 *it Assessment Systems*, Beth Stroble, Dean, College of Education, University of Akron, September, 2000

When she wrote this paper, Dean Stroble was concluding her position as Associate Dean at the University of Louisville. Louisville is a large urban university with a co-educational enrollment approximating 22,000. Its school of education has about 175 completers each year, virtually all at the master's level. Louisville insists that teacher candidates have a strong grounding in the liberal arts and sciences, requiring that applicants earn a bachelor's degree in an academic major in addition to completing coursework necessary to be licensed in a teaching major. It provides candidates with the same kind of hands-on, performance-based experiences they will later design for their students. In addition to her personal involvement in developing the Louisville performance-based program, Dean Stroble has been a frequent consultant to institutions and states, including Indiana and Kentucky, on performance-based teacher education.

This paper is organized around seven considerations about good assessment systems:

- Embedding candidate assessments in the curriculum, instruction and experiences of teacher preparation
- Creating assessments aligned with teacher knowledge and skill content standards
- Evaluating teaching skills, knowledge about teaching, and dispositions
- Sampling work of the students of teacher candidates and making evaluations of candidates from it
- Establishing rubrics
- Scoring and judging credibility of an assessment system (consistency, accuracy, fairness, avoidance of bias)
- Forming evaluations of candidate proficiencies from samples of assessment information

The paper is distinctive through its annotated bibliographic quality. It provides references to many works, and in several places draws principles or criteria from different authors, so that points of similarity or difference can be observed. This begins in the introductory section titled “attributes of good assessment systems,” which describes principles of good assessment practice from the AAHE Assessment Forum, NCATE 2000 standard 2 on Assessment, the state of Indiana Professional Standards Board, the Kentucky Educational Professional Standards Board, Art Wise and Donna Gollnick. Dean Stroble provides many examples, as well, from her own experience at the University of Louisville, and in Indiana.

Together these bibliographic references and personal experiences make this paper a compendium of views on the principal current issues in the field of higher education assessment as applied to teacher preparation—whether it is state perspectives on teacher candidate assessment, building portfolios around “established purposes,” aligning assessments with standards, assessing dispositions, using student work as a basis for focusing teacher preparation, the nature and use of rubrics, or forming evaluations of candidate proficiencies from samples of assessment information.

UNIT ASSESSMENT SYSTEMS

Beth Stroble, Ph. D.
Dean, College of Education
University of Akron

ATTRIBUTES OF GOOD ASSESSMENT SYSTEMS

An assessment system should be consistent with the best thinking about the nature of assessment and the principles of good practice in assessing student learning. As the director of the American Association for Higher Education (AAHE) Assessment Forum, Angelo developed this definition with feedback from forum colleagues:

Assessment is an ongoing process aimed at understanding and improving student learning. It involves making our expectations explicit and public; setting appropriate criteria and high standards for learning quality; systematically gathering, analyzing, and interpreting evidence to determine how well performance matches those expectations and standards; and using the resulting information to document, explain, and improve performance. When it is embedded effectively within larger institutional systems, assessment can help us focus our collective attention, examine our assumptions, and create a shared academic culture dedicated to assuring and improving the quality of higher education (Angelo, 1995, p. 7).

Principles of good assessment practice have been developed by AAHE Assessment Forum colleagues (Astin, et. al , 1992) and expanded by Banta, et. al (1996).:

1. "The assessment of student learning begins with educational values.
2. Assessment is most effective when it reflects an understanding of learning as multidimensional, integrated, and revealed in performance over time.
3. Assessment works best when the programs it seeks to improve have clear, explicitly stated purposes.
4. Assessment requires attention to outcomes but also and equally to the experiences that lead to those outcomes.
5. Assessment works best when it is ongoing, not episodic.
6. Assessment fosters wider improvement when representatives from across the educational community are involved.
7. Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.
8. Assessment is most likely to lead to improvement when it is part of a larger set of conditions that promote change.
9. Through assessment, educators meet responsibilities to students and to the public.
10. Assessment is most effective when undertaken in an environment that is receptive, supportive, and enabling."

The intention of the NCATE 2000 Unit Standard 2. *Assessment System and Unit Evaluation*, is entirely consistent with these principles:

The unit has an assessment system that collects and analyzes data on the applicant qualifications, the candidate and graduate performance, and unit operations to evaluate and improve the unit and its programs (NCATE 2000 Unit Standards, p. 1).

The accompanying rubrics for the unit assessment system (See Figure 1) outline the particular requirements for the system; for data collection, analysis, and evaluation; and for the use of data for program improvement. The supporting explanation addresses the need for program evaluations that are purposeful and ongoing, with attention to faculty, curriculum, instruction, and candidate performance. The measures must be fair, consistent, accurate, and free of bias. Assessments must tap various sources such as the unit, field sites, faculty across campus, candidates, graduates, and employers.

The unit has a professional responsibility to ensure that its programs and graduates are of the highest quality. Meeting this responsibility requires using information technologies in the systematic gathering and evaluation of information and making use of that information to strengthen the unit and its programs (NCATE 2000 Unit Standards, p. 11).

Figure 1
NCATE 2000 Standard 2: Assessment System and Unit Evaluation

Rubrics

Element of Standard	Unacceptable	Acceptable	Target
Assessment System	<p>The unit has not involved its professional community in the development of an assessment system. The unit's system does not include a comprehensive and integrated set of evaluation measures to provide information for use in monitoring candidate performance and managing and improving operations and programs. The assessment system does not reflect professional, state, and institutional standards. Decisions about continuation in and completion of programs are not based on multiple assessments and the assessments used are not related to candidate success. The unit has not taken effective steps to examine or eliminate sources of bias in its performance assessments, or has made no effort to establish fairness, accuracy and consistency of its assessment procedures.</p>	<p>The unit has developed an assessment system with its professional community that reflects the conceptual framework(s) and professional and state standards. The unit's system includes a comprehensive and integrated set of evaluation measures that are used to monitor candidate performance and manage and improve operations and programs. Decisions about candidate performance are based on multiple assessments made at admission into programs, at appropriate transition points, and at program completion. Assessments used to determine admission, continuation in, and completion of programs are predictors of candidate success. The unit takes effective steps to eliminate sources of bias in performance assessments and works to establish the fairness, accuracy and consistency of its assessment procedures.</p>	<p>The unit, with the involvement of its professional community, is implementing an assessment system that reflects the conceptual framework(s) and incorporates candidate proficiencies outlined in professional and state standards. The unit continuously examines the validity and utility of the data produced through assessments and makes modifications to keep abreast of changes in assessment technology and in professional standards. Decisions about candidate performance are based on multiple assessments made at multiple points before program completion. Data show the strong relationship of performance assessments to candidate success. The unit conducts thorough studies to establish fairness, accuracy and consistency of its performance assessment procedures. It also makes changes in its practices consistent with the results of these studies.</p>
Data Collection, Analysis, and Evaluation	<p>The unit does not regularly and comprehensively gather, compile, and analyze assessment and evaluation information on the unit's operations, its programs, or candidates. The unit does not use current information technologies to maintain its assessment system. The unit does not use multiple assessments from internal and external sources to collect data on applicant qualifications, candidate proficiencies, graduates, unit operations, and program quality.</p>	<p>The unit maintains an assessment system that provides regular and comprehensive information on applicant qualifications, candidate proficiencies, competence of graduates, unit operations, and program quality. Using multiple assessments from internal and external sources, the unit collects data from applicants, candidates, recent graduates, faculty, and other members of the professional community. These data are regularly and systematically compiled, summarized, and analyzed to improve candidate performance, program quality, and unit operations. The unit uses</p>	<p>The unit is implementing its assessment system and providing regular and comprehensive data on program quality, unit operations, and candidate performance at each stage of a program, including the first years of practice. Data from candidates, graduates, faculty, and other members of the professional community are based on multiple assessments from both internal and external sources. Data are regularly and systematically collected, compiled, summarized, analyzed, and reported publicly for the purpose of improving</p>

Element of Standard	Unacceptable	Acceptable	Target
		information technologies to maintain its assessment system.	candidate performance, program quality, and unit operations. The unit is developing and testing different information technologies to maintain its assessment system.
Use of Data for Program Improvement	The unit makes limited, or no use, of data collected, including candidate and graduate performance information, to evaluate the efficacy of its courses, programs, and clinical experiences. The unit fails to make changes in its courses, programs, and clinical experiences where evaluations indicate that modifications would strengthen candidate preparation to meet professional, state, and institutional standards. Candidates and faculty are not regularly provided formative feedback based on the unit's performance assessments.	The unit regularly and systematically uses data, including candidate and graduate performance information, to evaluate the efficacy of its courses, programs, and clinical experiences. The unit analyzes program evaluation and performance assessment data to initiate changes where indicated. Candidate and faculty assessment data are regularly shared with candidates and faculty respectively to help them reflect on their performance and improve it.	The unit has fully developed evaluations and continuously searches for stronger relationships in the evaluations, revising both the underlying data systems and analytic techniques as necessary. The unit not only makes changes where evaluations indicate, but also systematically studies the effects of any changes to assure that the intended program strengthening occurs and that there are no adverse consequences. Candidates and faculty review performance data on their performance regularly and develop plans for improvement.

The state of Indiana has undertaken development of unit assessment systems prior to the adoption of the NCATE 2000 standards. The assessment criteria developed for Indiana teacher education institutions' unit assessment systems provide one example of the features of a good assessment system:

- "The unit assessment system incorporates stakeholders' involvement in its development and management. Minimally, stakeholders should include content faculty, P-12 faculty and administrators, candidates in the program and program alumni.
- The unit assessment system includes evidence that the conceptual framework(s) for the unit's programs incorporate all Indiana Professional Standards Board (IPSB) standards. IPSB standards include the INTASC principles and the core IPSB content and development standards for each licensure area.
- The unit assessment system includes a coherent, sequential, assessment system for individual candidates that include performance assessments. The unit assessment system is shared with candidates, utilizes, for both formative and summative purposes, a range of performance-based assessment strategies throughout the program, and has multiple decision points.

- The unit assessment system uses the collective presentation of candidate assessments and related data to document the quality of programs to prepare candidates to meet the IPSB standards.
- The unit assessment system uses aggregated assessments from individual candidates and other sources to refine and revise the conceptual framework and programs.
- The unit insures that its assessment system is continuously managed.
- The unit assessment system provides for review and revision of the assessment system" (Criteria and Expected Evidence, July 23,1999, pp. 1-3).

The Indiana institutions' unit assessment systems operate at two levels: to measure program quality and individual candidate quality. As the IPSB Teacher Education Committee has worked with institutions to consider revisions of teacher education curriculum and assessments that focus on "what a teacher can do," they imagine this: a complete teacher preparation curriculum that incorporates all IPSB standards along with a comprehensive, systemic, assessment plan that permits multiple formative and summative decision points (Ingersoll & Scannell, 1998, pp. 6-7).

The mark of exemplary progress on this work will be a "performance driven system implemented with a range of assessments for summative and formative purposes." Psychometric properties needed for high stakes decisions will have been established. Systematic articulation of appropriate standards will be evident. Candidates will have sufficient and non-redundant opportunities to experience and demonstrate appropriate standards. They will routinely demonstrate proficiency in outcomes appropriate to a beginning teacher. And the program itself will be subject to routine, internal, formative review and revision. (Ingersoll & Scannell, 1998, pp. 6-7).

Similarly, Kentucky's Educational Professional Standards Board (EPSB) has conceptualized unit assessment systems as continuous assessment plans. Such plans are "an institution's internal quality control mechanism to ensure that teacher preparation programs consistently address and integrate the appropriate performance standards and the EPSB's policies. A continuous assessment plan honors each institution's human, financial, and physical resources and reflects the institution's singular mission statement and student population. Each institution's plan is the product of intense reflective analysis by faculty, administrators, and staff, and supports the institution's claim of a quality teacher preparation program. In that each teacher preparation program is unique, so too is each assessment plan. Continuous assessment goes beyond individual student assessments, which are necessary but insufficient components of the assessment loop. Student assessments are part of the total assessment plan that continually operates to improve the quality of the institution's programs and its program graduates" (Guidelines for the Submission of Continuous Assessment Plans, 1997, p. 1). As institutions have submitted their continuous assessment plans for approval to a standards board committee, they described:

- How the plan relates to the institution's conceptual framework model including the incorporation of Kentucky standards
- Types of data generated at reference points: admission, mid-checkpoint, exit, and follow-up
- How the data are used to improve the effectiveness of candidates and the overall program.

As Wise and Gollnick describe the requirements of performance-based accreditation, they note that faculty must first:

develop consensus on the knowledge, dispositions, and skills that candidates should possess before being recommended for licensure or completing a program. In developing this consensus, they must consider their institution's mission, their unit's philosophy, their state's licensing standards and standards set by the profession. But the new system takes this planning step one step further. Faculty must address the types of evidence they will use to determine whether candidates can pass muster. (Wise & Gollnick, 2000, p. 1).

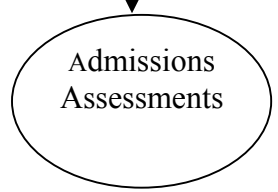
The components of a unit assessment system that will meet the specific requirements of the NCATE 2000 performance-based accreditation standards are graphically displayed in Figure 2. Developing the component parts and assuring that they honor the best thinking about assessment involves these considerations:

- Embedding candidate assessments in the curriculum, instruction, and experiences of teacher preparation
- Creating assessments aligned with teacher knowledge and skill content standards
- Evaluating teaching skills, knowledge about teaching, and dispositions
- Sampling work of the students of teacher candidates and making evaluations of candidates from it
- Establishing rubrics
- Scoring
- Judging credibility of an assessment system (consistency, accuracy, fairness, avoidance of bias)
- Forming evaluations of candidate proficiencies from samples of assessment information.

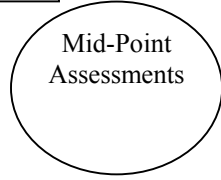
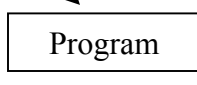
The following sections of this paper provide guidelines, examples, and references for ways in which requirements for a unit assessment system can be met.

Figure 2
Unit Assessment System Model
Institutional Mission
Unit Conceptual Framework
State Licensing Standards and Assessments
Content and Pedagogical Standards for National Professional Organizations
Standards for P-12 Students
NCATE Standards

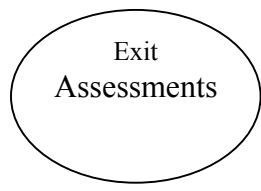
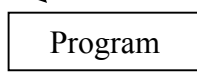
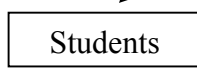
Expectations for Candidate Performance: Learning Outcomes



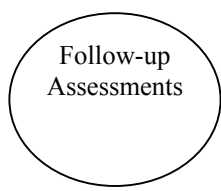
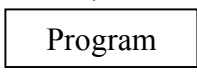
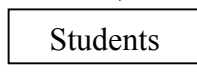
Reviewed by unit and A&S faculty and school partners with feedback to



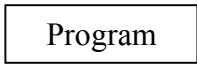
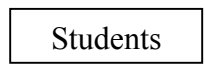
Reviewed by unit and A&S faculty and school partners with feedback to



Reviewed by unit and A&S faculty and school partners with feedback to



Reviewed by unit and A&S faculty and school partners with feedback to



Embedding candidate assessments in the curriculum, instruction, and experiences of teacher preparation

The process of embedding assessments in the curriculum, instruction, and experiences of teacher preparation can have multiple starting points. The goal is to have all the components in place as depicted in Figure 2. While an ideal process might begin at the top of the diagram and move downward to bring coherence to the conceptual framework, learning outcomes, assessment, and analysis of data for the benefit of candidates and programs, the actual process used by institutions may have many starting points. The process of development and refinement is not necessarily linear. As McTighe writes about performance assessment in P-12 schools: "Not only assessment needs to change. Curriculums and instructional strategies, too, must reflect a performance orientation" (1997, p. 6). Wiggins (1994), too, explains the logic of designing assessment tasks, but cautions against thinking this is a sequential undertaking:

- "desired outcome (to be assessed)
- criteria implied (conditions to be met for the outcome to be realized)
- indicators (concrete signs of criteria being met)
- contexts in which outcome occurs (the different kinds of settings/challenges related to this outcome)" (p. 9).

He points out that the logic of design does not equate with chronology in design—that the design process is recursive. Institutions may begin with the design of assessments and then work backwards to revise learning outcomes and conceptual frameworks and then create databases to support data gathering and analysis. Those who are designing a unit assessment system may begin at any point and work backwards and forwards, later assuring that the logic of design is honored. As many institutions have begun to conceptualize a unit assessment system that is embedded in their teacher preparation programs, they often find that pieces of the system have already been implemented, and the process is not one of creating an entire system from scratch. Experiences of Kentucky and Indiana institutions, those most known to me, are pertinent here.

The implementation of portfolios as a primary tool for performance assessment at the University of Louisville School of Education began with individual professors in courses who worked to build more authentic assessments of secondary education courses housed in professional development schools. Because the teacher education programs had adopted a performance orientation, the assessments followed course. Portfolios began in individual courses and moved to programmatic assessments. As such, they became a catalyst for programmatic change (Lyons, Stroble, & Fischetti, 1997). And, as they became part of required continuous assessment plans for the programs and the unit, their assessment focused not only on individual candidates but also to examining trends in

candidates' development within and across programs. Further changes came in programmatic portfolios as the conceptual framework was revised to include program themes that extended beyond the Kentucky Teacher Standards. Additional evidence was needed to show candidate's understanding of the complex lives of students and adults in school and society, for example. Initial purposes for portfolio assessment focused on assessing candidate progress for formative feedback to candidates. Later, program faculty used portfolios to make summative decisions about students—at admission, advancement, program completion, and recommendation for licensure. Because the institution is required to verify that candidates have met Kentucky standards prior to recommendation for initial and advanced licenses, the portfolio is used for this purpose as well. And many candidates find the portfolio a useful tool for gaining employment and advancement. Thus, the portfolio that began as a performance assessment in a course for the purpose of meeting course requirements has added purposes and functions which are consistent with the intention of a unit assessment system.

In Indiana, some institutions began unit assessment system design by reconsidering their conceptual frameworks to include INTASC principles and the Indiana standards. Others mapped existing frameworks to the principles and then determined gaps for revisions. Other Indiana institutions considered performance assessments already in place, whether field observations, microteaching, journals, or portfolios, and then matched those to INTASC and Indiana standards as a starting place. Some Indiana institutions began with revision of teacher education courses, requiring all faculty to indicate the match of course topics and assessments to the principles and standards. Because Danielson's (1996) book provides a useful framework for linking INTASC principles and proposed PRAXIS III assessments with teacher portfolios, many found her book a valuable resource for aligning programs with required standards and performance assessments.

Many models exist for embedding assessments in the programs. Florida has worked with educators to develop standardized performance tasks for use in programs across the state. Asbury College in Kentucky developed a system of "gates" linked to the three knowledges and three functions of knowledge of their conceptual framework as well as to the Kentucky Teacher Standards. The resulting matrix identifies data summaries for checkpoints: feedback to students and feedback to the program at each gate. Western Kentucky University requires that each course include a required assessment that will ultimately become part of an electronic portfolio for candidates. While the models vary, the logic of the design of a unit assessment system, as displayed in Figure 2, can serve as a benchmark for the ongoing design and necessary revisions of system components.

Creating assessments aligned with teacher knowledge and skill content standards

Darling-Hammond, Wise, & Klein (1995) identified the shortcomings of the first generation of pedagogical examinations, finding that they were “unable to meet the goal of a licensing assessment system to represent a knowledge base and, thus improve the quality of preparation” (p. 51) of teacher candidates. They faulted paper-and-pencil tests

for decontextualized measurement of facts about teaching rather than allowing “demonstrations of teacher knowledge, judgment, and skills in the kinds of complex settings that characterize real teaching” (pp. 51-52). Similarly, they found that performance assessments of teaching skills often measured trivial aspects of teaching, limited assessments to one classroom setting, and confused employment with licensing decisions.

Portfolios have become an increasingly popular method of assessing candidate performance, particularly because of the multiple purposes such assessments can serve when they are conceived of as more than scrapbooks of assigned artifacts. With increasing frequency, teacher education faculties require students to assemble professional portfolios to accomplish multiple purposes. Models for and study of portfolio implementation are plentiful (See, for example, Campbell, et al., 1997; Martin, 1999; Yancey & Weiser, 1997; California Council on Teacher Education, Winter 1998). Barton and Collins (1993) assert:

The development of a portfolio begins with the act of establishing purposes. Students, with the help of an advisor, develop purposes for their studies by establishing what they need and want to learn in order to become master teachers. Once they establish these purposes, students seek to find and create practices that meet the need. Because the portfolio emphasizes purpose, students have real reasons to look for connections between theory and practice. (p. 202).

The portfolio design that distinguished their candidates’ portfolios includes:

1. "explicitness of purpose
2. integration of academic course work and field experiences
3. multisourced, containing a variety of evidence for assessment
4. authentic links between classroom instruction and portfolio evidence
5. dynamic, showing growth and change in candidates over time because candidates select work at various points in their learning
6. candidate ownership as evidenced by personal reflections and self-evaluations
7. multipurposed: for course completion, candidate evaluation, program evaluation, and job placement."

The decision about how to organize portfolio evidence is one best shared with candidates because the decision—whether by chronology, theme, topic, kind or source of evidence, or standard—affects the candidates' learning since the organization entails an evaluation of the evidence, one’s learning, and how to make a compelling case (Barton & Collins, 1993). And sharing the decisions about the criteria by which portfolios are assessed becomes an important professional growth activity for teacher candidates (Stroble, 1996). As faculties and candidates work more with the organization of portfolios, they often find that they are easily indexed to standards for content and pedagogical knowledge. But the most powerful organizers for the contents of the portfolio may transcend individual standards in ways that involve a meta-cognitive analysis of candidates' reflections on teaching and learning.

Based on the piloting of the new performance-based elementary standards, a review team has developed a review process for those standards. The NCATE web site, www.ncate.org, provides guidance for institutions preparing program reports for elementary teacher preparation beginning in the Spring of 2001. This web site provides answers to five key questions:

1. How is "performance-based" elementary program review different from the previous NCATE/ACEI program review?
2. What should be included in a submission of performance-based program evidence?
3. How will a performance-based program report be judged?
4. What are the expectations for evidence if the reviews are scheduled before an institution has an assessment system fully in place?
5. What are the characteristics of assessment systems that can provide sound evidence of candidate proficiencies in the standards?

Pertinent information from the answers helps us anticipate how to prepare for the review of performance assessment evidence from all programs. It is important to note several statements in relation to the second question. Candidate proficiency data should be aggregated and interpreted. Descriptions of rubrics or criteria should be included as well as the proportion of program enrollees or completers who attain each performance level. A few samples of candidate work should be included to demonstrate the different levels of performance. The aggregated and sampled data must show candidate proficiency in relation to the content standards in reading/language arts, mathematics, science, and social studies knowledge and teaching skills. Multiple data sources should be illustrated, from different points in the program related to the knowledge, skills, and dispositions having positive effects on student learning. "The intent is to inform reviewers about candidate proficiencies in relation to the standards."

In relation to the fourth question, the answers note that the performance plan must include three pieces:

- A context statement
- A summary of performance data currently available
- A description of the program's plan for a performance assessments together with proposed measures of candidate proficiencies and stages of implementation.

This latter piece is elaborated to include the types of assessments and the sources, with examples provided.

Finally, this document details the characteristics of an assessment system that can provide sound evidence of candidate proficiencies. Such a system:

- Results from planned, purposeful, and continuing evaluation of candidate proficiencies, drawing on diverse sources

- Represents the scope of the standards for elementary teacher preparation
- Measures the different "attributes" of standards in appropriate and multiple ways
- Results from rigorous and systematic efforts by the institution to set performance levels and judge accomplishments of its candidates
- Provides information that is credible--accurate, consistent, fair, and avoiding bias
- Makes use of appropriate sampling and summarizing procedures
- Uses data to advise candidates and to strengthen teaching, courses, experiences, and programs.

This document is invaluable in providing detailed guidance for the preparation and implementation of a unit assessment system for all professional programs.

Evaluating teaching skills, knowledge about teaching, and dispositions

The evaluation of teaching skills and knowledge about teaching have long been part of the strategies used by teacher educators, particularly in field settings. Before programs instituted early, diverse field experiences, they may have relied heavily on the paper and pencil tests and simulated tasks that Darling-Hammond, Wise, & Klein faulted for their limited and inauthentic bases. The NCATE 2000 Unit Standards indicate that assessments may exist in multiple forms:

End-of-course evaluations, written essays, or topical papers, as well as from tasks used for instructional purposes (such as projects, journals, observations by faculty, comments by cooperating teachers, or videotapes) and from activities associated with teaching (such as lesson planning, identifying student readiness for instruction, creating appropriate assessments, reflecting on results of instruction with students, or communicating with parents, families, and school communities" (p. 12).

Institutions may continue to use traditional assessments of knowledge about teaching because of their direct and efficient nature in assessing factual knowledge. Indeed, required standardized assessments of teacher knowledge, such as PRAXIS II, will likely continue to serve as important measures of candidate knowledge. With greater opportunity to observe candidates' skills, knowledge, and dispositions in real settings, however, the opportunities for authentic assessment grow. The NCATE 2000 standards requirement for assessments of candidate performance with a focus on the candidates' positive effects on student learning necessitates including what has become known as "authentic" assessment. Authenticity in assessment, as described by Wiggins (1993), means:

- Engaging, worthy problems which require use of knowledge in effective and creative applications
- Replicas of or tasks that are analogous to those of the profession
- Options, constraints, and access to resources that are faithful to real-life contexts rather than arbitrary or efficient

- Real problems that are nonroutine and multistage
- Tasks that require quality products and/or performances
- Transparent or demystified criteria and standards
- Interactions between assessor and assessee
- Concurrent feedback and possibility of self-adjustment during assessment
- Trained assessor judgment related to clear and appropriate criteria
- Search for patterns of response in diverse settings to observe consistency of work and assessment of habits of mind in performance (pp. 228-230).

Unit assessment systems should provide opportunities to gather evidence of candidates' skills, knowledge, and dispositions in ways that are consistent with these markers of authentic assessment. Even assessments that were once thought of as simulated, mock-ups of real teaching can provide authentic experiences for planning, teaching, and assessing learning when these principles are incorporated. Microteaching experiences in professional development school sites in the University of Louisville Department of Secondary Education provide one example.

While microteaching can be seen as a simulated teaching episode in which candidates teach “mock” lessons on P-12 content to their peers in controlled and videotaped settings, the University of Louisville secondary education faculty have implemented microteachings that allow more authentic teaching and learning to take place. Candidates are expected to teach mini-lessons to peers that are a genuine educational experience for them. Rather than designing lessons that mimic high school curriculum, candidates teach their peers content or skills that are new learning for them. They might, for example, model a teaching technique, such as how to use writing to support learning in the content areas, or discuss a topic of importance to teachers, such as resources for school safety or unique needs of students with specific disabilities. Because the microteaching lesson is designed to teach peers new information or skills, the candidate is expected to conduct an assessment of the lesson’s impact on peers’ learning. And because the lesson is also self-assessed and assessed by peers and the course instructor with feedback for future lessons, candidates gain collegial experiences as professional coaches for one another's' teaching. In this way the microteaching lesson becomes more authentic preparation for the kind of instructional planning and assessment candidates will implement in secondary classrooms.

Many institutions find it challenging to assess dispositions. The NCATE 2000 Unit Standard 1 describes dispositions in this way:

Candidates preparing to work in schools as teachers or other professional school personnel know and demonstrate the content, pedagogical, and professional knowledge, skills, and dispositions necessary to help all students learn” (p. 1). Further, the support explanation notes that "dispositions are not usually assessed directly; instead they are assessed along with other performances in candidates' work with students, families, and communities" (p. 8).

Rubrics can be created to assess dispositions that the faculty value and that are delineated in professional, state, and institutional standards. At the University of Southern Maine, teacher educators and clinical faculty have designed a rubric that is used as an alert device to trigger conferences with candidates about these “professionalism” topics:

- Respect for diversity
- Confidentiality
- Responsibility
- Dress/Appearance
- Respect for School Rules/Policies/Norms
- Appropriate Relations with Students
- Accepting Feedback Positively
- Positive Relations with Colleagues.

The Interstate New Teacher Assessment and Support Consortium (INTASC) proposes a number of dispositions as part of the model standards for beginning teacher licensing and development. These three are linked to principle #7: the teacher plans instruction based upon knowledge of subject matter, students, the community, and curriculum goals.

- The teacher values both long term and short term planning.
- The teacher believes that plans must always be open to adjustment and revision based on student needs and changing circumstances.
- The teacher values planning as a collegial activity.

These examples of dispositions or “commitments” (Schalock, 1999) illustrate ways in which dispositions can be observed through performances as clinical faculty and teacher education observe and interact with candidates in planning settings. In fact, these dispositions could be observed even in the more limited microteaching described above. Evidence of candidates' dispositions can be found in teaching visit journals, in letters to the reader in portfolios, in reflections on field experience activities. As teacher education faculties and professional colleagues agree on the dispositions they most value in candidates and build those into their conceptual frameworks, they will also need to gather evidence of candidates' growth in those commitments. And they will need to consider what attributes of the teacher preparation enterprise will most successfully develop those dispositions.

Sampling work of the students of teacher candidates and making evaluations of candidates from it

Dwyer and Stufflebeam’s discussion of teacher evaluation in the 1996 Handbook of Educational Psychology points out the controversial nature of using student work for teacher evaluation, noting that the controversy has centered on “technical issues such as gain scores, taking context and resources into account, and so on” (p. 781). They mention projects in Oregon, Texas, and Tennessee as examples of attempts to measure the effects of instruction on student achievement. Millman's (1997) Grading Teachers, Grading Schools: Is Student Achievement a Valid Evaluation Measure raises important

questions about the potential and problems of initiatives across the county. Although difficulties may accrue from assessing teacher instruction by student learning, the NCATE 2000 standards are clear: "Teachers and teacher candidates have student learning as the focus of their work" (p. 8).

The Oregon Teacher Work Sample Methodology, while critiqued by Darling-Hammond (1998) for over-reliance on measures of limited kinds of performances and the inability to link definitively teaching to specific outcomes, offers a model that intends to:

encourage teachers and teacher educators to go beyond simple acquisition of the knowledge, skills, and dispositions thought to engender success as a teacher and to focus more strongly on ways to integrate and apply these enabling factors in order to foster students' learning progress. (Schalock, Schalock, & Myton, 1998, p. 470).

Shalock (1999) affirms that the work in Oregon and Washington, and as called for in the NCATE 2000 standards, moves beyond the "presumptive" approach to teacher preparation and licensure.

Instead of presuming teacher effectiveness as a facilitator of K-12 student learning on the basis of teacher demonstrated knowledge and skills, these two states, and NCATE, are requiring additionally that intending teachers demonstrate they are in fact able to foster learning gains in students taught (in NCATE language, "have a positive impact on student learning" (p. 2).

The work sample methodology has been adopted by a group of ten Renaissance Partnership institutions funded by a Title II grant directed by Roger Pankratz at Western Kentucky University. Their adaptation requires that teachers show specific connections of work samples and assessments to state and national standards (Performance Assessment in Progress, pp. 4-5, 8).

Performance assessments in place in other states include a case study combined with an interview with clinical and school faculty at the University of Indianapolis. The ratings on the interviews have been monitored for acceptable interrater reliability to enable decisions about a candidate's continuation in the program. In subsequent years, the measures become more classroom-based because the program courses move to P-12 classrooms. At Maryville University in St. Louis, rating scales are used to assess applicants' promise as candidates while they serve a practicum of 120 hours in P-12 schools. At Alverno College in Wisconsin, performance assessments are embedded across the college curriculum and are not limited to the teacher education program. Performances are judged by peers, faculty, and sometimes the public.

As Darling-Hammond (1998) defends the INTASC licensure assessments for beginning teachers and the NBPTS assessments of accomplished teaching as more appropriate for high-stakes decisions, she notes these characteristics of the assessments and their scoring:

- "Substantially standardized.
- Linked to standards of practice.

- Include on-demand performance tasks and samples from classroom work.
- Validated and tested for reliability in scoring and standard setting.
- Vetted for issues of bias" (1998, p. 472).

Whether institutions choose from the above methods or others to insure that their focus and that of their candidates is on student learning, these characteristics are worthy of attention.

Establishing rubrics

Rubrics, the scoring guidelines for evaluating candidates' work, answer these questions:

1. "By what criteria should performance be judged?
2. What does the range in the quality of performance look like?
3. How should the different levels of quality be described and distinguished from one another?" (Center on Learning, Assessment, and School Structure, 1995, p. 1).

Rubric developers must make decisions about the holistic or analytic nature of the scoring guide—whether the guide should describe the performance as a whole or contain sub-scales for multiple dimensions of the performance. CLASS, (The Center on Learning, Assessment, and School Structure) offers these criteria for the best rubrics:

- "Discriminate among performances, validly, not arbitrarily—by the central features of the performance, not by the easiest to see, count, or score.
- Do not try to combine independent criteria in one rubric
- Are based on an analysis of many work samples, and based on the widest possible range of work samples—including valid exemplars.
- Rely on descriptive language, highlighting the distinctive features of samples representing a level—as opposed to relying heavily on mere comparatives or value language (e.g. 'not as thorough as,' or 'excellent product') to make the discrimination.
- Provide useful and apt discrimination to enable sufficiently fine judgments—but not using so many points on the scale as to threaten reliability (typically involving, therefore, 6-12 points on a scale)
- Use descriptors that are sufficiently rich to enable student performers to verify their score, accurately self-assess, and self-correct.
- Highlight the judging of the 'impact' of performance as opposed to over-rewarding the processes used, the format used, and/or the good-faith effort made" (1995, p. 2).

This last criterion is important as assessments of candidates focus on the outcomes of their performances, especially the effect on student learning, rather than on more traditional input criteria such as the structure of a lesson plan format or the correct form for lesson objectives. CLASS describes four kinds of rubric criteria: impact, process, form, and rule. A balance among these kinds of criteria can provide information in

formative and summative ways about candidates' performances for the candidates and for programs. Validity issues of performance assessments are partially played out in the development of criteria (Wiggins, 1994; Herman, Ashbacher & Winters, 1992). If a performance can meet the criteria yet not honor the intention of the desired outcome, then validity becomes an issue. Similarly, if a performance can be deemed successful in demonstrating the outcome yet not measure well on the criteria, another measurement issue threatens the credibility of the assessment. Credibility is tied to the "interplay of evidence and consequences" (Messick, 1994). Messick argues that performance assessments must represent thoroughly the construct being assessed and diminish any variance that is irrelevant to the construct being assessed. Further, he emphasizes the need for attention to the positive and negative consequences in the validation of performance assessments. In operational terms, issues of credibility are often played out not only in the design of the task, with the need for evidence of construct validity, but also to the construction of rubrics and scoring procedures.

Scoring and judging credibility of an assessment system (consistency, accuracy, fairness, avoidance of bias)

"The most obvious reason for consistent scoring is equity" (Herman, Ashbacher, & Winters, 1992, p. 80). When decisions about candidates and program rest on performance data scored by individuals who know the candidates, then it is especially important that scores, to the degree possible, minimize influences of the rater's ethnicity, gender, age, experience as it relates to the candidate. That such influences can affect raters' scores is easily documented (Baker & O'Neil, 1996; Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Supovitz, 1997). Yet, it is also possible to achieve consistency by having well-defined and defensible criteria and moderating judgements to achieve a consensus among raters about the meaning of the criteria (Frederiksen, Sipusic, Sherin, & Wolfe, 1998; Herman, Ashbacher, & Winters, 1992). While training should include practice scoring sessions using the rubric, discussions to reach consensus about the meaning of the criteria, and careful examination of benchmarks from actual candidates' work, "drift" may occur across time for an individual scorer or across different kinds of candidates' work. Continual moderation and record-keeping of consistency is necessary when high stakes are attached to performance assessment scoring. And indeed, protocols may need to be developed by scorers to deal with unexpected performance circumstances. The work of Wiggins (1993) and Herman, Ashbacher, & Winters, 1992) provides practical assistance for the procedures of establishing rubrics and insuring their meaningful use. Decisions must be made about the number and characteristics of scorers and issues of maintaining candidate confidentiality for the purposes of reporting data in program reviews.

Herman, Ashbacher, & Winters recommend checking the reliability of the rating process by checking for:

- "documented, field-tested scoring guide
- clear, concrete criteria
- annotated examples of all score points
- ample practice and feedback for raters

- multiple raters with demonstrated agreement prior to scoring
- periodic reliability checks throughout
- retraining when necessary
- arrangements for collection of suitable reliability data" (pp. 93-94).

Of course, an additional concern arises that performance assessments, when used for high-stakes decisions, may lose their power because concerns for reliability will eventually narrow the allowable performances. As Jones and Whitford note about the P-12 assessments implemented in Kentucky:

Over the course of KIRIS testing, the growing pressure for increased reliability and tighter alignment with a specified body of content has translated into a shift away from "open-endedness" and performance in the assessments. The logic is clear. The more open and performance-based an assessment is, the more variety in the responses; the more variety in the responses, the more judgment is needed in scoring; the more judgement in scoring, the lower the reliability. Hence, less open-endedness and less focus on performance" (1997, p. 278).

As Wiley and Haertel (1996) caution, there are difficulties in distinguishing reliability from validity when considering the differences between multiple choice and extended assessment tasks. They conclude that it is:

important to insist that performance records be adequately designed to reflect the important (intended) skills. Ambiguity in criteria must be minimized. . . .(and) scoring criteria must be communicated and understood by the scorers (p. 86).

Forming evaluations of candidate proficiencies from samples of assessment information

Darling-Hammond, Wise, & Klein (1995) propose that a system for assessing new teachers' readiness to teach should be parsimonious, that it "should strive to create a few good and useful measures of quality rather than many poor ones" (p. 94). Given the need to make credible, summative decisions about candidate proficiencies across contexts and knowledge, skill, and disposition domains, institutions are likely to find themselves adding embedded performance assessments to more traditional assessment information, such as admission and exit tests, GPA, and so on. What is useful is a template by which assessment measures and decisions are tied to events in a candidate's program and to standards and concepts underlying the program. A variety of organizers have been used by institutions to focus on the programmatic level of assessment evaluations.

One such template used by the University of Louisville School of Education is a Conceptual Framework Consistency Chart. Each certification program completes a chart with these columns as a component of program review documents, formerly called folios:

			COURSE NUMBERS	
PROGRAM CONCEPTS	GOALS/ OBJECTIVES	CONTENT	FIELD EXPERIENCES	ASSESSED

Another template used by every program on campus at the University of Louisville is an affinity diagram (designed by Associate Provost John Welsh). These columns are included:

IDENTIFIED STUDENT OUTCOMES	ASSESSMENT STRATEGIES/ MEASURES/ CRITERIA	SCHEDULE FOR REPORTING OUTCOMES/ PLANS FOR FUTURE REPORTS	HOW RESULTS ARE USED FOR PROGRAM IMPROVEMENT
-----------------------------	---	---	--

These templates are filed with the provost's office and guide annual data summaries for the provost's office and for seven-year academic program reviews across campus. They are also used as documents in the program review documents for NCATE purposes.

These templates and many like them in use in institutions across the country are useful for building consensus about program-level student outcomes, assessments, and the specific means by which data will be gathered and used for defined purposes, whether feedback to candidates or for program review and improvement. They also serve as important ways to communicate the logic of the program and assessment design for internal and external audiences.

Sampling both situations and tasks and knowledge and skills is challenging. Those in the health professionals have found (Swanson, Norman, & Lin., 1995) the sampling methods difficult as well as a lack of predictability of performance in simulated performances, scoring difficulties which fail to reward alternative answers, and inappropriateness of context-specific measures for high-stakes testing. These difficulties highlight the importance of a well-crafted plan that allows for evaluation decisions to be shared across colleagues who work with candidates in multiple settings. Further, institutions find it necessary to assemble databases that can record data from multiple assessments for individual candidates and that can aggregate data across candidates to enable informed decisions.

Finally, institutions will find it necessary to sample assessment data for the purposes of representing it in program reviews—whether for a particular cohort of students in a program or from various data points and topics of assessment. The sampling methods for assessing individual candidates, for providing assessment data for program reviews, and for archiving sample assessments for program reviews and on-site visits should become a part of the unit assessment plan.

SUMMARY

As I have worked with institutions in several states to develop performance-based programs and assessment systems, I have found these elements to be necessary for this work to succeed:

- Perceived Need for Content and Performance-Focused Curriculum
- Trust: Collaborative Leadership and Ownership by Stakeholders
- Clear Expectations for Performance Linked to Units' Missions
- Coherence Among Program and Assessment Elements
- Variety of Assessments for a Variety of Purposes for a Variety of Audiences
- Reasonable Timelines with Opportunities for Networking and Feedback
- Support for the Individuals Involved in Development and Implementation.

The profession will be served well if we--teacher education faculty, professional colleagues, BOE members, Unit Accreditation Board Members, program reviewers, and candidates--support one another in the learning necessary to develop unit assessment systems.

References

Angelo, T. (1995). Reassessing (and redefining) assessment. AAHE Bulletin, 48(3), pp 7-9.

Astin, A.W., Banta, T. W. , Cross, K. P., El-Khawas, E., Ewell, P. T., Hutchings, P., Marchese, T. J., McClenney, K. M., Mentkowski, Miller, M. A., Moran, E. T., Wright, B. D. (1996). Nine principles of good practice for assessing student learning. AAHE Assessment Forum.

Baker, R. & O'Neil, H. F. (1996). Performance assessment and equity. In Kane, M. B. & Mitchell, R. (Eds.). Implementing performance assessment: Promises, problems, and challenges. pp. 183-199. Mahwah, NJ: Lawrence Erlbaum.

Banta, T. W., Lund, Black, & Oblander. (1996). Assessment in practice: Putting principles to work on college campuses. San Francisco: Jossey-Bass.

Barton, J. & Collins, A. (1993). Portfolios in teacher education, Journal of Teacher Education, 44(3), 200-210.

California Council on the Education of Teachers (Winter 1998). Teaching Portfolios in Teacher Education theme issue of Teacher Education Quarterly, 25(1).

Campbell, D. M., Cignetti, P. B., Melenyzer, B. J., Nettles, D. H., & Wyman, R. M. (1997). How to develop a professional portfolio: A manual for teachers. Boston: Allyn and Bacon.

Center on Learning, Assessment, and School Structure (1995). Rubrics and Scoring Criteria: Guidelines and Examples. Princeton, NJ: Author.

- Danielson, C. (1996). Enhancing professional practice: A framework for teaching. Alexandria, VA: ASCD.
- Darling-Hammond, L. (February 1998). Standards for assessing teaching effectiveness are key: A response to Schalock, Schalock, and Myton. Phi Delta Kappan, 471-472.
- Darling-Hammond, L., Wise, A. E., & Klein, S. (1995). A license to teach. Boulder, CO: Westview Press.
- Dwyer, C. A. & Stufflebeam, D. Teacher evaluation. In Berliner, D. C. & Calfee, R. C. (Eds.) 1996. Handbook of educational psychology. New York: Macmillan Reference USA, pp. 765-786.
- Educational Professional Standards Board (1997). Guidelines for the submission of continuous assessment plans. Author.
- Frederiksen, J. R., Sipusi, M., Sherin, M., & Wolfe, E. W. (1998). Video portfolio assessment: Creating a framework for viewing the functions of teaching. Educational Assessment, 5(4), 225-298.
- Herman, J. L., Ashbacher, P. R., & Winters, L. (1992). A practical guide to alternative assessment. Alexandria, VA: ASCD.
- Indiana Professional Standards Board Teacher Education Committee. (1999). Criteria and expected evidence. Author.
- Ingersoll, G. & Scannell, D. (1998). Performance-based teacher preparation and licensure in Indiana. Quality Teaching 7(2), pp. 6-8.
- Interstate New Teacher Assessment and Support Consortium. (1992). Model standards for beginning teacher licensing and development: A resource for state dialogue. Washington, DC: Council of Chief State School Officers.
- Jones, K. & Whitford, B. L. (December 1997). Kentucky's conflicting reform principles: High-stakes school accountability and student performance assessment. Phi Delta Kappan, pp. 276-281.
- Lyons, N., Stroble, B. & Fischetti, J. (1997). The idea of the university in an age of school reform: The shaping force of professional development schools. In M. Levine & R. Trachtman (Eds.). Making professional development schools work: Politics, practice, and policy. (pp. 88-111). NY: Teachers College Press.
- Martin, D. B. (1999). The portfolio planner: Making professional portfolios work for you. Upper Saddle River, NJ: Merrill.

McTighe, J. (1997). What happens between assessments? (December 1996/January 1997), Educational Leadership, 6-12.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. Educational Researcher, 23(2), pp. 13-23.

Millman, J. (1997) (Ed.). Grading teachers, grading schools: Is student achievement a valid evaluation measure? Thousand Oaks, CA: Sage Press.

National Council for Accreditation of Teacher Education. NCATE 2000 unit standards. Washington, DC: Author.

Performance assessment in progress. Quality Teaching, 9(2), 4-5;8.

Schalock, D. (1999). Framing pages for the Monday AM Session of the Pete Workshop: Assessment evidence related to K-12 Student Learning. Presentation at Philadelphia, Adams Mark Hotel, August 15-17, 1999.

Schalock, D., Schalock, M. & Myton, D. (February 1998). Effectiveness—along with quality—should be the focus. Phi Delta Kappan, 468-470.

Sizer, T. (1992). Horace's School: Redesigning the American High School. Boston: Houghton-Mifflin.

Stroble, E. (1996). Portfolio pedagogy: Assembled evidence and unintended consequences. Teaching Education 7(2), 97-102.

Supovitz, J. A. (November 5, 1997). From multiple choice to multiple choices. Education Week on the Web.

Swanson, D., Norman, G., & Linn, R. (1995). Performance-based assessment: Lessons from the health professions. Educational Researcher, 24(5), 5-11; 35.

Wiggins, G. P. (1993). Assessing student performance: Exploring the purpose and limits of testing. San Francisco: Jossey-Bass Publishers.

Wiggins, G. P. (1994). Assessment reform. Princeton, NJ: Center on Learning, Assessment, and School Structure.

Wiley, D. E. & Haertel, E. H. (1996). Extended assessment tasks: Purposes, definitions, scoring, and accuracy. In Kane, M. B. & Mitchell, R. (Eds.). Implementing performance assessment: Promises, problems, and challenges. Pp. 61-89. Mahwah, NJ: Lawrence Erlbaum.

Wise, A. E. (Spring 2000). Performance-based accreditation: Reform in action. Quality Teaching, 9(2), pp. 1-2.

Wise, A. E. & Gollnick, D. M. (Spring 2000). Performance-based accreditation for the new millennium. NCATE Newsbriefs.

Yancey, K. B. & Weiser, I. (1997). (Eds.). Situating portfolios: Four perspectives. Logan, UT: Utah University Press.