CAATS -- Comprehensive Assessments Aligned with Teacher Standards –
A Five Step Design Model for Assessing Teachers Validly and Reliably

Paper Presented at the Annual Meeting of the
American Association of Colleges of Teacher Education
February 2005

Judy R. Wilkerson, Ph.D., wilkerso@tempest.coedu.usf.edu
William Steve Lang, Ph.D., wslang@tempest.coedu.usf.edu
University of South Florida St. Petersburg

Statement of the Problem

NCATE (2002) requires the measurement of knowledge, skills, and dispositions as part of its accreditation requirements for teacher education programs (Standard 1) and the use of unit assessment systems to aggregate and analyse data with a view toward program improvement (Standard 2). Data must indicate that candidates meet professional, state, and institutional standards. Institutions nationally are struggling with meeting these two standards.

In this presentation, we will provide ten recommendations then describe a five-step design model for developing a standards-driven, task based assessment system that can yield valid, reliable and fair decisions about teacher candidates' knowledge, skills, and dispositions on all three sets of standards. This model is integrally linked to sound measurement theory and practice, most notably the *Standards of Educational and Psychological Testing* (APA, AERA, and NCME, 1999). The recommendations are drawn directly from those standards. This model has evolved over a four year period.

Critical Flaws Impeding Validity

There are three critical flaws in the typical assessment process, which make meeting psychometric requirements virtually impossible. These flaws result in a hodgepodge or haphazard collection of evidence assembled without use of design frameworks or blueprints (AERA, APA, NCME Section 3, 1999):

- Evidence is typically drawn from a collection of class assignments, designed based on course objectives to determine a course grade and then used as summative assessments of standards to which they may be partially aligned. Thus, they are being used for a purpose for which they were neither designed not intended.

- The collection of artefacts, self-selected by the student or chosen solely on the basis of a tangential relationship to a standard rather than a predetermined alignment with all important aspects of a standard, typically fail to stand the test of construct representativeness (or domain sampling). Sampling (through a test blueprint) was not the design starting point.

- Decisions made about teacher competency based on a self-assessment through reflection do not stand the test of job-relatedness. In service teachers rarely analyse

their work against teaching standards once they are in the classroom full-time, and few schools require reflections to be turned in and reviewed by building administrators.

So, from the outset, some fundamental tenets of establishing validity are jeopardized in the assessment design processes used by most teacher preparation institutions. With an "I'll think about it tomorrow" attitude, institutions typically skip past validity, focus on inter-rater reliability, and conclude that they are consistently rating portfolios, forgetting that without validity, statistics about reliability are meaningless (Cureton, 1950; Wilkerson and Lang, 2003b). Furthermore, the correlations are artificially high and report a false positive, since few faculty members have the energy to adequately assess portfolios and even fewer are ready to fail a teacher candidate at a late date in their program. As Wright and Stone (2004) remind us, without difference there cannot be a valid measure of sameness. The ceiling effect, combined with a lack of adequate and appropriate evidence of job performance, leave the institution in a quandary.

More on Validity: Conflicting Paradigms and Purpose

The validity problem in teacher assessment begins with a common confusion about assessment purpose. We touched on that in the first flaw we noted in the previous section, but that is just the tip of the iceberg. Colleges of education need to respond to accreditation and approval requirements that are based on different purposes, and these purposes often remain undifferentiated. So the real beginning of an assessment system needs to be an understanding of purpose. NCATE accredits units, looking for evidence of overall program quality. That is their purpose. States approve programs, looking for evidence that individual teachers are minimally competent. Their purpose is to credential teachers through licensure or certification. NCATE conceptual frameworks focus on the unique aspects of graduates of an accredited program; state expectations focus on the consistency of graduate qualifications. While both types of agencies review results for teachers on the same or similar sets of teaching standards, they look at them through a different lens because their purposes are different (Wilkerson and Lang, 2004).

Despite these differences in purpose, institutions often attempt to meet both sets of requirements with the same data housed in the same containers, typically in a portfolio (often electronic) of student-selected work. The conflicting paradigms of ensuring minimal competence (protecting the public from unqualified practitioners) from the state perspective and preparing unique practitioners from the NCATE perspectives create a potential validity conflict. This lack of clarity about purpose or multiple purposes also often results in dissonance when faculty are trying to author a conceptual framework prior to an institutional review (Wilkerson and Lang, 2004).

Barrett (2004) describes the conflict of paradigms rooted in these two different purposes with two different needed products often rolled unsuccessfully into one – the assessment management system and the reflective portfolio. Difficulties in data aggregation result; and weaknesses in NCATE Standard 2 (NCATE, 2001) are then cited. The tension created by this conflict cannot be resolved until institutions recognize the need for different approaches based on different purposes. While there certainly will be overlap in the data collected for these purposes and approaches, there may also be differences. Successful assessment must simultaneously serve two or more different masters.

Recommendations to Help Achieve Validity

We have previously provided a series of recommendations with regard to the use of portfolios or other assessments in certification and licensure decisions (Wilkerson and Lang, 2003b). Two of those suggestions bear heavily on developing the assessments described in this article:

> Recommendation #1: The knowledge and skills to be demonstrated in the assessment must be essential in nature. They must represent important work behaviours that are job-related and be authentic representations of what teachers do in the real world of work.
>
> Recommendation #2: The entire assessment system must meet the criteria of representativeness, relevance, and proportionality.

The standards-based recommendations establish some points to consider in the planning stages of an assessment system

The following is a list of ten new recommendations now being proposed that have been culled from the *Standards* for all assessment systems. The recommendations establish some points to consider in the planning stages of an assessment system.

1. ***Identify the construct to be measured.*** In this case, the *Standards* provide an example of a construct as "performance as a computer technician." This can easily be converted for teacher educators as "performance as a teacher." (Chapter 1, p. 9, Validity)

2. ***Define the purpose***. Chapter 14 describes the requirements for credentialing. If the teacher preparation unit or the school district is advising the state on whether or not to license or certify, then this chapter applies. The *Standards* clarify that credentialing decisions are valid when they protect the public from unqualified practitioners, which then becomes the purpose. (Chapter 1, Validity)

3. ***Determine the use.*** Institutions need to decide if they will deny graduation to a teacher candidate based on the results of the assessment. Some states require this use; others do not require such a high stakes decision. Districts need to determine if they will fire a teacher based on the results of the assessment. In Florida, this is required. (Chapter 1, Validity)

4. ***Identify the measurable conceptual framework.*** Both NCATE and the *Standards* refer to observation of knowledge, skills, and dispositions when discussing a conceptual framework, so the framework can be all the teacher standards that define competency in these three categories. (Chapter 1, Validity)

5. ***Develop a blueprint or framework to guide the design process.*** Chapter 3 clarifies the need to build an assessment system, like any test, based on the domains to be measured – the conceptual framework. This is the reverse of what most teacher preparation institutions do. They start with what they have and hope it fits. (Chapter 3, Test Development and Revision)

6. ***Keep checking validity – both construct and content.*** Ensure that the system that is being built measures teacher performance, through job-related tasks (construct validity). Also show evidence that the set of assessments adequately represent the most important elements of the domains to be measured – with not too much and not

too little and nothing irrelevant targeted for any given standard (content validity). (Chapter 1, Validity)

7. ***Build assessments that can be studied for internal consistency.*** Rater agreement is important, but so are other sources of measurement error. A common scale on various tasks may help provide an adequate number of "items" to check for reliability. (Chapter 2, Reliability)

8. ***Develop systems to ensure fairness toward all those candidates assessed.*** This includes the policies and procedures to implement and monitor the system as well as specified checks for bias in the way tasks are written and differential results for protected populations. (Chapters 7-10 on Fairness in Testing)

9. ***Check the consequences of the decisions.*** Show evidence that (1) remediation attempts are appropriate and (2) the decisions made reduce to a minimum the number of poor teachers being certified ("false positives") and the number of good teachers being excluded ("false negatives"). (Chapter 1, Validity)

10. ***Build it once, and revise it***. Many institutions attempt to build parallel systems for each individual set of the many sets of standards. Align the standards from the beginning, and develop a single system to measure all of the standards. The system may have branches or tracks to fit multiple purposes, but all standards and all purposes should be considered at one time. Then revise based on experience, changes in institutional mission and standards, and problems identified related to validity, reliability, and fairness. (Chapter 3, Test Development and Revision)

The Need for Performance-Based Tasks

Much has been written about the shortcomings of licensure tests in sorting the qualified from the unqualified teacher (Pascoe and Halpin, 2001; Zirkel, 2000) and the need for including performance tasks with licensure tests to measure teacher competence (Lee and Owens, 2001; Rebell, 1991; Mehrens, 1991; Nweke & Noland, 1996). The National Commission on Teaching and America's Future (1996) makes it clear that continuous assessment is a major component of accountability and improvement, noting that "documentation efforts should include the extent to which graduates have developed and mastered the qualities of a highly qualified teacher" (p. 22). They recommended that licensure be based, not just on a single test, but also on demonstrated performance in the teaching skills that reflect the core competencies of a highly qualified beginning teacher.

In a study commissioned by the National Research Council (2001), the researchers concluded that even a set of well-designed licensure tests is inadequate to measure all of the prerequisites for a competent beginning teacher. Among other things, they recommended that licensure tests should be used only as part of an assessment system of of teacher competence. Similarly, researchers from the Southeast Center for Teaching Quality (2003) concluded that assessment systems need to use multiple methods, including student work samples and the demonstration of new knowledge and skills known to increase achievement. Hawley (1985) noted that tasks such as these may prove to be more reliable and valid for identifying and rewarding accomplished teachers.

Darling-Hammond, et al. (2002) also supported the use of a task-based system of teaching and assessing in their analysis of teacher education programs and pathways to certification. In that study, the authors identified some of the core tasks of teaching, such as

the ability to make subject matter knowledge accessible to students, to plan instruction, to meet the needs of diverse learners and to construct a positive learning environment. They concluded that many teachers do not feel that their programs adequately prepared them for certain teaching tasks.

The CAATS Model: Competency Assessments Aligned with Teacher Standards

The Competency Assessments Aligned with Teacher Standards (CAATS) model was designed to address the need for valid assessment systems comprised of standards-based, job-related (authentic) tasks to determine basic teacher competency. It consists of the following five steps:

Step 1: Define content, purpose, use, and other contextual factors.

A. Define the purpose(s) of the system.
B. Define the content of the system.
C. Define the use(s) of the system.
D. Review local factors that impact the system.

Step 2: Develop a valid sampling plan.

A. Organize standards into assessment domains.
B. Visualize the minimally competent teacher.
C. Brainstorm summative tasks.
D. Sort out formative tasks from summative tasks.
E. Build assessment frameworks.

Step 3: Create or update tasks aligned with standards and consistent with the sampling plan.

A. Determine the task format for data aggregation.
B. Create new tasks or modify existing tasks.
C. Conduct first validity study.
D. Set the standards for minimal competency.
E. Align tasks with instruction.

Step 4: Design and implement data tracking and management systems.

A. Select and develop a tracking system.
B. Develop a management system.

Step 5: Ensure psychometric integrity

A. Create a plan to test for validity, reliability, and fairness on a regular basis.
B. Implement the plan.

Application of the CAATS Model

We have successfully used the CAATS Model to design an assessment system for the Florida Alternative Certification Program (FACP), which is described in detail in the literature (Wilkerson and Lang, 2004). The FACP assessment system is comprised of 42 job-

related tasks that have been adopted by about 45 of the 68 Florida school districts. Variations of it are being used by several teacher preparation programs, as well. A list of the tasks is included in Appendix A.

We are now combining the tasks into thematic portfolios that can serve as four of the six to eight pieces of required evidence for review by the Specialty Professional Associations as part of NCATE accreditation. The four mini-portfolios are described in the following sections. Reference numbers are to the tasks in Appendix A.

Thematic Portfolio #1: Planning for Instruction and Assessment

In this thematic portfolio, teachers demonstrate their ability to align curriculum, instruction, and assessment, providing evidence of many aspects of planning. They identify objectives at multiple learning levels, incorporate specific interdisciplinary targets, and use a variety of instructional and assessment strategies.

The development process for this portfolio begins with planning for the entire grading period and includes more specific planning of a unit within the grading period (Task 10A) and the assessment system that will be used with it (1C). The teacher also manages instruction and assessment using technology (12B).

The complete portfolio will include lessons in which the teacher plans to teach critical and creative thinking skills through questioning and the use of higher order thinking objectives (4A/4B). There are also lessons that use cooperative learning (9B) and strategies based on specific theories of learning and development (7A/7B). Teachers are encouraged to incorporate these lessons into a single interdisciplinary unit (8A/10A), demonstrating their ability to integrate both literacy and mathematics skills into instruction (8C/8D). The teacher also develops and uses two assessments – one traditional and one alternative (1A and 1B) – ideally as part of this unit. The summative product in this portfolio is a record and analysis of the results of planning (10B).

For this portfolio, teachers may complete a single comprehensive and well-planned unit that includes most of the required tasks, or they may provide discrete evidence from multiple units, depending on their specific needs, curriculum, creativity, and interests. An example of task combining would be a science unit that includes research on the Internet. If the students work in cooperative groups and prepare a research report that incorporates graphs of data on their topic, this unit could meet the requirements of six separate tasks: 8A (interdisciplinary unit), 8D (math integration), 9B (cooperative learning), 4B (Critical Thinking), 12A (computer-enhanced instruction), and 1B (alternative assessment).

Thematic Portfolio #2: Interacting with Stakeholders

In this thematic portfolio, teachers will demonstrate their ability to work with children and their parents, individually and in groups, verbally and in writing. The tasks in this portfolio are mostly observational in nature; hence the portfolio is a predominantly a record of the observation results but also includes materials prepared in advance for the observations. The observations include general interactions between the teacher and students in the classroom (2C), focused observations with regard to diverse learners (5D), classroom management (9D), and interactions with parents and students in a conferencing context. In addition to the observations, teachers prepare a folder of written communication (2A) and a video-tape of their performance evaluated for professional behaviours (2B).

Thematic Portfolio #3: Supporting Learning in a Positive Environment

The third thematic portfolio provides evidence of the ability of the teacher to help all children learn both through documentation of students' progress and through the creation of an environment that supports their growth as individuals and collectively. As in the first portfolio, this portfolio contains many sub-parts or tasks, including planning and then working with diverse learners with special needs (5A-5C), understanding and improving students motivation and attitudes toward learning (7C and 9C), developing a classroom management plan that supports learning (9A), identifying an individual child who needs assistance and working with the child and (and parents and colleagues) demonstrating that student's growth (1D and 11D). The teacher also demonstrates a positive impact on student learning (1E) in a major unit (which could be the one planned in the first portfolio) by analyzing the results of multiple assessment (possibly including the two prepared for the first portfolio). The culminating work in this portfolio is a portfolio of K-12 Learning in which the teacher provides evidence of learning both content and critical thinking skills based on the products of his/her students (4C/8B).

Thematic Portfolio #4: Becoming a Professional

In this final thematic portfolio, the teacher explores some issues important to professional behaviours and attitudes, begins collecting resource materials, and initiates plans for continuing improvement. Specifically, the teacher examines ethical issues and the consequences of infractions(6A-6C), responsibilities to children experiencing personal crises (11C), and responsibilities for school improvement and work with parents (3B and 11A). This portfolio culminates with a professional development plan (3A) based on the results of all prior assessments.

Conclusions

Designing comprehensive assessment systems that provide opportunities for teachers to demonstrate the full breadth of their skills, as defined by national and state standards, is an important way to protect the public from unqualified practitioners and to ensure that all children have access to qualified teachers. This design model, and the tasks and mini-portfolios created using it, help achieve those goals. Quality decisions are based on quality data. This model helps assure quality in teacher training and assessment and addresses the most common and difficult problems of establishing validity evidence for teacher education.

Appendix A

FEAP #1 and INTASC #8: Assessment

- 01A: Unit Exam/ Semester Final Assessment
- 01B: Alternative Assessment
- 01C: Classroom Assessment System
- 01D: Case Study of a Student Needing Assistance
- 01E: Demonstration of Positive Student Outcomes

FEAP #2 and INTASC #6: Communication

- 02A: Written Communication from the Teacher
- 02B: Evaluation of Video-Taped Teaching
- 02C: Interaction between Teacher and Students

FEAP #3 and INTASC #9: Continuous Improvement

- 03A: Professional Development Plan
- 03B: School Improvement Team Involvement

FEAP #4 and INTASC #4: Critical Thinking

- 04A: Questioning Using a Taxonomy
- 04B: Lesson(s) to Teach Critical and Creative Thinking
- 04C: Portfolio of K-12 Student Work
- 04D: Critical Thinking Strategies and Materials File

FEAP #5 and INTASC #3: Diversity

- 05A: A Demographic Study of Your Students and a Plan to Meet Their Needs
- 05B: Documentation of Diversity Accommodations
- 05C: Individual Planning for Intervention
- 05D: Observation for Diversity

FEAP #6 and INTASC #9: Ethics

- 06A: Analysis of Slippery Situations
- 06B: Multiple Jeopardies and Infraction Penalties
- 06C: Potential Infractions and Teacher Responses

FEAP #7 and INTASC #2: Human Development and Learning

- 07A: Assessing Developmental Characteristics
- 07B: Assessing Learning Modalities
- 07C: Student Attitudes about School Learning

FEAP #8 and INTASC #1: Knowledge of Subject Matter

- 08A: Interdisciplinary Unit
- 08B: Portfolio of K-12 Student Work (cont.)
- 08C: Integrating Literacy Skills in Instruction
- 08D: Integrating Mathematics Skills in Instruction

FEAP #9 and INTASC #5: Learning Environment

- 09A: Classroom Management System
- 09B: Cooperative Learning Activity
- 09C: Case Study on Classroom Management and Motivation
- 09D: A Productive Classroom Environment

FEAP #10 and INTASC #7: Planning

- 10A: Semester/Year Curriculum Plan and Individual Unit Plan
- 10B: Semester Planning Record and Analysis
- 10C: Comprehensive Resource File

FEAP #11 and INTASC #10: Role of the Teacher

- 11A: Open House and Other Professional Involvement Plan
- 11B: Parent/Teacher/Student Conference
- 11C: Kids in Crisis
- 11D: Case Study of a Student Needing Assistance (cont.)

FEAP #12: Technology

- 12A: Computer-Enhanced Instructional Delivery
- 12B: Computer-Enhanced Management of Instruction
- 12C: Resource Materials from the Web

References

American Educational Research Association, American Psychological Association, and National Council of Measurement in Education (1999). *Standards for educational and psychological testing.*

Barrett, H. (2004, March). Differentiating electronic portfolio systems and online assessment management systems. Paper presented at the Annual Meeting of the Society for International Technology in Education (SITE), Atlanta, Georgia.

Council of Chief State School Officers. (1998). *Key state education policies in K-12 education: Standards, graduation, assessment, teacher licensure, time, and attendance: A 50-state report.* Washington, D.C.: Author.

Cureton, E. E. (1950). Validity, Reliability, and Baloney, *Educational and Psychological Measurement, 10*, 94-96.

Darling-Hammond, L., Chung, R., & Frelow, F. (2002). Variation in teacher preparation: How well do different pathways prepare teachers to teach?. *Journal of Teacher Education. 53:4*, 286-302.

Hawley, W.D. (1985). Designing and implementing performance-based career ladder plans. *Educational Leadership. 43:3*, 57-61.

Ingersoll, G. M. & Scannell, D. P. (2002). *Performance-based teacher certification: creating a comprehensive unit assessment system.* Fulcrum, Golden, CO.

Lee, W.W. & Owens, D. L. (2001). Court Rulings Favor Performance Measures. *Performance Improvement, 40:4, 35-40.*

Mehrens, W. A. (1991). Using Performance for Accountability Purposes: Some Problems. *ERIC Document Reproduction Service, ED333008.*

National Commission on Teaching and America's Future (2003). No Dream Denied: A Pledge to America's Children. New York: Author.

National Research Council (U.S.), Committee on Assessment and Teacher Quality (2001). *Testing teacher candidates: The role of licensure tests in improving teaching quality*. Committee on Assessment and Teacher Quality, Center for Education, Board on Testing and Assessment, Division on Behavioral and Social Sciences and Education, National Research Council, Mitchel, K.J., Robinson, D.Z., Plake, B.S., Knowles, K.T., editors. Washington, DC: National Academy Press.

Nweke, W. & Noland, J. (1996). *Diversity in Teacher Assessment: What's Working, What's Not?* Paper presented at the Annual Meeting of the American Association of Colleges for Teacher Education, Chicago, Ill. *ERIC Document Reproduction Service, ED393828.*

Pascoe, D. & Halpin, G. (2001). Legal issues to be considered when testing teachers for initial licensing. Paper presented at the Annual Meeting of the Mid-South Educational Research Association, Little Rock, AK. *ERIC Document Reproduction Service, ED460162.*

Rebell, M.A. (1991). Teacher Performance Assessment: The Changing State of the Law. *Journal of Personnel Evaluation in Education, 5*, 227-235.

Southeast Center for Teaching Quality (2003). Performance-based teacher compensation: Learning from the lessons of history. *Best Practices and Policies. 3:1*, May. Chapel Hill: Author.

Wilkerson, J.R. & Lang, W.S. (2004). A Standards-driven, task-based assessment approach for teacher licensure or certification with potential for college accreditation. *Practical Assessment, Research, and Evaluation*, 9(12).

Wilkerson, J.R. & Lang, W.S. (2003a). Florida Alternative Certification Program ssessment System: *Analysis of District Coordinators' Validity Questionnaire on Assessment Tasks.* Report: Bureau of Teacher Certification, Florida Department of Education, Tallahassee, FL.

Wilkerson, J.R., & Lang, W.S. (2003b, December 3). Portfolios, the Pied Piper of teacher certification assessments: Legal and psychometric issues. *Education Policy Analysis Archives, 11:45*. Retrieved December 20, 2003 from http://epaa.asu.edu/epaa/v11n45/.

Wilkerson, J., Lang, W.S., Hewitt, M., Egley, R., & Stoddard, K. (2002). *Florida Alternative Certification Program Assessment System.* Bureau of Teacher Certification, Florida Department of Education, Tallahassee, FL.

Wright, B.D., & Stone, M.H. (2004). *Making Measures*. Chicago, The Phaneron Press.

Zirkel, P. (2000). Tests on Trial. *Phi Delta Kappan. 8: 10*, 793-794.