assessment

project is also reviewed by Wheelock faculty members with expertise in assessment and instruction. Interrater agreement is evaluated once a year by double-scoring a 20 to 25 percent subsample of the completed projects (McKibbens et al., 2008). However, there is no published record of the levels of interrater agreement achieved by the two faculty who have been primarily responsible for developing, administering, and scoring the assignment.

Impact on Candidate Learning. The reaction of candidates to this assignment is that the project is quite challenging. One of the challenges reported with the use of this extensive semester-long project is the ability of teacher candidates to keep up with the pace of weekly assignments so that they receive immediate faculty feedback. Nonetheless, many candidates report that it is a ''pivotal assignment that changes the way they view themselves as teachers and are viewed by others. The point at which the candidates analyze student work and revise their teaching strategies based on those findings is typically noted by their P-12 students and cooperating practitioners as a significant point of progress in their ability to demonstrate competency and independence in the classroom'' (McKibbens et al., 2008, p. 207).

George Mason University's Reading, Writing, Spelling Analysis Task

Another variant of the child study is the literacy or mathematics learning assessment case studies that are used in a number of elementary credential programs. With a growing emphasis on testing literacy and math skills associated with No Child Left Behind accountability requirements, many programs have bolstered their preparation of teachers in the instruction and assessment of literacy and mathematics. George Mason University's Elementary Education Program uses the Reading, Writing, Spelling Analysis Task (RWS) to support its candidates in applying theories of literacy development and pedagogical strategies through an in-depth analysis of one child's literacy learning. The RWS also requires that a candidate design an instructional intervention aimed at supporting the student's literacy development.

Candidates collect information about the case study student by listening to the child reading and documenting the child's reading performance using anecdotal records, running records, miscue analysis, interviews, discussion, reading inventories, developmental reading assessments, or another research-based approach. The candidates also collect three writing samples that represent different writing genres, and using those writing samples, they assess the student's spelling level and word study strategies. The evaluative criteria for scoring the task are based on the diversity of assessments the teacher candidate uses, the accuracy of the developmental levels assigned, and

the appropriateness of instructional plans connected with theories of learning and literacy development learned in the course. Teacher candidates complete the RWS task twice, during each of the literacy courses taken in the elementary credential program. Each of the twelve components of the RWS task is scored on four-point rubrics (Castle, Groth, Moyer, & Burns, 2006). In other programs, candidates are asked to take this task a step further by implementing and evaluating the planned literacy or math intervention.

Technical Quality. The content validity of the RWS task is based on its alignment with INTASC, the International Reading Association, and program standards, as well as course outcomes. Its construct validity depends on a review of one external consultant and surveys of experienced teachers on the instrument's authenticity and fit with real work expectations for practicing teachers. Reliability estimates for rubric scores in the second pilot year indicate 85 to 95 percent agreement in scores, depending on the course section (Castle et al., 2006). Groth (2006) also reports that candidates' scores range from 1.5 to 5.0 (on a five-point scale), with a mean score of 4.5, and that scores on the RWS are consistent with candidates' course grades. Unlike the previous two examples of child case studies, the systematic way in which George Mason's teacher education program scores the student case studies on a common set of scoring criteria and evaluates interrater reliability demonstrates that a case study project could meet standards of psychometric quality for a high-stakes assessment.

Impact on Candidate Learning. George Mason has collected feedback from teacher candidates on their experiences with the RWS task over three years, and although students list the RWS as one of the most difficult assignments of their two literacy courses, 89 percent agree that it is a highly valuable assignment ''because we could witness student growth and also see theories discussed in class, in practice'' (Groth, 2006, p. 9), and that it is ''extremely helpful for planning and assessment and helps apply philosophy and techniques learned in class'' (Castle et al., 2006, p. 77).

Discussion of Child Case Studies

At face value, the activity of systematically observing one student in different settings; collecting detailed information about one student's prior achievement; learning about one student's social, emotional, and physical development; administering psychological or literacy assessments to one student; and writing up a lengthy and highly analytical case study for one student may seem to be a somewhat unrealistic academic exercise because teachers in real-world settings do not have the time to conduct such an intensive assessment (especially secondary school teachers, who can be responsible for as many as 150 to 180 students). However, the high marks that teacher candidates give these kinds of assignments as being one of the most powerful assignments for learning

how to design and plan instruction suggests that the kind of thinking and learning that child case studies can provoke are strongly related to teaching and learning. While it is possible that a child case study task (such as the one designed and used by George Mason University) could be designed and evaluated in such a way as to produce credible and reliable information about teachers' competencies, it appears that its value as a formative learning opportunity overshadows its summative evaluative purposes in many teacher education programs. This is not to imply that teacher education programs should ignore issues of technical quality in their implementation of child case studies, especially if they are used for high-stakes purposes, but that the strength of the assignment may be its powerful formative potential.

PORTFOLIOS

Portfolio assessments are widely used in preservice credential programs, most often as a form of capstone or culminating assessment (St. Maurice & Shaw, 2004). A survey study conducted by Salzman et al. (2002) on behalf of the AACTE found that 88 percent of respondents (representing a 50 percent response rate of 750 AACTE member institutions) reported the use of portfolios as one measure of candidate outcomes. Nearly 40 percent of those programs reported the portfolio as being required for receiving a license to teach. In most cases (95 percent), the portfolios were designed within the programs, but a small percentage (5 percent) reported that the portfolios were designed by the state.

Danielson and McGreal (2000) describe four common features of teacher portfolios currently in use for teacher evaluation (though not specifically in teacher preparation programs):

- Alignment with professional teaching standards as well as individual and school goals
- Selected examples of both student and teacher work
- Captions and commentaries that explain and reflect
- Mentored or coached experiences including conversations with colleagues and supervisors [p. 94]

When evaluating the usefulness of portfolios as a form of teacher evaluation, it is important to note the wide array of portfolio purposes and formats that are currently in use (see Zeichner & Wray, 2001, for a detailed analysis of variations in portfolios used at the preservice level). Tucker, Stronge, Gareis, and Beers (2003) describe the differences in portfolios based on their purpose:

> These subtle differences in purposes tend to fall on a continuum that can be described as low stakes to high stakes. The continuum ranges from an informal, less structured, and improvement-oriented process (e.g., self-assessment, formative performance reviews) to a more formal, structured, and accountability-oriented

> process (e.g., initial hiring decisions, promotion and awards, pay-for performance plans, summative evaluation). . . . Some definitions emphasize processes of teaching, whereas others emphasize products such as evidence of student achievement [pp. 574–575].

While this dichotomy of low-stakes/high-stakes purposes seems valid, we do not subscribe to the notion of a strict dichotomy of low-stakes/formative versus high-stakes/summative or that a portfolio used for summative purposes cannot provide opportunities for teacher learning and improvement. This issue is discussed further below in our description of the Performance Assessment for California Teachers.

In our examination of portfolios, we distinguish structured and unstructured portfolios. Structured portfolios require candidates to submit specific artifacts of teaching with standardized prompts that require direct responses. These artifacts and responses are then scored in a standardized way by trained raters using a common evaluation tool, usually a rubric. The National Board for Professional Teaching Standards portfolio is an example of a highly structured portfolio. In unstructured portfolios, what and how artifacts are selected depend on the purpose of the portfolios. In showcase portfolios used to accompany applications for employment, teachers select artifacts that represent their best work. In portfolios that are meant to be used as a tool for professional learning or for the evaluation of their teaching, candidates may be required to include specific artifacts, such as a statement of teaching philosophy, a videotape of their teaching, lesson plans or units, or original curriculum materials they have developed, with accompanying reflections. In unstructured portfolios, candidates often have more choice in what is selected for inclusion in the portfolio as evidence for evaluation. Sometimes the required elements of the portfolio are meant to provide evidence of meeting state, national, or program teaching standards. However, unstructured portfolios often lack clearly defined evaluative criteria, or the grading of these portfolios may be conducted in less standardized ways (with little training or calibration of faculty scoring the portfolios).

Danielson and McGreal (2000) define several criteria for the summative use of portfolios that are often difficult for local programs and school organizations to meet because of a lack of capacity: ''[A teacher's portfolio] can be used as a summative evaluation tool, but to do so requires a much more structured process and a complex set of assessment strategies. The assessment component requires clear criteria, an established set of reliable and valid scoring rubrics, and extensive training for the evaluators in order to ensure fairness and reliability. These considerations can all be met, but they are often beyond the capacity or the will of a local district'' (pp. 94–95).

Wilkerson and Lang (2003) detail the legal and psychometric issues in using teacher portfolios as a teacher certification assessment, and they warn institutions of the myriad legal challenges they may face if the portfolio assessment

system used for determining a teacher candidate's eligibility for state licensure does not meet accepted guidelines for the technical and psychometric quality of licensing tests (the Standards for Educational and Psychological Testing, APA/NCME/AERA, 1999).

Due to the lack of capacity within most teacher education programs to achieve the technical quality of their portfolio assessments, most teaching portfolios currently in use at the preservice level are of the unstructured variety. However, even unstructured portfolios have been perceived as useful tools for summative teacher evaluation. In one study evaluating the usefulness of a districtwide unstructured portfolio as an instrument for teacher evaluation (Tucker et al., 2003), administrators rated highly the fairness and accuracy of portfolios in assessing teacher performance, and focus group interviews revealed that some administrators felt that portfolios provided a more comprehensive view of teachers' instructional practice than even a series of observations could provide. While both teachers and administrators had less positive ratings of the usefulness and feasibility of the portfolios because of the workload and time involved in constructing the portfolios (a common complaint about portfolios), they also had high ratings of the usefulness of the portfolio for promoting teachers' self-reflection and identifying strengths and weaknesses, and the practicality of the portfolio for aiding them in carrying out their professional responsibilities and their evaluation review conferences. The formative function of portfolios for promoting teacher learning and reflection is explored in more depth below.

Unstructured Portfolios and Formative Assessment

The use of unstructured portfolios as evidence of teacher candidates' learning over the duration of a teacher credential program has gained popularity over the last two decades. (See Chapter One, this volume, for a discussion of their use.) A survey study of teacher education programs conducted over a decade ago (Anderson & DeMeulle, 1998) had found that twenty-four programs had been using portfolios for a range of 6 months to 17 years, with an average of 4.75 years. Ninety-six percent of those programs reported that the purpose of portfolios was to promote student learning and development and 92 percent to encourage student self-assessment and reflection. Studies on the use of portfolio assessment in teacher education programs in Maine (Lyons, 1999) and California (Snyder, Lippincott, & Bower, 1998) have used the reflections of teacher candidates and evidence collected in their portfolios to make inferences about what they have learned through the process of developing their portfolios. Such studies, including an early study by Richert (1987), and a number of other studies on the use of portfolios in preservice teacher education (Anderson & DeMeulle, 1998; Darling-Hammond & Macdonald, 2000; Darling-Hammond & Snyder, 2000; Davis & Honan, 1998; Shulman, 1998; Stone, 1998; Whitford

et al., 2000), have found that portfolios can facilitate teachers' reflections about the content of their lessons and their pedagogical strategies.

The secondary teaching program at the University of Southern Maine (USM) has been using unstructured portfolio assessment since 1994. Teacher candidates complete their professional portfolios during their student teaching intern placements in connection with a student teaching seminar; they present evidence of their having met the teacher certification standards in a final exhibition, which is framed more as an opportunity to ''celebrate and share accomplishment'' than as a high-stakes assessment event (University of Southern Maine, 2008, p. 16). The portfolio assessment requires candidates to select evidence that will serve as the basis for a final judgment about certification, made by a panel of school-based and university faculty following the portfolio exhibition. However, the body of evidence on which the certification decision is made includes informal and formal observations by the university supervisor and mentor teacher, standards review conferences at the middle and end of each placement, videotaped teaching and reflections, lesson and unit plans, evidence lists, and the professional teaching portfolio and exhibition. Thus, completing and presenting the portfolio is less of a high-stakes endeavor because of the inclusion of multiple sources of evidence.

Based on the reflections of teacher candidates and evidence collected in their portfolios, Lyons (1999) found that through the portfolio development process, intern teachers developed habits of mind that helped them define good practice, reflect on their own teaching and learning, and support the reflection of their students. In a three-year longitudinal study of ten preservice teachers at USM's Extended Teacher Education Program, including undergraduates preparing to enter the program, preservice teachers in the postbaccalaureate program taking part in a year-long intensive internship, and graduates of the program in their first or second year of teaching, Lyons (1998a) examined the meaning teacher candidates gave to the portfolio assessment process through case studies involving analysis of their portfolios and open-ended interviews. Cross-sectional and longitudinal analyses of the data yielded several emergent themes:

- Initial efforts at reflection are simple and become elaborated over time.
- Critical conversations interrogating portfolio entries provide a scaffold that fosters teacher consciousness of their own knowledge of practice.
- Significant aspects of teaching practice, such as teaching philosophy, are identified and emerge.
- The sometimes painful process of public, collaborative inquiry results in teachers' learning about themselves and the values they hold for teaching and learning.

Another study, examining the use of portfolio assessment in the fifth-year postbaccalaureate teacher education program at the University of California at Santa Barbara (Snyder et al., 1998), explored the tension between the use of performance assessment for new teacher support and development and as an assessment tool for accountability purposes. At that time, the UCSB program required that all teacher candidates construct a portfolio documenting successful attainment of the California standards for teacher credentialing (the credential portfolio), as well as a second portfolio for candidates working toward a master's of education degree (the M.Ed. portfolio). The collection of artifacts for the credential portfolio was left to the discretion of teacher candidates but had to be selected to show evidence of meeting each of the ten California teaching credential standards. The artifacts could be test or testlike events, such as class papers, unit plans designed for a course, or standardized test scores; observation-based evaluations; or performance tasks or other work samples such as lesson plans, videotapes of teaching, or student work samples. Over the course of the year, candidates had weekly opportunities during their student teaching seminars and in three-way conferences with their university supervisor and cooperating teacher to share the emerging contents of the portfolio and reflect on their work over time. In the last three-way conference in June, the teacher candidate, cooperating teacher, and university supervisor sign off on an official form that the portfolio artifacts as a whole provided evidence that the student had met all of the state's teaching standards.

The M.Ed. portfolio was designed to allow candidates freedom from externally imposed standards in order to encourage reflection on individual practice with a focus on supporting a candidate's ability to ''learn from teaching'' (Snyder et al., 1998, pp. 46–47). This portfolio could be considered a type of inquiry project, built around an issue of the candidate's choosing. Candidates start by collecting three artifacts to identify an issue and examine their instruction in relation to these three artifacts and educational theories learned in the program. Students form self-selected support groups with a university- or school-based facilitator, meet regularly to share and receive feedback on their portfolio artifacts and their reflection on the selected issue, and complete their portfolios in the summer following their full-time student teaching experience. Two checkpoints are required for successful completion: approval by every member of the candidate's support group and a public conversation in which feedback is received from five ''critical friends'' (two school-based educators, two university faculty, and a community member).

Snyder and colleagues followed two cohorts of student teachers (eighteen candidates) over the course of two years through their professional preparation year into their first year of teaching. They found that although the nature of the artifacts selected for the two types of portfolios differed, because the M.Ed. portfolios started with a personal issue and credential portfolios began

with the state standards, both summative and formative portfolios can elicit reflection in student teachers when the assessments provide opportunities for the collection of multiple sources of evidence over time. However, it is unclear whether it is possible for teachers to have the same kinds of learning benefits when one portfolio is used for both formative and summative purposes. It is also unclear whether preservice teachers would have learned as much from their credential portfolios had there been no opportunity to develop a M.Ed. portfolio (Snyder et al., 1998).

Discussion of Unstructured Portfolios

Although these examples of unstructured portfolios appear to serve important formative assessment purposes by promoting teacher reflection and teacher learning, it is clear that such instruments are insufficient for making high-stakes summative decisions about whether preservice teachers should be granted a license to teach. Without clear evaluative criteria and instruments such as scoring rubrics, evidence regarding the reliability and validity of these unstructured portfolio assessments, a system for training and calibrating scorers, and a process for the moderation of scores and eliminating bias, basing high-stakes decisions on such assessments is questionable. In addition, while a great deal of time and effort goes into the construction of these portfolios (on the part of the teacher candidates) and into the evaluation of the work (on the part of the program faculty), the time invested serves a primarily formative purpose. It then becomes a question of whether programs deem the formative purposes of the portfolio worth the effort, time, and resources. The programs that have continued to incorporate them into their programs seem to believe that it is worth the investment.

Structured Portfolios

There are several national and state-level examples of structured portfolios that have been developed for the purpose of assessing the teaching performance of preservice or beginning teachers. Four of the best-known examples are the INTASC Teacher Portfolio; Teacher Work Sampling (Western Oregon University; Renaissance Group); the California Teaching Performance Assessment (developed by the California Commission on Teaching Credentialing in concert with the Educational Testing Service); and the Performance Assessment for California Teachers, developed by a consortium of California preservice credential programs. Although there are likely to be other examples of structured portfolio assessments in current use at the preservice level, these four examples have the most documentation regarding their reliability and validity, systems for the training and calibration of scorers, and moderation processes for ensuring that high-stakes consequences are warranted.

INTASC Teaching Portfolio. A project begun in 1987 (in the same year as the National Board for Professional Teaching Standards) by the Council of Chief State School Officers, the Interstate New Teacher Assessment and Support Consortium (INTASC) was commissioned to encourage collaboration among states to rethink teacher assessment for professional licensure. The ten INTASC principles, or model core standards, for new teachers were based on the National Board's five propositions about effective teaching (Arends, 2006a).

The INTASC Teacher Portfolio was designed to be used for evaluating teachers who have already received a preliminary license, not to evaluate preservice teachers. However, we include a brief description here because it laid the foundation for the preservice portfolio assessment systems described next. In 1995, the INTASC principles were incorporated into the NCATE standards. In 1998, NCATE decided to shift from an input-based model of accreditation to an emphasis on the assessment of teaching performance as evidence of program effectiveness. As a consequence, the INTASC principles and Teacher Portfolio had ripple effects on the standards and assessments used for preservice teachers.

The earliest versions of the INTASC Teacher Portfolio in English language arts, mathematics, and science were created and field-tested by the Performance Assessment Development Project (a joint effort of fifteen states) from 1994 to 2000 and were modeled after the National Board portfolio as well as prototypes of a portfolio assessment that were being developed by the Connecticut Department of Education for the advanced certification of beginning teachers (Arends, 2006a; INTASC, 2009b). The INTASC Teacher Portfolio requires that teachers submit

- Materials used in instruction
- Examples of student work
- Videotapes of teaching and learning in a candidate's classroom
- Written records of activities and assessments in a candidate's classroom
- Written commentaries that explain a candidate's thinking about teaching and learning

The portfolios are scored using a set of validated rubrics by trained scorers, and the reliability of the scores is checked through double-scoring of portfolios. (For a description of the INTASC scorer training and scoring process, see Moss, Schutz, & Collins, 1998.)

As of 2008, several states had incorporated the use of a teacher portfolio similar to the INTASC Teacher Portfolio into their licensure process for beginning teachers (among them are Connecticut, Indiana, California, and North Carolina) or master teachers (Wisconsin is one of them) (Coggshall et al., 2008). However, few states have required a portfolio assessment for

initial certification (administered at the preservice level). California's efforts to implement its new teaching performance assessment requirement through preservice credential programs are described in more detail below.

Technical Quality. The content validity of the INTASC teacher portfolio and its evaluative criteria rests on its coherence with the INTASC general and content-specific standards, research on teaching knowledge and practice, and self-reported practices of teachers (Moss et al., 1998). Moss and colleagues describe and evaluate the scoring method that was under development, a collaborative evaluation by two raters, as a ''dialogic and integrative'' (p. 141) way of reading and interpreting teacher portfolios. They then argue for a more hermeneutic approach to construct a coherent interpretation that is challenged and revised as more and more available evidence is accounted for (in contrast to a psychometric approach that calls for independent readings of isolated performances). While the INTASC scoring approach is holistic, it is based on ''a series of explicit steps involving data reduction and integration guided by questions from a detailed evaluation framework. At each stage, the steps of data reduction and integration are recorded for consideration at the next stage. By the time they are ready to reach consensus on an overall conclusion, readers have produced a written record of steps'' (Moss, 1998, p. 209). While Moss and colleagues have articulated a validity research agenda, no additional information on validity or score reliability has been reported by INTASC to date.

In one recent study of the predictive validity of the Connecticut BEST portfolios in relation to the achievement of students taught by teachers with varying scores on the portfolios (Wilson, Hallam, Pecheone, & Moss, in press), there appeared to be a small but significant relationship between teachers' scores on the BEST portfolio and their students' achievement scores (there were no significant relationships between teachers' scores on Praxis I and II and student achievement scores). More large-scale studies that examine the predictive validity of large-scale portfolio assessments are needed to provide additional evidence of the validity of portfolio assessments.

Impact on Teacher Learning. Little research has been published on the impact of the INTASC Teacher Portfolios on teachers' professional learning. (For state case studies on Indiana, North Carolina, and Connecticut, see Kimball, 2002; White, 2002; Kellor, 2002; and Wilson, Darling-Hammond, & Berry, 2001.) The little research that has been published indicates that the process of constructing these structured portfolios, sometimes to meet high-stakes licensing requirements, also served a formative function by promoting teacher reflection and learning. In surveys of beginning teachers who completed the portfolio requirement in Connecticut as part of the Beginning Educator Support and Training program, 72 percent of respondents reported that the

portfolio construction process had improved their ability to reflect, 60 percent reported that the process had helped them focus on important aspects of teaching, and 50 percent reported that the portfolio had improved their teaching practice (Wilson, et al., 2001). In another unpublished paper (Lomask, Seroussi, & Budzinsky, 1997), teachers who participated in a pilot science portfolio assessment provided written feedback indicating that most teachers found the process of portfolio development and the program's support seminars as an opportunity for reflection and professional growth.

Practicality and Feasibility. Implementation of the Connecticut BEST program has required significant resources ($3.5 million per year) (Kellor, 2002), suggesting that statewide teacher certification by portfolio could be cost-intensive for most states. In fact, recent state budget cuts and legal challenges have forced Connecticut to revise the BEST program to rethink the portfolio requirement. In Indiana, the cost of training scorers was $58,000 per training session and the cost of scoring each portfolio was $120 (Kimball, 2002). In North Carolina, the total state cost of administering the teacher portfolio assessment for the 2000–2001 academic year was $500,000 and the cost of scoring each portfolio was estimated at $168 (White, 2002). The cost-benefit ratio regarding the use of portfolio methodology to support teacher learning and to assess teacher quality needs further study.

Teacher Work Sampling. In their survey study for AACTE, Salzman et al. (2002) found that 66 percent of respondents reported the use of teacher work sampling and 28 percent reported the use of measures of P–12 pupil learning as one measure of candidate outcomes. Teacher work sampling is a specific form of structured portfolio that is similar in format to the INTASC portfolios in that it requires teacher candidates to collect and submit specific artifacts that represent their teaching practice and respond to prompts that are aimed to help teachers elucidate the thinking behind their instructional decisions. Del Schalock and his colleagues at Western Oregon University are usually cited as the originators of the Teacher Work Sample Methodology (TWSM) (see Schalock, Schalock, & Girod, 1997). The TWSM used at Western Oregon University requires teacher candidates to develop a three- to five-week instructional unit that they implement during their student teaching placements. The evidence on which teachers are evaluated includes the following (McConney et al., 1998):

- A description of the teaching and learning outcomes to be accomplished
- A description of the teaching and learning context
- Instructional plans
- Pre- and postassessments developed and used to measure student progress

- Evidence of student learning gains
- Interpretation of and reflection on the success of the teaching-learning unit, including progress made by students in their learning and significance for the teacher's future practice and professional development

Each of these artifacts is assessed by supervising and research faculty on ratings instruments, and the teacher candidate's implementation of the instructional plans is observed and rated by university and school supervisors on an observation instrument. What is unique about this structured portfolio assessment is the requirement that teachers use pre- and postassessments to measure student progress during the course of the unit and that evidence of students' learning gains is used to evaluate teachers' ability to improve student learning. What is unclear from publications on the TWSM is the technical quality and comparability of the student pre- and postassessments that are designed and implemented by preservice teachers. While the quality, variety, and range of teachers' assessments and their alignment with stated learning objectives are evaluative criteria in the TWSM rubrics (see McConney & Ayers, 1998), how this relates to the validity of the teacher-designed pre- and postmeasures as evidence of student learning raises questions about the validity of judgments about a teacher's ability to impact student learning.

This series of assessments becomes part of a teacher profile that is used both formatively to provide feedback to teacher candidates and summatively to inform the credential decision. In the Western Oregon University program, candidates completed two TWSM cycles, the second with greater independence. But both were completed with ongoing feedback from supervising faculty. Evidence from interviews and focus groups with graduates of the program strongly supports the idea that a performance assessment that is a capstone assessment can also be formative by prompting teachers to articulate the rationale of their instructional decisions, learn how to plan for instruction and adapt their instruction based on their preassessments of students, and reflect on their instruction in light of student learning (Schalock, 1998; Girod & Schalock, 2002).

Technical Quality. McConney et al. (1998) argued that the TWS instrument met criteria for authenticity and content and for face validity (based on alignment with program and state standards for teaching competency), but they acknowledged that the instrument's ability to meet psychometric standards for reliability and freedom from bias was mixed. While there were high levels of agreement in the observation ratings of school and college supervisors (between 81 and 98 percent agreement), interrater reliability for the other measures in the TWS could not be reported. The validity of the pre- and postassessment measures of student learning developed by teachers themselves was found

to be strongly related to other measures of teaching quality in the TWS. To date, we know of no updates to information about the interrater reliability of the ratings instruments used for evaluating preservice teachers on the TWSM used by Western Oregon University. However, there is more published research on the technical quality of the teacher work sampling approach adapted, piloted, researched, and validated by the Renaissance Partnership for Improving Teacher Quality.

Renaissance Teacher Work Sample. The Renaissance Teacher Work Sample (RTWS) instrument was developed between 1999 and 2005 by a partnership of eleven members of the Renaissance Group, a consortium of colleges and universities committed to improving teacher education within their institutions. Funded in part by a Title II Teacher Quality Enhancement Grant, Western Kentucky University was the grantee and leading institution in the project. An adaptation of the Western Oregon University TWSM, the RTWS instrument assesses preservice teachers along seven teaching dimensions (Pankratz, 2008):

- Use of the student learning context to design instruction
- Development of clear instructional unit goals aligned with state and national content standards
- Design of an assessment plan to include pre-, post-, and formative assessments that guide and measure student learning
- Use of formative assessment to make sound instructional decisions
- Measurement and analysis of student learning that resulted from a unit of instruction (individual students, subgroups, and total class)
- Reflection on and evaluation of teaching and learning with respect to the unit of instruction

Technical Quality. During the period of the portfolio assessment's development and piloting, the Renaissance Partnership used specific processes to train scorers to score accurately and without bias. Pankratz (2008) reports that the RTWS had dependability coefficients of 0.80 or better with three scorers and 0.60 or better with two scorers based on Shavelson and Webb's (1991) test of generalizabilty.[3] The validity of the RTWS rests on high levels of face validity, based on Crocker's (1997) criteria for validity: criticality of tasks, frequency of task performance, realism, alignment with state and national standards, and representativeness. The RTWS was designed to align with INTASC's Model Standards for Beginning Teacher Licensing, Assessment and Development (Pankratz, 2008). Evidence on scorer dependability and validity was published in Denner, Salzman, and Bangert (2002) on a modified TWS instrument with

modified holistic scoring procedures used at Idaho State University. Denner and colleagues found acceptable dependability coefficients ranging from 0.75 to 0.86 for two raters. This study, replicated with RTWS work samples from across nine of the eleven institutions (Denner, Norman, Salzman, & Pankratz, 2003), found that dependability coefficients of 0.77 to 0.82 could be achieved with three raters. Denner, Norman, and Lin (2009) also investigated the fairness and consequential validity of the TWS instrument using score and demographic data from two of the Renaissance Group institutions by examining whether the instrument had a disparate impact on candidates based on gender, age, or race/ethnicity. They did not find any disparate impact or adverse consequences based on these demographic backgrounds. The authors also investigated the relationship between candidates' TWS scores and their satisfaction of college entry requirements, Praxis I scores, and grade point average (GPA). They found that the GPA for the education core courses, the cumulative GPA, and Praxis I writing scores were significant predictors of the total TWS scores.

Impact on Candidate Learning. When the Renaissance Partnership project ended in 2005, more than six thousand teacher work samples had been produced across the eleven partner institutions, and nine of those institutions have required the completion of the RTWS for graduation. While the initial reactions of most teacher candidates to the RTWS were that it required too much paperwork, that they did not have enough time in their student teaching placements to produce a high-quality TWS, and that they were overwhelmed by the high standards of performance on the RTWS, these negative attitudes tended to dissipate as the candidates completed their units and saw evidence of student learning outcomes in their classes. Some appreciated the realization that they were able to make a difference with their students and that the process helped them feel like professionals (Pankratz, 2008).

One of the questions that often arises in relation to the practicality of portfolio assessment is how long it takes to score this body of work, with the implication that the time required would make this form of assessment less feasible on practical grounds. The RTWS directs candidates to write some twenty or more pages of narrative, plus the artifacts of their teaching of an instructional unit. Denner, Salzman, & Harris (2002) reported that the average time it took to complete their modified TWS scoring protocol at Idaho State University was 13.5 minutes for each TWS. Another study reported an average scoring time of 24 minutes for teacher work samples that were not considered benchmarks (those representative of specific scoring levels) (Denner, Salzman, & Harris, 2002).

Another question, mentioned in the description of the Western Oregon University TWSM, was whether the evaluation of teacher quality based on the TWS pre- and poststudent learning measures (designed, implemented, and

analyzed by teachers) is dependent on the quality of the student assessments themselves. Denner et al. (2003) analyzed the relationship between the quality of the student assessments and the overall evaluation of a TWS. They found significant positive correlations between independent evaluations of assessment quality and the teachers' RTWS scores, suggesting that the RTWS rubrics are able to distinguish the quality of the pre- and postassessments and that teachers are not given credit for showing improvements in student learning when the quality of the assessments is poor. Denner and Lin (2008) found a significant relationship between teachers' RTWS scores and the reported percentage gains in student achievement on the pre- and postassessments of student learning that comprise one required component of the RTWS. They also found greater gains in student achievement for candidates in the second intern-teaching experience than for the same candidates who constructed a RTWS during their pre-intern-teaching experience. The authors conclude that the RTWS provides evidence of the impact of teacher candidates' instruction on student learning and that ''teacher preparation programs make a difference to the teaching abilities of their teacher candidates'' (p. 16). (Additional research on the TWSM can be found at http://edtech.wku.edu/rtwsc.)

California Teaching Performance Assessment. In 1998, the California legislature passed Senate Bill 2042 with the goal of transforming the teacher licensing system in the state and reforming teacher preparation. One of the new requirements for the initial teaching credential introduced by SB 2042 was a teaching performance assessment (TPA) that would be completed during preservice preparation. Programs were given the option to administer the TPA developed by the state (through a contract with the ETS) or to design and administer their own TPA, provided that it meets the state's standards for psychometric quality. The TPA designed by the California Commission on Teacher Credentialing (CCTC), in partnership with ETS, was designed to measure the state teaching standards for beginning teachers.

The Cal TPA, a hybrid performance assessment that includes responses to classroom scenarios and portfolio components (in which teachers plan and teach lessons and collect student work samples), is designed to be administered during teacher education course work throughout the duration of the program.[4] It has four tasks:

- *Subject-specific pedagogy:* Candidates are given four case studies of specific classes and learners (specified for each credential—for example, elementary or secondary) and are prompted to develop teaching methods and lesson plans focused on the content, analyze and adapt assessment plans focused on the content, adapt lessons for English learners, and adapt lessons for students with special needs.

- *Designing instruction:* Candidates plan a lesson for an actual class of K–12 students, including adaptations for English-language learners and students presenting other instructional challenges.
- *Assessing learning:* Candidates plan an assessment based on learning goals, administer the student assessments, adapt the assessments for English learners and for students with other instructional challenges, and analyze and use the assessment results to plan instruction.
- *Culminating teaching experience:* This task integrates the three previous tasks by having candidates learn about their students, plan a lesson and assessment, adapt instruction and assessment for English learners and students with other instructional challenges, teach the lesson and administer the assessments, and analyze the lesson and assessment results to plan further instruction.

In addition, candidates reflect on what was learned in completing the task at the end of each task. As each task in the series is completed, it is scored holistically on a four-point rubric by the program faculty or other trained and qualified assessors. (The CCTC offers a one-day orientation training for scorers and a two-day training for scoring each of the four performance tasks. The commission also offers lead assessor training to provide local turnkey training sessions.) Candidates must earn a combined total of twelve points across the four tasks and must have a minimum score of 2 on any one task. Candidates are permitted to resubmit the tasks as many times as is necessary to earn the minimum number of points (California Commission on Teacher Credentialing, n.d.).

Technical Quality. The Cal TPA was designed to measure the California TPEs for beginning teachers, which was created based on a job analysis for beginning teachers and validated by a committee of educators and stakeholders across the state. Thus, the content validity of the assessment rests on its alignment with these TPEs, as well as with the two regional focus review groups that were used to support the development, pilot testing, and review of the TPA prototype. Once the TPA prototype was finalized, it underwent an ETS sensitivity and fairness review process and was pilot-tested and scored in spring 2002. The purpose of the session (scoring a subsample of candidates' responses) was to collect information about the tasks, reactions to the tasks, and recommendations for modifying the tasks (California Commission on Teacher Credentialing, 2003a).

Following the pilot test, a larger field review of the four tasks was completed from October 2002 to April 2003. Forty-two assessors were convened centrally for training and calibration in June 2003 and scored the tasks. Based on score data from 104 teacher candidate performances with all four tasks scored, assessor agreement was calculated to range from 91 to 98 percent across the four tasks (exact agreement plus differences of one point on the score

scale), and assessor reliability (using the Spearman Brown prophecy reliability statistic) was reported as ranging between 0.63 to 0.83 across the four tasks and 0.87 overall (California Commission on Teacher Credentialing, 2003b). These results indicate acceptable levels of interrater reliability when raters are trained and scoring is conducted centrally (raters are trained and score under direct supervision of the CCTC and ETS). Programs have the option of participating in centralized training of raters or using a trainer of trainers model for local scorer training. Local trainers receive specialized training from the CCTC's lead trainers before they may train local raters. There have been no additional official reports of rater consistency or reliability under this decentralized model of training and scoring.

Costs of Implementing a Statewide Teaching Performance Assessment. While approximately $10 million was appropriated by the state legislature for the development and validation of the Cal TPA, it is unclear what the total annual costs of implementing this performance-based assessment system for preservice teacher credential programs across the state would be. Some estimates have ranged from two hundred to four hundred dollars per teacher candidate (depending on whether the costs associated with implementing the assessment system include costs beyond payments to trainers and assessors for scoring). For large programs in the California State University system, which produce hundreds of teaching credentials per year, the pressure on program budgets is tremendous. There are divergent views about the responsibility of higher education institutions and other credentialing agencies for engaging in the assessment of candidates for beginning licensure. There are strong sentiments in the state legislature that assessment of graduates from credential-granting programs for purposes of quality control is inherent in the role of higher education programs and that the associated costs should be built into program budgets. The implication is that all program faculty, including supervisors, and other nontenured instructors would be required to support teacher candidates completing the TPA as well as participate in scoring the TPA as part of their job responsibilities (with no additional compensation). Given the current state budget crises and cuts to the state's education spending, few resources were appropriated for the purpose of funding the state's TPA mandate (as it went into effect in July 2008). Programs are struggling but making do with the few resources they were allocated, and they continue to lobby the state for additional funds on the grounds that the current mandate is unfunded.

Tensions in the Formative and Summative Purposes of the TPA. The Cal TPA model is designed to be implemented and scored in the context of teacher education course work. Teachers are to be prepared for and complete each of the four tasks as part of their course requirements and receive formative feedback and support on their tasks from course instructors. While on one level

it seems more supportive of candidate learning and success to embed the tasks in the context of their teacher education course work, one of the complaints commonly raised about this model is that it ''colonizes'' the curriculum of teacher education. Many program faculty members across the state have objected to the increasing encroachment by state regulation on their academic freedom and being forced to teach to the test. Thus, while it appears that the integration of the Cal TPA into program course work is likely to improve its usefulness as a formative assessment, this integration is in tension with the assessment's high-stakes function as one gatekeeper to the initial credential. The high-stakes nature of the assessment and its integration into course work force program faculty to teach to the test. If higher education faculty members are required to adopt and integrate a TPA into their courses to this extent, it is imperative that research about the predictive validity of the TPA and its value as both a summative and formative assessment be documented.

Performance Assessment for California Teachers. The Performance Assessment for California Teachers (PACT), another form of a structured portfolio assessment, is currently used in thirty-two preservice credential programs in California and was recently adapted for use in Washington State as part of its initial teaching licensure requirement.[5] After California elected to require teacher preparation programs to use standardized performance assessments in making credentialing decisions (along with other measures), it contracted with the ETS to develop such an instrument, but gave teacher education institutions the option of using a different instrument if it met the CCTC's Assessment Quality Standards.

A coalition of California institutions of higher education formed PACT to develop such an alternative assessment method. The PACT Consortium was initially composed of twelve universities: University of California (UC) Berkeley, UCLA, UC San Diego, UC Santa Cruz, UC Santa Barbara, UC Riverside, UC Davis, UC Irvine, San Jose State University, San Diego State University, Stanford University, and Mills College. The consortium has since grown to include thirty-two preservice credential programs (including one district intern program). Based on the latest available data from the CCTC for 2005–2006, PACT institutions produced 3,877 or 31.6 percent of multiple subject (elementary), 2,544 or 35.1 percent of single subject (secondary), and 6,421 or 32.9 percent overall of the candidates receiving preliminary California teaching credentials. From 2002–2003 up to 2007–2008, the PACT Teaching Event was piloted across the consortium; 2008–2009 is the first year of full enactment of the law (performance on the TPA now counts for teacher licensure).

The development of and research on the PACT was funded by grants from the University of California Office of the President, the Flora and Sally Hewlett Family Foundation, the Hewlett Foundation, and the Morgan Family

Foundation.[6] A key motivation for the PACT Consortium was to develop an integrated set of rigorous, transparent, subject-specific, standards-based certification assessment instruments that would be consistent with the curricular and professional commitments of the member institutions. The goal of the PACT Consortium has been to strengthen the quality of teacher preparation by using curriculum-embedded assessment instruments developed by each member institution in combination with a standardized teaching performance assessment to recommend licensure for prospective teachers.

PACT Assessment Design. The PACT assessment system consists of two interconnected components: a standardized portfolio assessment, the Teaching Event (TE), and locally developed Embedded Signature Assessments (ESAs). The Teaching Event is an evidence-based system that uses multiple sources of data: teacher plans, teacher artifacts, student work samples, video clips of teaching, and personal reflections and commentaries. The TEs are subject-specific assessments integrated across four tasks—planning, instruction, assessment, and reflection—with a focus on the use of academic language embedded across the tasks.[7] To meet the needs of the range of credential programs offered by PACT campuses, there are six versions of the multiple-subject Teaching Event (including two for bilingual emphasis candidates and two for candidates concurrently earning a special education credential) and eighteen single-subject TEs. For each Teaching Event, candidates must plan and teach a learning segment of thee to five hours of instruction (an instructional unit or part of a unit), videotape and analyze their instruction, analyze student learning, and reflect on their practice. The Teaching Events are designed to measure and promote candidates' abilities to integrate their knowledge of content, students, and instructional context in making instructional decisions and reflecting on practice.

Individual PACT credential programs have also developed and administered ESAs, customized assessments to measure additional teaching competencies that are central to their program mission and goals. PACT is still tackling the technical challenge of combining scores from varied locally designed and customized assessments with scores from the Teaching Event, which serves as a standardized anchor assessment. The ultimate goal is to use both sources of evidence to contribute to the final pass/fail decision for the PACT teaching performance assessment. Until the measurement challenges have been resolved, the ESAs will be part of required course work and used formatively to build the teaching capacity of prospective teachers and for program evaluation or accreditation purposes.

PACT Scoring System. The proposed scoring system for the PACT Teaching Event by itself includes both a local and centralized scoring model. In most years, scoring is conducted at each local campus by a group of subject-specific

trainers who are trained centrally each year. These trainers train, calibrate, and monitor scorers and oversee the local scoring process, including implementation of a plan for double-scoring selected TEs. All failing and borderline TEs are double-scored and checked by the lead trainer to confirm the decision. An additional random 10 percent sample stratified across passing score levels is double-scored by local scorers. The consistency of local scoring is managed through a centralized audit of 10 percent of local scores, with intervention aimed at campuses that are identified as producing unreliable scores. Every third year, a central standardized scoring model will be used to provide another check on the consistency of training and the scoring process and the reliability and validity of scores. It takes between two and four hours to score a single TE, depending on the experience of the rater. Scores from the pilot indicate that candidates across all subject areas tended to perform at a higher level on the planning and instruction tasks than on the assessment and reflection tasks. In addition, candidates tended to perform at a lower level of performance on the academic language–related rubrics.

Technical Quality. To meet the assessment quality standards of the California Commission on Teacher Credentialing, which unanimously approved the PACT for use in meeting the requirements of the statute, the PACT Consortium spent considerable resources to collect evidence on and document the validity and reliability of the instrument. To document content validity, teacher educators who participated in the development and design of the assessments, as well as teacher educators not involved in the design of the assessment and who scored the portfolios, were asked to judge the extent to which the content of the TEs was an authentic representation of important dimensions of teaching. Another study examined the alignment of the Teaching Event tasks to the California teaching performance expectations (TPEs). Overall, the findings across all content validity activities suggest a strong linkage of the TPE standards, the Teaching Event tasks, and the skills and abilities needed for safe and competent professional practice (Pecheone & Chung, 2007).

Bias Reviews and Analysis. Bias reviews, following guidelines put forth by the Educational Testing Service (2002) for conducting bias/sensitivity reviews of assessments, were conducted to examine the TE handbooks and rubrics used in each certification area to evaluate the text for offensive or potentially offensive language and to identify any areas of bias due to race, gender, ethnicity, or cultural-linguistic backgrounds. The findings from this process were used to flag areas of potential bias, which informed subsequent revisions of the TE handbooks, rubrics, and scoring process. Second, the performance of candidates on the PACT assessment was examined to determine if candidates performed differentially with respect to specific demographic characteristics.

For the 2003–2004 pilot, there were no significant differences in scores by race/ethnicity of candidates, percentage of English language learner students in candidates' classrooms, grade level taught (elementary versus secondary), academic achievement level of a candidate's students, and months of previous paid teaching experience. There were statistically significant differences between male and female candidates (with females scoring higher) and between candidates teaching in schools in different socioeconomic contexts (with candidates in suburban schools scoring slightly higher than those in urban or inner-city schools). The PACT Consortium plans to continue to monitor and reexamine the scorer training process, the design of the Teaching Event assessments, and differences in candidate scores based on demographic differences to uncover any potential sources of bias due to varying socioeconomic contexts (Pecheone & Chung, 2007).

Finally, score consistency and reliability were examined. Analysis of the consistency between 2003–2004 local campus scores and audit scores (in which a sample of locally scored TEs was rescored at a central scoring session) provided evidence about consistency across pairs of scores by computing consensus estimates within each subject area. Across content areas, 91 percent of score pairs were exact matches or within one point. Interrater reliability was also calculated using the Spearman Brown prophecy reliability statistic. For the 2003–2004 pilot year, the overall interrater reliability for all rubrics across tasks was 0.88 (Pecheone & Chung, 2007).

High Stakes and Formative. The formative potential of the PACT Teaching Event for teacher candidates, individual faculty members, and programs as a whole is in large part related to the analytical nature of the rubrics and the specific information that the rubric scores provide about the strengths and weaknesses of preservice teachers' instructional practice. The design of the rubrics and the way in which they are written allow some transparency in interpreting the score results by providing concrete images of beginning teacher practice at various levels. This is supportive of program faculty who want to provide formative feedback to candidates as they construct their Teaching Events, as well as to programs engaging in an analysis of aggregate scores for the purpose of program review and revision.

Several research studies have been conducted that examine the impact of completing the PACT on the learning experiences of preservice teachers who have completed the PACT Teaching Event. Chung (2005, 2008), one of the authors of this chapter, collected both quantitative and qualitative evidence (surveys, ratings of teaching, interviews, observations) to examine whether preservice teachers who completed the PACT Teaching Event reported learning from their experiences with the portfolio assessment and whether there was any evidence of changes in their teaching practice consistent with their reports of

learning. In a study conducted during the first pilot year (2002–2003) at a large urban university participating in the PACT Consortium, Chung (2008) interviewed and observed the classroom instruction of preservice teachers before, during, and after completion of the Teaching Event. She found that engaging in the process of planning, teaching, and documenting a unit of instruction for the Teaching Event afforded a number of important learning experiences related to the novelty of some of these experiences for the two case study teachers. For example, the teachers cited learning from the opportunity to formally investigate the characteristics and learning needs of their students, independently plan lessons (rather than implement lessons designed by their cooperating teachers), attend to the needs of English learners, analyze their students' learning and use that information to make adjustments to subsequent lessons, and reflect on their teaching effectiveness. In addition, some of the changes that teachers reported in their teaching practices were observed in their subsequent classroom instructional practice.

In another comparison group study of teacher candidates in another large urban university conducted during the second pilot year (2003–2004), Chung (2005) compared the learning experiences of those who had completed the Teaching Event and those who had not. She found that preservice teachers in cohorts that had completed the Teaching Event began with significantly lower average self-ratings and lower supervisor ratings of their teaching knowledge and skills than teachers in cohorts that had not completed the Teaching Event, but by the end of the program, they had closed the gap in their self-ratings and supervisor ratings. Through case studies of teachers in the piloting and nonpiloting cohorts, Chung analyzed more closely the learning gains of teachers in both groups and was able to disentangle to some extent the learning gained from experiences with the Teaching Event and program experiences overall. Teachers in the piloting cohorts were more likely to report improving their ability to reflect on their teaching decisions and assess student learning. However, candidates' reports of learning were moderated by the quality of implementation at the university during the second pilot year, which led many of the piloting teachers to feel that the requirements of the Teaching Event were too burdensome. These constraints detracted from their abilities to learn from their planning and teaching experiences.

In the first and second pilot years of the PACT, the Teaching Event at most universities in the PACT Consortium was not completed as a high-stakes assessment with scores counting toward the credential decision. However, in most cases, piloting teachers were required to complete the assessment as part of a course or program requirement and contributed in some way to the credential decision. During the first two pilot years, the PACT Consortium administered a survey to all candidates completing Teaching Events across the consortium. They found that approximately 90 percent felt that the Teaching Event validly measured important elements of their teaching knowledge

and skill, and two-thirds felt that they had learned important skills through their experiences with the Teaching Event. In particular, survey respondents reported that they had learned the importance of using student work analysis to guide their instructional decisions and reflect more carefully about their teaching. In addition, preservice teachers have reported that the experience of investigating their students' backgrounds for the instructional context task of the Teaching Event has prompted them to pay greater attention to their students' specific learning needs in designing instruction, as this comment from a California State University teacher candidate illustrates:

> So, you know, at the beginning, the PACT lesson has you analyze: What's the context? Who are the kids? What needs do they have? Do you have English language learners? . . . What kind of English language learner are they and how much, where are they on the spectrum? Are they beginning language learners? Are they advanced language learners? And then, to take that information about all the kids in your class, and then think about teaching to every single one of them. That was kind of a new experience for me. It was actually the first time in my teaching experience that everything came together from beginning to end, and made sense. It made sense.

In more recent years, many of the piloting programs have required a passing score on the Teaching Event for successful completion of the credential program. Nonetheless, the reports of preservice teachers have been no less positive regarding the learning gained from their experiences with the Teaching Event. One California State University teacher candidate who had gone back to graduate school to earn a public school credential after having taught for ten years in a private school setting recently commented on her experience with PACT:

> It made a huge change in the way that I assessed. And that was kind of a surprise for me when I got to the end of the PACT. . . . When I went to assess, I kind of floundered for a second. I'm kind of looking at the kid's work in front of me, and I'm kind of spinning my wheels for a second. What am I assessing? What am I assessing? And it kind of took me an hour or two to figure out, wait a second, I need to go back to my big idea to be up front. And that's the first time I—and I've been teaching for ten years—I ever thought of assessment in that sense. And when I pulled out the big idea, it allowed me to assess not only the students, but to assess my own teaching, and that was kind of a new experience for me too. I never used objectives or the big idea to assess my own teaching to particular students. And it just kind of opened up a whole new world that made more sense to me as to what my role was in the classroom, what my role is as a teacher to these students, and a tool that allowed me to analyze my own teaching over three lessons to see if I was really effective in teaching a big idea. It allowed a focus—before there was so much data coming at me, there were multiple objectives, it was too much—and this way, it kind of gave me a focus to say, this big idea, did I teach this well? Did the students know this, and if not, why not?

Using survey data from candidates completing the PACT, the PACT Consortium has found that the more teachers reported being supported in their completion of the assessment, the better prepared they felt by their course work and field placements, and the more likely they were to report having learned from the Teaching Event. Higher ratings of program supports and preparation and higher levels of reported learning have been associated with higher scores on the PACT Teaching Event (Chung, 2007; Chung & Whittaker, 2007). (See Tables 3.1 to 3.4.)

These findings suggest that even in a high-stakes environment, preservice teachers can have positive learning experiences associated with a summative evaluation of their teaching when their programs are able to provide the supports for candidate learning and their prior program experiences are supportive of success on the Teaching Event.

Bunch, Aguirre, and Tellez (2008) emphasize the formative value of preservice assessments that explicitly prompt preservice teachers to attend to the academic language development of their students. Their analysis of the Teaching Events of eight teacher candidates, focused on teachers' understanding of

Table 3.1. Association Between Candidate Ratings of Support for Completing the TE and Their Perceptions of Learning from the TE, 2003–2004

Total Support Score[a]	*Number*	*Mean Agreement Level on Learning*	*Standard Deviation*	*Standard Error*	*95 Percent Confidence Interval*	
					Lower Bound	*Upper Bound*
Group 1: 1–6	25	2.04	.889	.178	1.673	2.407
Group 2: 6–12	156	2.37	.924	.074	2.226	2.518
Group 3: 13–18	220	2.58	.843	.057	2.465	2.689
Group 4: 19–24	136	2.84	.762	.065	2.709	2.968
Group 5: 25–30	34	3.03	.870	.149	2.726	3.333

Note: The mean differences between groups 1 and 4, groups 1 and 5, groups 2 and 4, groups 2 and 5, and groups 3 and 4 are statistically significant at the .05 level. Dependent variable: ''I learned important skills from the process of constructing the Teaching Event'' (1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree).

[a]The total support score is a composite score (the sum) of six items on which candidates rated various potential sources of support for completing the Teaching Event on a scale of 1 (not very helpful) to 5 (very helpful), including other credential candidates, university supervisor, cooperating or master teacher, school site administrator, university instructors and professors, and the teacher education program director.

Table 3.2. Association Between Candidate Ratings of Course Work Preparation for Completing the Teaching Event and Their Perceptions of Learning from the Teaching Event, 2003–2004

Course Work Preparation for Teaching Event	*Number*	*Mean Agreement Level on Learning*	*Standard Deviation*	*Standard Error*	*95 Percent Confidence Interval*	
					Lower Bound	*Upper Bound*
1) Strongly disagree	25	1.80	.913	.183	1.423	2.177
2) Disagree	65	2.05	.856	.106	1.834	2.258
3) Agree	331	2.57	.792	.044	2.488	2.660
4) Strongly agree	161	2.93	.891	.070	2.787	3.064

Note: The mean differences between groups 1 and 3, groups 1 and 4, groups 2 and 3, groups 2 and 4, and groups 3 and 4 are significant at the .01 level.

Table 3.3. Association Between Candidate Ratings of Course Work Preparation for Completing the Teaching Event and Their Scores on the Teaching Event, 2003–2004

Course Work Preparation for Teaching Event	*Number*	*Mean Teaching Event Scores*	*Standard Deviation*	*Standard Error*	*95 Percent Confidence Interval*	
					Lower Bound	*Upper Bound*
1) Strongly disagree	15	2.469	.765	.198	2.045	2.893
2) Disagree	26	2.395	.572	.112	2.164	2.626
3) Agree	169	2.533	.584	.045	2.444	2.622
4) Strongly agree	76	2.768	.606	.070	2.629	2.906

Note: The mean difference between groups 2 and 4 and groups 3 and 4 are significant at the .01 level.

academic language and its role in the teaching of mathematics to elementary students, found that candidates responded in a range of ways and exhibited varying levels of sophistication in their understanding of how academic language and mathematics content understandings are interrelated. They suggest that even high-stakes performance assessments like the PACT that have an explicit focus on academic language can formatively support candidate learning in this area and could provide useful information to programs about their candidates' understandings of academic language and its role in the teaching of content.

Table 3.4. Association Between Candidate Perceptions of Learning from the Teaching Event and Their Scores on the Teaching Event, 2003–2004

''I learned important skills from completing the Teaching Event''	*Number*	*Mean Teaching Event Score*	*Standard Deviation*	*Standard Error*	*95 Percent Confidence Interval*	
					Lower Bound	*Upper Bound*
1) Strongly disagree	47	2.44	.668	.098	2.248	2.641
2) Disagree	84	2.56	.598	.065	2.429	2.689
3) Agree	175	2.58	.587	.044	2.491	2.666
4) Strongly agree	45	2.75	.686	.102	2.541	2.952

Note: The mean difference between groups 1 and 4 is significant at the .05 level.

In another study that examined preservice teacher learning, Nagle (2006) described the use of the PACT Teaching Event artifacts as the prompts for preservice teachers' inquiry into their own practice through guided collaborative discussions in a small seminar for science credential candidates:

> Through this study it is evident that state mandated assessments like the Teaching Event can afford learning opportunities for preservice teachers to examine their practice in an in-depth and collaborative manner. The tasks of the Teaching Event provide a structure or scaffold to investigate the primary areas of teaching—planning, instruction, assessment and analysis. As this study illustrates the tasks of the Teaching Event complement the learning goals of the teacher education program. The Teaching Event affords direction for the preservice teachers to investigate their teaching practice, but the specific issues that the preservice teachers eventually investigated were influenced primarily through the theoretical foundations of the teacher education program. Two-thirds of the issues investigated and presented by the preservice teachers in the Student Teaching Seminar involved student learning, one of the primary goals of the teacher education program [p. 15].

Nagle points out that the Teaching Event in itself is insufficient to ensure successful in-depth collaborative examination of practice and that structured, theory-guided discussions, as well as a culture of trust and community that was previously built during the program, are needed to facilitate honest examination of practice that leads to learning in practice.

The positive association that the PACT Consortium found between teachers' ratings of their program supports and preparation and their scores on the Teaching Event also suggests that the aggregated results of the scores from the PACT Teaching Events from each campus could serve as an important

indicator of program quality. But more important, surveys of program leaders and faculty have found that the results of PACT Teaching Event scores have been used in some campuses to formatively guide program review and revision (Pecheone & Chung, 2006). There is evidence from programs that the PACT scores have helped to guide program review and revision by making more clearly visible the strengths and common weaknesses of candidate performance in particular areas of teaching, such as assessment and the instruction of English learners. Perceived weaknesses in the ability of teacher candidates to design and analyze assessments of student learning have led to a greater focus on assessment literacy in preservice program courses across the PACT Consortium. Likewise, historically lower scores on the rubric dimension Academic Language, which focuses on teachers' understanding of the language demands of their lessons and their strategies to support English learners' acquisition of academic language and content, have led many programs to work toward a common understanding of academic language among program faculty members, and to attend to teacher candidates' weaknesses in this area. One California State University program director said:

> We've always gone out into the classrooms and observed and we've always assessed their work inside the classroom. Now this is asking them to write and think about, in one assessment, in an official culminating assessment process, what they're actually thinking about as they're teaching. What's important about that is that the rubrics are sufficiently detailed enough so that when we analyze the data, it gives us important information about the strengths of our program. And in terms of what students know, what do we think we're teaching, yet the students aren't getting? In other words, where are the holes in our program? And that has been really valuable for us . . . . One of the areas that we've found we need to work on is: What do student teachers do next? Once they've analyzed the student work in the assessment of student work [task], what do they need to do next to actually improve on the student learning in their classroom? And I know, because of the consortium, that's actually a weakness in many of the student teacher programs.

Another California State University faculty member and program coordinator explained:

> A major change in our program has been a stronger emphasis on academic language. It's always been a goal for the program, but I think there has been historically a conception that ''Oh well, that course will handle it and the students will remember what they learned in that one course and it will carry forward.'' But we've been doing a lot of professional development with our supervisors, in particular around academic language, to have them think about what linguistic demands are embedded in their candidates' lessons and to help the candidates

> understand that they need to think about academic language development while they're planning, not after they've done a plan and then modify the plan, that it is a consideration from the very beginning of their lessons. And when they look at student work, they want to ask, ''To what extent does the student performance reflect an academic language issue, in addition to a content learning issue?''

Peck, Gallucci, Sloan, and Lippincott (2008) describe the inquiry process used by the Teacher Education Program at the University of California-Santa Barbara to engage faculty in collaborative examination of candidate portfolio work and to address gaps in teacher performance through innovations in program design. For this purpose, all of the UCSB program faculty (including administrators) score the PACT Teaching Events (whereas at many other campuses, supervisors or nonladder faculty score Teaching Events) because it allows them to examine together the evidence of student performance in a way that creates a common language and common understandings about what preservice teachers should know and be able to do by the time they graduate from the program. It also allows an evidence-based discussion around gaps in candidate knowledge or skills.

PACT Embedded Signature Assessments (ESAs). The introduction of the ESAs was based on the finding that almost all PACT institutions had designed and developed unique assessments of teaching to support course instruction or meet programmatic or state or national accreditation requirements. After much deliberation, PACT made a strategic and practical decision to build on rather than supplant existing assessment practices. In an examination of teaching assessments across programs, clear patterns emerged:

- Teacher education programs developed assessments that were purposefully aligned to the California TPEs and were emblematic of tasks that appeared to be representative of the universities' goals and mission.
- The teaching assessments were embedded in university courses and often contributed to course grades.
- The teaching assessments occurred throughout the program (from entry to completion).
- The teaching assessments were most often used formatively to enable both instructors and teacher candidates to identify areas of strengths and weaknesses, as well as to monitor individual progress toward meeting state teaching standards.

In sum, the embedded assessments used by colleges and universities were customized campus-specific records of practice (assignments), developed by instructors using standard criteria to track a candidate's growth over time.

The PACT development committee struggled with the question of how to take into account the assessment work in which universities were already engaged and bring some rigor to the developmental process. The ESA definition was developed to signify signature features of the customized assessments. The assessments that represented signature assessments were course-embedded assignments that all candidates in a particular course of study (for example, multiple subjects, science, special education) were administered, and reliability and validity evidence was systematically collected for these signature tasks. That is, not all teaching assignments are designated as ESAs. To be considered an ESA, evidence of reliability and validity needs to be gathered for each ESA teaching assignment within a specific course of study, and candidate scores can be aggregated to inform program evaluation or accreditation. Examples of ESAs could include case studies, lesson plans, observations, classroom management plans, and other assignments or activities that fulfill the selection criteria.

In summary, the PACT system is based on the synergistic alignment of evidence that occurs at two points: during student teaching as a capstone demonstration of teacher competence through the Teaching Event and formatively throughout the program by means of ESAs. Table 3.5 contrasts the key features that distinguish the two PACT components.

Discussion of Structured Portfolios

Structured portfolio assessments provide promising examples of performance-based assessments that collect and assess multiple sources of evidence about student teacher performance that are based on valid constructs and can be scored reliably with a systematic process for training assessors to score and a moderation process to ensure that the scoring process is fair. In addition, some of these assessment systems (for example, the TWSM, RTWS, and PACT) have created scoring rubrics that describe in detail the indicators of performance at each level and across multiple scoring criteria. This increases the potential of these assessment instruments to provide detailed feedback to candidates and programs about the quality of the candidate performance on the portfolio assessment and contribute to the educative, formative purpose of the process. At least three of these cases (TWSM, RTWS, and PACT) offer substantial evidence of the formative benefits of the portfolio assessment process for candidate learning and development, and some evidence of the potential of the assessments to serve a formative purpose in the context of program self-study and revision. However, we also know that the quality of candidates' learning experiences in relation to a summative portfolio assessment is strongly shaped by the quality of program implementation of these assessments. Programs must provide multiple opportunities for candidates to practice and hone the kinds of skills measured by these assessments (for example, analyzing evidence of

Table 3.5. Key Features of the PACT Teaching Event and Embedded Signature Assessments

Key Feature	*Teaching Event*	*Embedded Signature Assessment*
Evaluation purpose	Summative.	Formative.
Timing	Capstone event usually during the student teaching placement.	Continuous; administered throughout program.
Records of practice	Common tasks focused on three to five days of instruction organized around planning, instruction (videotape), assessment (student work for whole class and two students), reflection, and academic language.	Customized university-based teaching assignments, linked to one or more teaching standards or specific to the program mission; for example, ESAs may include course assignments, fieldwork, case studies, observations.
Context	During student teaching, guided by detailed subject-specific handbooks and implementation guidelines for faculty and staff.	Embedded in university courses; includes a range of customized teaching activities aligned to standards or the mission or goals of the teacher preparation program.
Scoring system	Trained scorers within disciplines who meet calibration standards; benchmark Teaching Events within each discipline; standardized rubric within each discipline; failing TEs are judged by multiple raters; and overall there is a 10 percent program audit within each IHE.	Customized rubrics aligned to standards (TPEs) and matched to specific tasks; generally scored by faculty, instructors, clinical supervisors.
Stakes	High stakes for California preliminary credential. Requirement that all candidates meet or exceed a proficiency standard for the Teaching Event in order to be eligible for an initial teaching license.	Low to moderate stakes: formative feedback to monitor progress on the TPEs or is included in course grades and supports program evaluation.

what students have learned in pre- and postassessments, supporting students' acquisition of academic language) by embedding these skills in their course work and fieldwork experiences and providing ongoing feedback to candidates on these skills.

The work and the costs associated with creating and validating a portfolio assessment, administering and scoring the assessment on a large scale, as well as the additional work it creates for teacher candidates and credential programs supporting candidates as they complete the assessment, need to be addressed. We cannot ignore the effort and resources that went into creating and validating the Cal TPA, as well as the PACT assessment system. This is why a state law and public funding of the work (in the case of the Cal TPA), and combining the resources and expertise across consortia (for example, the Renaissance Group and PACT Consortium), have been so indispensable to getting the work done well. It seems imperative that for assessment systems like those described here to be adopted, a targeted investment in both fiscal resources and the tools and technologies needed to streamline the work involved will be necessary.

DISCUSSION

U.S. policy for developing a competitive, highly effective teaching force is rooted in state-by-state licensure requirements that generally focus on measuring basic core knowledge (reading, writing, and mathematics) and content knowledge. These are proxy measures of teacher quality that are designed to assess minimum competency to meet legal requirements for licensure assessment. Importantly, a few states have statutes that require teachers to demonstrate their teaching competence in the classroom. The importance of gathering evidence about classroom practice in making teacher certification decisions was highlighted in a report from the National Research Council (Mitchell, Robinson, Plake, & Knowles, 2001). The authors of the report concluded that ''paper and pencil tests provide only some of the information needed to evaluate the competencies of teacher candidates'' and called for ''research and development of broad based indicators of teaching competence,'' including ''assessments of teaching performance in the classroom'' (p. 172). Thus, our investigation of curriculum-embedded assessments of teaching led us to identify broader university policies and practices that focus on how existing evidence-based embedded assessments of teaching were aligned, interconnected, and used to support candidate learning and program change. Because instruments to assess teaching differ widely across programs in quality and practice, it is difficult to identify particular reforms or innovative methodologies that can serve as a unifying model for raising the level and quality of teacher education programs.

Fostering change in teacher assessment within teacher education programs can be viewed from three perspectives that together provide a framework for understanding the evolution and current state of the art of assessment in teacher education: a design perspective, a sociocultural perspective, and a policy perspective. These are depicted in Figure 3.1.

The design perspective includes those educational aspects of the assessment of teaching that have been identified as key indicators of effective performance. In the case of this review, these include identifying those research-based teacher-preparation assessment practices that can stand alone or be combined to support valid judgments of a candidate's teaching performance. Promising practices that have been identified in this review include observation protocols (CLASS, ITQ), teacher work samples (Renaissance Group), and portfolio assessment (PACT). Of the three assessment types, both the teacher work sampling

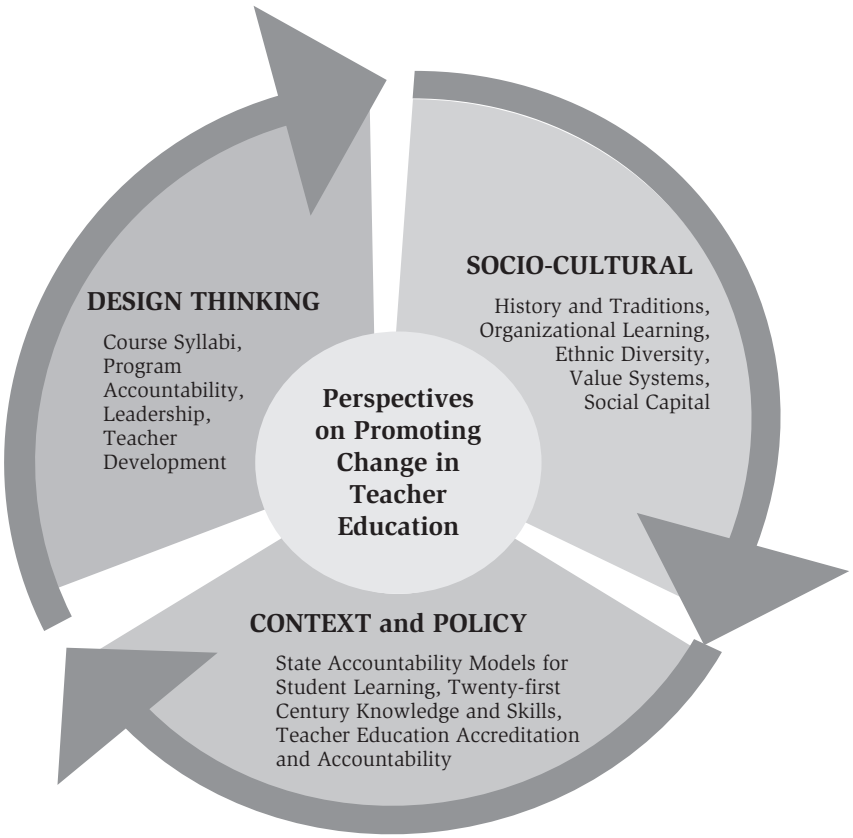

Figure 3.1 Three Perspectives for Promoting Organizational Change in Teacher Education

methodology and PACT incorporate the use of multiple measures of teaching and learning to evaluate teacher competence. That is, both systems include the collection of similar records of practice: lesson plans, teacher assignments, videotapes or observations of teaching, and student work samples. However, the PACT assessment also provides universities additional opportunities to gather both formative (ESAs) and summative evidence (TE) to enable candidates and faculty to both monitor candidate progress and use this information for program improvement and initial licensure.

Other considerations in designing credible and defensible systems to assess teaching focus on decisions regarding whether to proceed with development using a generic or subject-specific framework for design, decisions about training and scoring, decisions about calibration and benchmarking, decisions about feedback and support, and decisions about standard setting. These decisions around assessment design and development are essential because they provide a window into the thinking behind what a program values regarding their views of effective teaching and learning. Therefore, the assessment of classroom practice can serve either to reinforce the value, respect, complexity, and challenges presented in teaching or, if conceptualized too narrowly, to deconstruct teaching in ways that marginalize and trivialize teaching and learning. Choosing the right path is embedded in the values, norms, and conversations within universities around their philosophy and definition of effective teaching.

The sociocultural perspective addresses in part the social, cultural, and historical features of teacher education programs. The implicit theory underlying current assessment policies and practices in teacher education is that programs will improve when objective evidence of candidate outcomes becomes the basis for making judgments of teacher quality (Fallon, 2006; McDiarmid & Peck, 2007). But it is unclear whether teacher education programs will actually act on objective evidence of candidate performance when they have it. In fact, research on evidence-based decision making in other fields suggests the opposite effect (Estabrooks, 2007; Putnam, Twohig, Burge, Jackson, & Coix, 2002). What appears to be needed in teacher education is to investigate what changes in social practice, collaboration, and organizational learning should be put in place to support organizational innovation and change. Teacher education is a complex system in which multiple roles and practices are orchestrated and enacted across many stakeholders: tenure line faculty, adjunct faculty, clinical supervisors, cooperating teachers, and prospective teachers. Therefore, the organizational norms and structure of the work, including the protocols and tools used to inform practice, the unique ways it is carried out by individuals and programs, and the social capital that is needed to effectively use assessment data to inform change, are at best challenging and often not addressed in making assessment decisions. These organizational structures must be considered if evidence-based assessment is to become the policy lever used to

drive changes in teacher education. Clearly we have a lot to learn about how teacher education programs will respond to high-stakes assessment initiatives (licensure and accreditation) and how these initiatives can be used to promote innovation and positive change.

The policy context perspective is situated in the national trend of adopting high-stakes accountability policies for students, schools, and districts (for example, No Child Left Behind). Success or failure of schools and teachers is now determined by standardized tests and external evaluations that often delimit educational outcomes to a small subset of content knowledge and skills, such as standardized state tests of literacy, mathematics, and science. Consequently, this laser-like focus on closing the gap in student achievement is closely tied to the processes of accrediting, promoting, rewarding, or punishing schools and teachers. It is in this policy context where the outcomes of assessment practice in teacher education are being discussed and evaluated. Research in teacher education appears to be following a similar accountability model that is being used to judge student performance and school performance with value-added statistical modeling. The use of value-added methodologies to evaluate teacher education programs is an emerging trend, fueled by Carnegie Foundation–funded grants from the Teachers for a New Era project that sponsored value-added studies to examine the relationship between teacher education program practices and student learning. While maintaining a focus on learning in teacher education is important, defining learning around basic knowledge and skills can lead to narrowing teaching to content and methods beneficial to attaining predefined results. The challenge ahead is to broaden our definition of teaching and learning to focus on raising standards for learning to better prepare all students for college and the workplace, giving significant attention to all aspects of a teacher's professional competencies: dispositions, cultural sensitivity, content knowledge, and teaching skills.

Assessment practices in teacher education do not occur in isolation; they are influenced and shaped by the sociocultural and policy demands on the system. In addition, the technical aspects of assessment design are essential tools in the development of a curriculum-embedded evidence-based system that promotes organizational learning and change as evidenced by improved educational outcomes for all children.

Notes

1. According to Wiggins (1989), an authentic assessment is one that replicates the challenges and standards of performance that typically face actual practitioners. In addition, legitimate assessments are responsive to individuals and contexts. Accurate and equitable evaluation also entails dialogue with the person being assessed, allowing clarification of questions and explanations of answers.

2. The survey had a 65 percent response rate, representing 240 institutions and 65,000 education degrees awarded each year.
3. While generalizability coefficients depend on the relative standing or ranking of individual scores, dependability coefficients are used when making absolute decisions about the level of performance (as in pass/fail decisions) (Shavelson & Webb, 1991).
4. In California, there is a wide range of teacher credential program designs, including blended undergraduate programs, district intern programs, and two-year master's programs. However, most teacher credential programs in California are one-year postbaccalaureate programs, sometimes leading to a master's degree. Most preservice teachers are enrolled for two or three semesters or quarters of teacher education course work, with concurrent student teaching experiences.
5. Both authors of this chapter have participated in the development of the PACT assessment, and we devote more space to this than to the other assessments. As members of the technical development group leading the design and implementation of the PACT, we draw from the technical report and other published articles that we have written about the PACT assessment for this chapter. The PACT handbooks and the PACT technical report can by found online at http://www.pacttpa.org.
6. Even with funding, this work could not have been completed without the in-kind contributions (the sweat equity) put forth by volunteer faculty and staff of the consortium members, and the investment of member programs to provide the release time and cover travel costs so that faculty and staff could contribute thousands of hours of work to the project.
7. *Academic language* is defined as ''the language needed by students to understand and communicate in the academic disciplines'' (Pecheone & Chung, 2007, p. 9). It includes specialized vocabulary, conventional text structures within a field (for example, essays, lab reports), and other language-related activities typical of classrooms (for example, expressing disagreement, discussing an issue, asking for clarification). Academic language includes both productive and receptive modalities.

REFERENCES

Anderson, R. S., & DeMeulle, L. (1998). Portfolio use in twenty-four teacher education programs. *Teacher Education Quarterly, 25*(1), 23–32.

APA/NCME/AERA (American Psychological Association, National Council on Measurement in Education, & American Educational Research Association). (1999). *Standards for educational and psychological testing 1999.* Washington, DC: Author.

Arends, R. I. (2006a). Performance assessment in perspective: History, opportunities, and challenges. In S. Castle & B. D. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 3–22). Lanham, MD: Rowman and Littlefield.

Arends, R. I. (2006b). Summative performance assessments. In S. Castle & B. D. Shakle (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 93–123). Lanham, MD: Rowman and Littlefield.

Athanases, S. Z. (1994). Teachers' reports of the effects of preparing portfolios of literacy instruction. *Elementary School Journal, 94*(4), 421–439.

Bunch, G. C., Aguirre, J. M., & Tellez, K. (2008, January 19). *Language, mathematics, and English learners: Pre-service teachers' responses on a high stakes performance assessment.* Paper presented at the Center for the Mathematics Education of Latinas and Latino Research Symposium, Santa Cruz, CA.

California Commission on Teacher Credentialing. (2003a, July). *California TPA field review report.* Sacramento: Author.

California Commission on Teacher Credentialing. (2003b, July). *Scoring analysis for the field review of the California TPA.* Sacramento: Author.

California Commission on Teacher Credentialing. (n.d.). *California Teaching Performance Assessment (CalTPA).* Retrieved March 10, 2009, from http://www.ctc.ca.gov/educator-prep/TPA-files/CalTPA-general-info.pdf.

Castle, S., Groth, L., Moyer, P. S., & Burns, S. (2006). Course-based performance assessments. In S. Castle & B. D. Shaklee (Eds.), *Assessing teacher performance: Performance-based assessment in teacher education* (pp. 49–92). Lanham, MD: Rowman and Littlefield.

Castle, S., & Shaklee, B. D. (Eds.). (2006). *Assessing teacher performance: Performance-based assessment in teacher education.* Lanham, MD: Rowman and Littlefield.

Chung, R. R. (2005). *The Performance Assessment for California Teachers and beginning teacher development: Can a performance assessment promote expert teaching?* Unpublished doctoral dissertation, Stanford University School of Education.

Chung, R. R. (2007, April). *Beyond the ZPD: When do beginning teachers learn from a high-stakes portfolio assessment?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

Chung, R. R. (2008). Beyond assessment: Performance assessments in teacher education. *Teacher Education Quarterly, 35*(1), 7–28.

Chung, R. R., & Whittaker, A. K. (2007, April). *Preservice candidates' readiness for portfolio assessment: The influence of formative features of implementation.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.

CLASS (Classroom Assessment Scoring System). (n.d.). *Technical appendix.* Retrieved June 2, 2008, from http://www.brookespublishing.com/class2007/CLASS_Pre-K.pdf.

Coggshall, J., Max, J., & Bassett, K. (2008, June). *Key issue: Using performance-based assessment to identify and support high-quality teachers.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved June 13, 2008, from http://www.tqsource.org.

Crocker, L. (1997). Assessing content representativeness of performance assessment exercises. *Applied Measurement in Education, 10*, 83–95.

Danielson, C., & Dwyer, C. A. (1995). How Praxis III supports beginning teachers. *Educational Leadership, 53*(6), 66–67.

Danielson, C., & McGreal, T. (2000). *Teacher evaluation to enhance professional practice*. Alexandra, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2001). Standard setting in teaching: Changes in licensing, certification, and assessment. In V. Richardson (Ed.), *Handbook of research on teaching* (4th ed., pp. 751–776). Washington, DC: American Educational Research Association.

Darling-Hammond, L., & Macdonald, M. B. (2000). Where there is learning, there is hope: Bank Street College of Education. In L. Darling-Hammond (Ed.), *Studies of excellence in teacher education: Preparation at the graduate level* (pp. 1–95). New York: National Commission on Teaching and America's Future; Washington, DC: American Association of Colleges for Teacher Education.

Darling-Hammond, L., & Snyder, J. (2000). Authentic assessment of teaching in context. *Teaching and Teacher Education, 16*(5–6), 523–545.

Darling-Hammond, L., Wise, A. E., & Klein, S. P. (1999). *A license to teach: Raising standards for teaching.* San Francisco: Jossey-Bass.

Davis, C. L., & Honan, E. (1998). Reflections on the use of teams to support the portfolio process. In N. Lyons (Ed.), *With portfolio in hand: Validating the new teacher professionalism* (pp. 90–102). New York: Teachers College Press.

Denner, P. R., & Lin, S. (2008). *Evidence for impact on student learning from the teacher work samples at Idaho State University.* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, New York.

Denner, P. R., Norman, A. D., & Lin, S. (2009). Fairness and consequential validity of teacher work samples. *Educational Assessment, Evaluation, and Accountability.* Retrieved May 1, 2009, from http://www.springerlink.com/content/g604442h281g0535/fulltext.pdf.

Denner, P. R., Norman, A. D., Salzman, S. A., & Pankratz, R. S. (2003, February 17). *Connecting teacher performance to student achievement: A generalizability and validity study of the Renaissance Teacher Work Sample Assessment.* Paper presented at the Annual Meeting of the Association for Teacher Educators, Jacksonville, FL.

Denner, P. R., Salzman, S. A., & Bangert, A. W. (2002). Linking teacher assessment to student assessment: A benchmarking, generalizability, and validity study of the use of teacher work samples. *Journal of Personnel Evaluation in Education, 15*(4), 287–307.

Denner, P. R., Salzman, S. A., & Harris, L. B. (2002, April). *Teacher work sample assessment: An accountability method that moves beyond teacher testing to the impact of teacher performance on student learning.* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, New York, February 2002. (ERIC Document Reproduction No. ED 463 285)

Dwyer, C. A. (1994). *Development of the knowledge base for the Praxis III: Classroom performance assessments assessment criteria.* Princeton, NJ: Educational Testing Service.

Dwyer, C. A. (1998). Psychometrics of Praxis III: Classroom performance assessments. *Journal of Personnel Evaluation in Education, 12*(2), 163–187.

Educational Testing Service. (2002). *ETS Standards for Quality and Fairness 2002*. Princeton, NJ: Author.

Estabrooks, C. (2007). A program of research on knowledge translation. *Nursing Research, 56*(4), 4–6.

Fallon, D. (2006). *Improving teacher education through a culture of evidence.* Paper presented at the Sixth Annual Meeting of the Teacher Education Accreditation Council, Washington, DC. Retrieved January 15, 2008, from http://www.teac.org/membership/meetings/Fallon remarks.pdf.

Girod, G. (Ed.). (2002). *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology.* Washington, DC: AACTE Publications.

Girod, G. (2008). Western Oregon University: Cook School District simulation. In A. E. Wise, P. Ehrenberg, & J. Leibbrand (Eds.), *It's all about student learning: Assessing teacher candidates' ability to impact P-12 students* (pp. 213–215). Washington, DC: National Council for the Accreditation of Teacher Education.

Girod, M., & Girod, G. (2006). Exploring the efficacy of the Cook School District simulation. *Journal of Teacher Education, 57*(5), 481–497.

Girod, G., & Schalock, M. (2002). Does TWSM work? In G. Girod (Ed.), *Connecting teaching and learning: A handbook for teacher educators on teacher work sample methodology* (pp. 347–358). Washington, DC: AACTE Publications.

Gropper, N. (2008). Bank Street College of Education child study (for observation and recording, ''O&R'') In A. E. Wise, P. Ehrenberg, & J. Leibbrand (Eds.), *It's all about student learning: Assessing teacher candidates' ability to impact P-12 students* (pp. 191–202). Washington, DC: National Council for the Accreditation of Teacher Education.

Groth, L. A. (2006, October). *Performance based assessment in a preservice literacy class: The reading, writing, spelling analysis.* Paper presented at the Annual Meeting of the College Reading Association, Pittsburgh, PA.

Haynes, D. (1995). One teacher's experience with national board assessments. *Educational Leadership, 52*(8), 58–60.

Howes, C., Burchinal, M., Pianta, R., Bryant, D., Early, D., Clifford, R., et al. (2008). Ready to learn? Children's pre-academic achievement in pre-kindergarten programs. *Early Childhood Research Quarterly, 23*(1), 27–50.

Institute for Learning. (2002). *Principles of learning.* Pittsburgh, PA: Author.

Interstate New Teacher Assessment and Support Consortium (INTASC). (2009a). *Test of teaching knowledge.* Retrieved March 14, 2009, from http://www.ccsso.org/projects/interstate_new_teacher_assessment_and_support_consortium/Projects/Test_of_Teaching_Knowledge/.

Interstate New Teacher Assessment and Support Consortium (INTASC). (2009b). *INTASC portfolio development.* Retrieved January 15, 2009, from http://www.ccsso.org/Projects/interstate_new_teacher_assessment_and_support_consortium/projects/portfolio_development/792.cfm.

Junker, B., Weisberg, Y., Matsumara, L. C., Crosson, A., Kim Wolf, M., Levison, A., et al. (2006). *Overview of the instructional quality assessment.* Los Angeles: Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing.

Kellor, E. M. (2002). *Performance-based teacher licensure in Connecticut.* Madison: Consortium for Policy Research in Education, University of Wisconsin-Madison. Retrieved June 2, 2008, from http://cpre.wceruw.org/tcomp/research/standards/licensure.php.

Kimball, S. (2002). *Performance-based teacher licensure in Indiana.* Madison: Consortium for Policy Research in Education, University of Wisconsin-Madison. Retrieved June 2, 2008, from http://cpre.wceruw.org/tcomp/research/standards/licensure.php.

La Paro, K. M., & Pianta, R. C. (2003). *Observational assessment of teaching practices.* Retrieved March 17, 2009, from http://www.virginia.edu/provost/tneuva/about.html#.

Lomask, M., Seroussi, M., & Budzinsky, F. (1997). *The validity of portfolio-based assessment of science teachers.* Paper presented at the Annual Meeting of the National Association of Research in Science Teaching, Chicago.

Long, C., & Stansbury, K. (1994). Performance assessments for beginning teachers: Options and lessons. *Phi Delta Kappan, 76*(4), 318–322.

Lyons, N. P. (1996). A grassroots experiment in performance assessment. *Educational Leadership, 53*(6), 64–67.

Lyons, N. P. (1998a). Reflection in teaching: Can it be developmental? A portfolio perspective. *Teacher Education Quarterly, 25*(1), 115–127.

Lyons, N. P. (1998b). Portfolio possibilities: Validating a new teacher professionalism. In N. P. Lyons (Ed.), *With portfolio in hand* (pp. 247–264). New York: Teachers College Press.

Lyons, N. P. (1999). How portfolios can shape emerging practice. *Educational Leadership, 56*(8), 63–65.

Mashburn, A. J., Pianta, R., Hamre, B. K., Downer, J. T., Barbarin, O., Bryant, D., et al. (2008). Measures of classroom quality in pre-kindergarten and children's development of academic, language and social skills. *Child Development, 79*(3), 732–749.

Matsumura, L. C., Garnier, H., Slater, S. C., & Boston, M. B. (2008). Measuring instructional interactions ''at-scale.'' *Educational Assessment, 13*(4), 267–300.

Matsumura, L. C., Slater, S. C., Junker, B., Peterson, M., Boston, M., Steele, M., et al. (2006). *Measuring reading comprehension and mathematics instruction in urban middle schools: A pilot study of the Instructional Quality Assessment.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.

McConney, A. A., & Ayers, R. R. (1998). Assessing student teachers' assessments. *Journal of Teacher Education, 49*(2), 140–150.

McConney, A. A., Schalock, M. D., & Schalock, H. D. (1998). Focusing improvement and quality assurance: Work samples as authentic performance measures of prospective teachers' effectiveness. *Journal of Personnel Evaluation in Education, 11*, 343–363.

McDiarmid, B., & Peck, C. (2007, March). *Theories of action and program renewal in teacher education.* Paper presented at the Annual Meeting of the Northwest Association for Teacher Education. Seattle, WA.

McKibbens, D. E., Shainker, S., & Banks-Santilli, L. (2008). Focus child assessment project. In A. E. Wise, P. Ehrenberg, & J. Leibbrand (Eds.), *It's all about student learning: Assessing teacher candidates' ability to impact P-12 students* (pp. 203–212). Washington, DC: National Council for the Accreditation of Teacher Education.

Mitchell, K. J., Robinson, D. Z., Plake, B. S., & Knowles, K. T. (2001). *Testing teacher candidates: The role of licensure tests in improving teacher quality.* Washington, DC: National Academy Press.

Moss, P. A. (1998). Rethinking validity for the assessment of teaching. In N. Lyons (Ed.), *With portfolio in hand* (pp. 202–219). New York: Teachers College Press.

Moss, P. A., Schutz, A. M., & Collins, K. M. (1998). An integrative approach to portfolio evaluation for teacher licensure. *Journal of Personnel Evaluation in Education*, *12*(2), 139–161.

Nagle, J. F. (2006, April). *Collaborative examination of practice: Using a state mandated assessment as part of teacher inquiry.* Paper presented at the Annual Conference of the American Educational Research Association, San Francisco.

Office of the Superintendent of Public Instruction. (2004). *Performance-based pedagogy assessment of teacher candidates.* Olympia, WA: Author. Retrieved March 15, 2009, from http://www.k12.wa.us/certification/profed/pubdocs/PerfBasedPedagogyAssessTchrCand6–2004SBE.pdf.

Ohio Legislative Service Commission. (2007, October). *HB 347 Fiscal note and local impact statement.* Retrieved March 15, 2009, from http://www.lbo.state.oh.us/fiscal/fiscalnotes/127ga/HB0347IN.htm.

Pankratz, R. (2008). Renaissance Partnership for improving teacher quality: Renaissance Teacher Work Sample. In A. E. Wise, P. Ehrenberg, & J. Leibbrand (Eds.), *It's all about student learning: Assessing teacher candidates' learning to impact P-12 students* (pp. 45–74). Washington, DC: NCATE.

Pecheone, R., & Chung, R. R. (2006). Evidence in teacher education: The Performance Assessment for California Teachers. *Journal of Teacher Education, 57*(1), 22–36.

Pecheone, R. L., & Chung, R. R. (2007). *The Performance Assessment for California Teachers (PACT) technical report.* Stanford, CA: PACT Consortium.

Peck, C., Gallucci, C., Sloan, T., & Lippincott, A. (2008). Organizational learning and program renewal in teacher education: A socio-cultural theory of learning, innovation and change. *Educational Research Review, 4*(1), 16–25.

Pianta, R. C. (2003). *Standardized classroom observations from pre-K to third grade: A mechanism for improving quality classroom experiences during the P-3 years.* Unpublished manuscript. Retrieved March 16, 2009, from http://www.fcd-us.org/usr_doc/StandardizedClassroomObservations.pdf.

Pianta, R., Mashburn, A. J., Downer, J. T., Hamre, B. K., & Justice, L. (2008). Effects of Web-mediated professional development resources on teacher-child interactions in pre-kindergarten classrooms. *Early Childhood Research Quarterly, 23*, 431–451.

Putnam, W., Twohig, P., Burge, F., Jackson, L., & Coix, J. (2002). A qualitative study of evidence in primary care: What practitioners are saying. *Canadian Medical Association Journal, 166*(12), 1525–1530.

Resnick, L. B., & Hall, M. W. (2001). *The principles of learning: Study tools for educators.* [CD-ROM, version 2.0]. Available at www.instituteforlearning.org.

Richert, A. E. (1987). *Reflex to reflection: Facilitating reflection in novice teachers.* Unpublished doctoral dissertation, Stanford University School of Education.

Rotberg, I. C., Futrell, M. H., & Lieberman, J. M. (1998). National Board certification: Increasing participation and assessing impacts. *Phi Delta Kappan, 79*(6), 462–466.

Salzman, S. A., Denner, P. R., & Harris, L. B. (2002, February). *Teacher education outcomes measures: Special study survey.* Paper presented at the Annual Conference of the American Association for Colleges of Teacher Education, New York. (ERIC Document Reproduction Service NO. ED465791)

Schalock, M. D. (1998). Accountability, student learning, and the preparation and licensure of teachers: Oregon's Teacher Work Sample Methodology. *Journal of Personnel Evaluation in Education, 12*(3), 269–285.

Schalock, H. D., Schalock, M., & Girod, G. (1997). Teacher work sample methodology as used at Western Oregon State College. In J. McMillan (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 15–45). Thousand Oaks, CA: Corwin Press.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer.* Thousand Oaks, CA: Sage.

Shulman, L. (1998). Teacher portfolios: A theoretical activity. In N. P. Lyons (Ed.), *With portfolio in hand* (pp. 23–37). New York: Teachers College Press.

Snyder, J., Lippincott, A., & Bower, D. (1998). The inherent tensions in the multiple uses of portfolios in teacher education. *Teacher Education Quarterly, 25*(1), 45–60.

Stein, M. K., & Matsumura, L. C. (2008). Measuring instruction for teacher learning. In D. Gitomer (Ed.), *Measurement issues and assessment for teacher quality* (pp. 179–205). Thousand Oaks, CA: Sage.

St. Maurice, H., & Shaw, P. (2004). Teacher portfolios come of age: A preliminary study. *NAASP Bulletin, 88*(639), 15–25. Retrieved November 2, 2009, from http://bul .sagepub.com/cgi/content/ abstract/88/639/15.

Stone, B. A. (1998). Problems, pitfalls, and benefits of portfolios. *Teacher Education Quarterly, 25*(1), 105–114.

Tisadondilok, S. (2006). *Investigating the validity of the Washington State performance-based pedagogy assessment process for teacher licensure.* Unpublished doctoral dissertation, Oregon State University. Retrieved February 1, 2009, from http://ir.library.oregonstate.edu/dspace/handle/1957/2232.

Tracz, S. M., Sienty, S., & Mata, S. (1994, February). *The self-reflection of teachers compiling portfolios for national certification: Work in progress.* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, Chicago.

Tucker, P. D., Stronge, J. H., Gareis, C. R., & Beers, C. S. (2003). The efficacy of teacher portfolios for teacher evaluation and professional development: Do they make a difference? *Educational Administration Quarterly*, *39*, 572–602.

University of Southern Maine. (2008). *Teachers for Elementary and Middle Schools (TEAMS) Program: 2008–09 handbook.* Retrieved March 14, 2009,

from http://www.usm.maine.edu/cehd/TED/pdfs/TEAMS percent20Handbook percent2008–09.pdf.

Wasley, P. A., & McDiarmid, G. W. (2004, June 28–30). *Connecting the assessment of new teachers to student learning and to teacher preparation.* Prepared for the National Commission on Teaching and America's Future, National Summit on High Quality Teacher Preparation, Austin, TX.

Wiggins, G. (1989). A true test: Toward more authentic and equitable assessment. *Phi Delta Kappan, 70*(9), 703–713.

White, B. (2002). *Performance-based teacher licensure in North Carolina.* Madison, WI: Consortium for Policy Research in Education, University of Wisconsin. Retrieved June 2, 2008, from http://cpre.wceruw.org/tcomp/research/standards/licensure.php.

Whitford, B. L., Ruscoe, G., & Fickel, L. (2000). Knitting it all together: Collaborative teacher education in southern Maine. In L. Darling-Hammond (Ed.), *Studies of excellence in teacher education: Preparation in the undergraduate years* (pp. 173–257). New York: National Commission on Teaching and America's Future; Washington, DC: American Association of Colleges for Teacher Education.

Wilkerson, J. R., & Lang, W. S. (2003, December 3). Portfolios, the pied piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives*, *11*(45). Retrieved August 30, 2008, from http://epaa.asu.edu/epaa/v11n45/.

Wilson, M., Hallam, P. J., Pecheone, R., & Moss, P. (in press). Using student achievement test scores as evidence of external validity for indicators of teacher quality: Connecticut's Beginning Educator Support and Training Program. *Education Evaluation and Policy Analysis*.

Wilson, S. W., Darling-Hammond, L., & Berry, B. (2001). *A case of successful teaching policy: Connecticut's long-term efforts to improve teaching and learning.* Seattle: University of Washington, Center for the Study of Teaching and Policy.

Wineburg, M. (2006). Evidence in teacher preparation: Establishing a framework for accountability. *Journal of Teacher Education, 57*(1), 51–64.

Wise, A. E., Ehrenberg, P., & Leibbrand, J. (Eds.). (2008). *It's all about student learning: Assessing teacher candidates' ability to impact P-12 students.* Washington, DC: National Council for the Accreditation of Teacher Education.

Zeichner, K., & Wray, S. (2001). The teaching portfolio in US teacher education programs: What we know and what we need to know. *Teaching and Teacher Education, 17*, 613–621.