Practical Assessment, Research & Evaluation

A peer-reviewed electronic journal.

Volume 9, Number 12, June, 2004 ISSN=1531-7714

A Standards-Driven, Task-Based Assessment Approach for Teacher Credentialing with Potential for College Accreditation

Judy R. Wilkerson & William Steve Lang
University of South Florida St. Petersburg

When the Southeast Center for Teaching Quality (2003a) studied what teachers would want to tell policy makers about highly qualified teacher requirements, they summarized their findings in this quote: "Come to my classroom, and not just for a day." (p. 2). What could be a more obvious invitation to improve and expand teacher assessment? The standardized objective tests and occasional formal observational evaluations being used to measure teacher competence today have been contested for decades as ineffective according to politicians and invalid according to researchers. While these two long-standing and useful assessment strategies are important components of an overall assessment system, neither – alone or paired – is sufficient to identify and remediate new teacher deficiencies. This article includes a series of recommendations, organized in steps, for developing an assessment approach that is task-based, standards-driven, and job-related that would serve as a major component of a comprehensive beginning teacher assessment system. These recommendations are based on a two-year effort that resulted in Florida's Alternative Certification Program Assessment System. This system has now been adopted by about two-thirds of the Florida school districts and is beginning to be adopted by colleges of education preparing teachers through the traditional route. The design takes into account three sets of standards: the Florida requirements for program approval the NCATE requirements for accreditation, and the *Standards of Educational and Psychological Testing* (referred to as the *Standards;* AERA, APA, and NCME, 1999). These standards define the ultimate purpose of decisions about initial teacher competence: protecting the public from unqualified practitioners.

Context

In the starkest terms, the failures of policies and practices, whether in federal or state government, in university preparation programs, or in school districts, are being shouldered by children. This is unconscionable… It is unacceptable, as a matter of public policy, to hold students to academic standards that some of their teachers are unable to help them meet. It is time for full public disclosure. States and school districts should ensure that every teacher in every classroom has met teaching standards aligned with K-12 learning standards. The Commission believes it is time to make accountability for results a reality for everyone involved. The chain of accountability should include states, teacher preparation programs, and school districts. They all should be held responsible for enforcing high standards for all entrants to teaching coming from all forms of teacher preparation. All links in the chain should deny teaching appointments to unlicensed and unqualified individuals. -- National Commission on Teaching and America's Future, January, 2003

This harsh, but continuing, condemnation of teaching and teacher assessment provides the backdrop for the teacher assessment design model proposed in this article. The assumptions from the Commission Report that will serve as a framework for this article are: (1) the perceived failures of college and district-based teacher preparation to prepare teachers who have demonstrated the knowledge and skills set forth in national and state standards for both teachers and children and (2) the need to ensure that teachers have met those standards.

Perceptions often come from a mix of truth and fiction, and the "failures" in schooling have become a highly politicized topic. Clearly, all is not broken, but, equally clearly, all is not perfect. Our accreditation agency, the National Council for Teacher Education (NCATE) and all of the Specialty Professional Associations (SPA's) in the content areas have recognized the need for focused, targeted performance assessments that provide data for standards-driven continuous improvement to address this problem.

In this paper, we attempt to tackle the issue of meeting standards because standards can help us improve as a profession, if we approach the assessment process in a systematic way so that we can objectively identify the areas we need to target for improvement -- teacher by teacher, program by program, college by college, district by district, state by state. The hypothesis under girding this research is that the problems in teacher assessment are largely a function of designing systems without adequate evidence or operational understanding of standards. So, these two elements of the backdrop are inextricably linked.

Essentially, decisions about entrance into the profession have to be made based on a demonstration of standards, and these decisions need to be valid. As measurement professionals, we believe that the greatest challenge to those assigned the responsibility of ensuring teacher competence is to design systems with a constant eye on validity. As we mentally process the NCTAF challenge, the conclusion is clear: nationwide, we need to ensure that the right decisions about teacher credentialing are being made consistently.

There are three sets of standards that have influenced the approach outlined in this article. As is typical nationwide, both national and state requirements for accreditation and program approval form the basis for defining teacher competency (knowledge, skills, and dispositions) and then creating the assessment systems required to measure competency. Added to these typically used standards are the ones that these authors believe to be equally critical in assessment design – the *Standards of Educational and Psychological Testing* (AERA, APA, and NCME, 1999). A merger of these three sets of standards establishes a stronger opportunity to measure teacher competence with confidence, precision, and accuracy.

As long-term consultants working with the Florida Department of Education and the colleges of teacher education in this state and elsewhere, we have found that the teacher education community, whether it be college or district-based, is driven largely by value and belief systems that are personal in nature and are only tangentially relevant to the basic teacher certification decision. System designers are largely unaware of the AERA, APA, NCME *Standards* and are typically unconcerned with, and uniformed about, the critical importance of validity. The tendency, then, is to develop assessment systems without the aid of measurement professionals and without regard to these *Standards*.

The hodgepodge or haphazard evidence collected without use of design frameworks or blueprints (AERA, APA, NCME Section 3, 1999) in the name of NCATE standards and/or state laws and rules makes meeting psychometric requirements virtually impossible. There are three critical flaws in the typical assessment process. Each poses a major threat to validity:

- Evidence drawn from a collection of class assignments are used for a purpose for which they were not intended; i.e., summative assessment of standards. Course assignments and tests are typically designed to meet individual course objectives to derive a course grade as a record of participation and achievement for a transcript, not as a measure of meeting a standard.
- Collections of artifacts, self-selected by the student or chosen solely on the basis of a tangential relationship to a standard, or reflective/showcase portfolios of candidates' best work, usually fail to stand the test of construct representativeness (or domain sampling), since 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 analyze 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 outlined in the *Standards* are violated.

While states and NCATE say that they require evidence of credibility or psychometric integrity, this is an issue that college faculty and administrators, like Scarlet O'Hara, say they will think about tomorrow. If they do acknowledge the need to ensure any aspect of psychometrics, it tends to be a study of inter-rater agreement. This agreement study is often based on a weakly designed rubric for a portfolio that virtually every student successfully completes because few faculty members have the energy to adequately assess it (Wilkerson and Lang, 2003b). As Wright and Stone (2004) remind us, without difference there cannot be a valid measure of sameness.

Conflicting Paradigms and Purpose

In the previous section we briefly stated some of the major threats to validity. Reliability, too, is an issue, but it is not addressed in depth in this article, although we note briefly that reliability without validity is meaningless (Cureton, 1950). One could ask how assessment systems have gotten so far off track. The answer appears rather simple.

The validity problem in teacher assessment begins with a common confusion about assessment purpose. Colleges of education need to respond to accreditation and approval requirements that are based on different purposes, and these purposes often remain undifferentiated. As consultants, we see this almost daily in Florida. Deans and faculty talk about "NCATE" and mean both "NCATE and DOE." There are NCATE Coordinators, not NCATE/DOE Coordinators, here and elsewhere. So the real beginning of this article, like 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.

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.

College faculties do not want to think about minimal competence, because, by virtue or human nature, they want to believe they are preparing teachers better than anyone else, especially districts using alternative routes. The NCATE conceptual framework and years of defining vision and purpose for other institutional and accreditation purposes feed this value. Thinking about graduation requirements focused on minimal qualifications and protection of the public is, to put it bluntly, distasteful. However, the AERA, APA, & NCME Standards make it clear that in credentialing decisions, job-related competency is precisely the role that assessment needs to play. Perhaps the crisis cited by NCTAF and others who worry about the failures of American schooling is a direct result of failing to focus on the measurement of minimum competence, as defined by the agreed upon details of good teaching.

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.

The remainder of this article will be focused predominantly on the more stringent assessment purpose: credentialing for the State. Clearly, much of what is done to achieve state program approval will more than satisfy NCATE teams looking for assessment systems. It is proposed herein that if teacher educators do a solid job of designing for certification; it will require little additional effort to add a component to the assessment system that meets NCATE expectations.

After a brief review of the literature on testing and credentialing, we will describe a model we used in creating a standards-driven, task-based assessment system here in Florida, and we will conclude with a brief description of the system and some initial evidence regarding validity.

Literature on Testing and Licensure/Certification: A Need for Performance 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. They concluded:

Most states test prospective teachers, but many are still not using true performance-based assessments that provide valid measures of teaching competence. In short, teacher licensure tests simply don't measure up; many are weak indices of the depth of knowledge and skills all teachers must have. States also differ substantially in how they set passing scores. States have raised teaching standards substantially in the past decade; now they need to improve the measures of teaching competence that make standards credible. (p. 23)

In a study commissioned by the National Research Council (2001), the role of licensure tests in improving teacher quality was examined. The researchers concluded that even a set of well-designed tests is inadequate to measure all of the prerequisites for a competent beginning teacher. These researchers recommended that states use multiple forms of evidence in making decisions about teacher candidates, including a comprehensive set of high quality indicators and concluded that:

- New and developing assessment systems warrant investigation for addressing the limits of current licensure tests and for improving the licensure decision-making process.
- Licensure tests should be used only as part of a coherent developmental system of preparation, assessment, and support that reflects the many features of teacher competence.
- Among the policies that may have the greatest potential to improve the quality of pre-service teacher education as well as alternative certification would be to require passing assessments of teaching performance that included both frequent and substantial evaluation.

Researchers from the Southeast Center for Teaching Quality (2003c) also concluded that the lesson is simple: 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, since successful career development models use assessments that teachers view as being fair and appropriate for frequent use.

The notion of a task-based system of teaching and assessing was further supported by Darling-Hammond, et al. (2002) in an 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. These authors make the link between preparedness, core tasks, and retention.

The approach described here is based on the use of a series of job-related, standards-driven performance tasks that are either product or behaviorally based. These tasks can be combined with other teacher competency measures, such as teacher certification exams, supervisor observations, evidence of completion of training requirements (courses and/or workshops), and showcase or reflective portfolios, but the anchored core of the system would be observable task performance.

Performance Assessment Tasks As An Operational Definition of Standards

In a world driven by accountability and standards, decision-makers are faced with a great challenge when they attempt to measure teacher competence with regard to the teaching standards they are required to apply. The standards answer the fundamental question, "What does an effective teacher need to know, be able to do, and believe?" Users then need to create a comprehensive set of operational definitions of the standards which address the knowledge, skills, and dispositions required for competent teaching. To create such a system designers must answer the next fundamental question: "What does each standard ***look like*** in practice, when performed by a good teacher?", or "How can we ***see*** that the teacher has acquired the knowledge, skills, and dispositions called for in the standards at some minimal level of competence?"

Often, this is a matter of common sense. If a standard requires that the teacher be able to manage a classroom, the evaluator most likely will want to ***see*** a classroom management plan and ***watch*** the teacher working with students in the classroom. This is a relatively simple example; others may be more complex, such as maintaining integrity and acting ethically, working collaboratively, reflecting and improving, involving students in decision making, or using technology. In each case, the assessor needs to ***visualize*** a performance that provides evidence of competency. This means that the elements of the standards need to be operationally defined in terms of a job-related performance, whether it be something written or something performed. The term "tasks" is used in this article to refer to these job-related, performances that are either product or behaviorally based evidence and can be scored.

The dilemma of answering the "what does it look like?" question is complicated by the sheer numbers of standards that have been written. Standards are written and applied at both the state and national levels, and even sometimes at the district levels. Chief among them is the set of standards developed by the Council of Chief State School Officers (CCSSO, 1998) through its committee, the Interstate New Teacher Assessment and Support Consortium (INTASC). These standards are called the INTASC Principles and form the basis for many other standards written by states and Specialty Professional Associations affiliated with the National Council for Accreditation of Teacher Education (NCATE). Some of the SPAs include the National Councils of Teachers of Mathematics or English (NCTM or NCTE), the Association for Childhood Education International (ACEI), and the Council for Exceptional Education (CEC).

Design Requirements for a Psychometrically Sound Assessment System

Standards and statutes toss out the terms "validity" and "reliability" and require teacher educators to think about them in some meaningful way. Then teacher educators, in their accreditation related reports, salt and pepper the text they write with the appropriate buzz words with more rhetoric than substance. But how does one get there, making use of the *Standards for Educational and Psychological Testing (1999)*?

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 behaviors 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 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)

Competency Assessment Aligned with Teacher Standards (CAATS) Model

The Competency Assessment Aligned with Teacher Standards (CAATS) model is described below. There are five developmental or procedural steps that can help the designers of teacher assessment systems progress toward the above ten recommendations. Some very brief examples of a few of the implementation strategies are provided for illustrative purposes in the subsequent section on Florida's system.

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

In this step, designers begin by determining what they want to know (assessment content), why they want to know it (assessment purpose), and what they will do with the information once they get it (assessment use). Each purpose and use are conceptualized and evaluated separately as a matter of validity.

- ***Purposes*** will vary based on need. Institutions may have dual or triple purposes – such as teacher certification, program improvement, and teacher reflection or growth. The first would be more aligned with state program approval and would tend to be rather prescriptive and analytic in nature. The second and third would be more aligned with the NCATE conceptual framework and could tolerate much more freedom of choice and more holistic decision-making. Thus, these are likely to require different data and data analysis strategies to fulfill the different purposes, although there may be substantial overlap.

- ***Content*** can be defined in many ways, but in the case of accreditation, it is reasonable to be consistent with NCATE Standard 1. This would require defining content in terms of the knowledge, skills, and dispositions required of teachers by various groups – national (professional), state, and institutional.

- ***Uses*** also dictate what will be done with the data. Some examples are: certify a teacher or allow him/her to graduate, identify certification-related program weaknesses and improve a program, identify weaknesses and make improvements in unit-defined areas of importance (e.g., conceptual framework).

Examples showing the relationship of purpose, content, and use are:

Table 1: Samples of Purpose, Use and Content

	Example 1	**Example 2**	**Example 3**
Purpose	The system ensures teacher competence in order to protect the public from unqualified practitioners.	The system provides data on program quality for program improvement.	The system provides opportunities for teachers to demonstrate competency in self-assessment and reflection.
Content	Specific tasks are required in the system for analysis of teacher competency. The content of each task varies but is aligned with institutional, state, and national professional standards, agreed to by the profession, showing depth and breadth of coverage.	Same as Example 1 plus any additional data targeted by the institution or district. Depth and breadth of coverage are important to the extent that they yield data on overall program quality that are useful for substantive improvement.	Portfolios of teacher-selected evidence with reflections on demonstration of standards and future plans for growth. Content varies for each teacher within a pre-defined institutional or district structure.
Use	Aggregated data from scoring rubrics on tasks for each teacher determine whether the teacher successfully completes the program, and, therefore, can be certified.	Data are aggregated from scoring rubrics on tasks for each standard to identify standards with which teachers have more difficulty than anticipated in order to target program improvements.	Teachers receive feedback on individual work, and scoring rubrics for portfolios are aggregated for the unit to evaluate teachers' ability to reflect and self-assess as a measure of curriculum strength.

Once the purpose(s), use(s), and content are determined, it is important to analyze all the local factors that would affect the system, e.g., conceptual framework, resources, faculty resistance/cooperation. It is also important to ensure that all aspects of the program to be measured, including the delineation of content and conceptual framework, be more motivated by faculty's ability to visualize what they look like when performed than any political or other motivations. For example, how is a "lifelong learner" visible unless one follows the graduate for a lifetime?

Step 2: Develop a valid sampling plan.

A critical next step is the identification of all relevant standards and the alignment of standards with each other into assessment domains. This is the beginning of the job analysis, and it is not as daunting a task as it sounds. There are common threads that run through all of the sets of standards because the literature has identified several important characteristics of effective teaching consistently over time. While there is much hair-splitting in the field about the details, essentially, the ten Principles written by INTASC identify those critical competencies.

For example, virtually every set of standards has something about content knowledge, planning, and assessment. This can easily be demonstrated through the alignment process, as is shown in Table 2 where these three aspects of good teaching are aligned for Florida's Accomplished Practices, INTASC Principles, Association of Childhood Education International (ACEI – elementary education) Guidelines, Council for Exceptional Children (CEC – special education) and the National Association for Education of Young Children (NAEYC – primary) Guidelines.

Table 2: Sample Alignment of Selected Standards

	Florida #	INTASC #	ACEI #	CEC #	NAEYC #
Content	8	1	2	1, 4	4
Planning	10	7	3	7	4
Assessment	1	8	4	8	3

When considered together, as a kind of content domain, one can clearly see the similarities and differences between and among the perceptions of what is important from each group of professionals. While this can add frustration to those assigned the responsibility of measuring competence, it also can add a depth that can be useful in articulating one's own professional commitments. It is really a matter of the viewpoint chosen: the glass is either half-empty or half full. To demonstrate the similarities, the precise standards language for Assessment (with some language in bold for emphasis) is provided below. The indicators developed by each standards-setting group provide more detail that is useful or annoying, again depending on one's point of view.

- Florida: The preprofessional teacher collects and **uses** data gathered from a variety of sources. These sources will include both **traditional and alternate assessment strategies**. Furthermore, the teacher can identify and match the student's instructional plan with their **cognitive, social, linguistic, cultural, emotional, and physical needs**.

- INTASC: The teacher understands and **uses formal and informal assessment** strategies to evaluate and ensure the continuous **intellectual, social and physical development** of the learner.

- ACEI: Candidates know, understand, and **use formal and informal assessment** strategies to plan, evaluate, and strengthen instruction that will promote continuous **intellectual, social, emotional, and physical development** of each elementary student.

- CEC: Assessment is integral to the decision-making and teaching of special educators and special educators **use multiple types of assessment information** for a variety of educational decisions. Special educators use the results of assessments to help identify exceptional learning needs and to develop and implement **individualized instructional** programs, as well as to adjust instruction in response to ongoing learning progress…

- NAEYC: Candidates **use multiple**, systematic observations, documentation, and other responsible **assessment strategies** as an integral part of their practice.

All of the standards could be correlated in a similar fashion, extending the matrix in both directions – vertically and horizontally. The point is that there is substantial overlap among standards sets. Aligning them allows for the creation of a solid set of assessment domains and reduces the overall workload in the end. Personnel working on assessment systems can tire easily if asked to recreate the assessments repeatedly, each time for a different master.

Some additional suggestions on establishing the job analysis follow:

- Develop a condensed working version of the standards – a set of institutionally relevant criteria --- that represent the "best of all worlds" in order to begin the process of defining the sub-constructs of teacher performance. This could be a major component of the NCATE-required conceptual framework.

- Visualize the competent teacher performing the critical skills identified, and brainstorm a list of tasks (products and performances) that can be used to measure the skills. This is the second component of the job analysis and answers that fundamental question raised earlier in this article:

"What does each standard ***look like*** in practice, when performed by a good teacher?"

or

"How can we ***see*** that the teacher has acquired the knowledge, skills, and dispositions called for in the standards at some minimal level of competence?"

- Construct a series of design frameworks that help formulate a balanced and appropriate sampling plan. The more comprehensive the planning, the more likely the sampling plan is to be well-developed. The frameworks (or blueprints or matrices) can have the standards or skills in one dimension and a number of options in the other to ensure the utility and feasibility of tasks from various perspectives. The second dimension could include one or more of the following:

  - Types of competency (knowledge, skills, dispositions, impact on K-12 learning)

  - Level of measurement inference (high, medium, low)

  - Timing (admission, pre-internship, internship, graduation, post-graduation)

  - Assessment method (tests, products, observed performances, interviews, scales, etc.).

In recommendation #5 above, the importance of starting with a blueprint was noted. Validity is attainable if the designers start in the right place – what they need and not what they have. Validity is very difficult, if designers work from a hodgepodge of assessment stuff.

- Sort formative from summative tasks. It is important to think through carefully which tasks should be used for formative purposes, and which ones should be used for summative decision-making. College faculties are often tempted to include everything they do, and this will lead to an unmanageably large system. Sampling is critically important to achieve the desired use and interpretation. The key question here is how much is enough – not too much and not too little to assure a representative and thorough sample of all critical skills without overdoing it.

If this were airline pilot school, one could make the following analogy: Landing the plane is important and needs to be measured. Finding the "bring me coffee" button (there probably is no such button) does not need to be assessed; the pilot will find it when he/she gets thirsty. Some pilot instructors may think it is important because previous pilots have complained, or it was their most used button when they were pilots, but it is not a critical skill.

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

If one is to view the series of tasks as necessary for a specific, unified decision – graduation or certification – then it is useful to agree upon a common format to make data aggregation across tasks easier. This format could technically be considered item specifications. A useful approach is to include the following:

- Standards alignment
- A brief description of the task for public consumption
- Comprehensive directions for task completion
- Detailed scoring rubrics or guides that include both criteria and a scale.

The more specific and clear the directions and rubrics, the less danger there is of some forms of construct irrelevant variance in performance.

It is helpful to have a common rating scale with an equal number of points on the scale (e.g., a three-point scale such as "target," "acceptable," and "unacceptable" or 3, 2, 1) for each task in the system, and each task should be evaluated for several criteria. Scoring in this analytic way helps to provide maximum feedback to teachers. If programs do not agree on the step values, a transformation to a common score format is often possible.

Once the tasks are written based on these specifications, several additional procedures need to be followed:

- Evidence of content validity should be gathered to ensure that the tasks are representative of the construct and its conceptual framework, proportional, and in fact job-related. Reviews of coverage of the standards, combined with surveys of (or discussions with) experts, can help ensure that the tasks are representative and job-related. This would constitute an important source of validity evidence.

- Standards for establishing minimal competency should be established – the point at which teachers will be denied graduation or certification. Methods of setting criterion decision points are widely documented (Impara, 2000, Rudner, 2001).

- Evidence that adequate instruction is provided for each task needs to be gathered, so that teachers have the maximum opportunity possible to complete successfully the tasks. This has been called instructional validity, although it is not a technical term in the S*tandards*.

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

The data must be accumulated and managed for decision making. At a minimum, it is helpful to agree on a rubric for decisions at the task level. Many approaches are possible for this. Assessors then need to share information about teachers to make informed summative decisions about progression.

Tracking systems, recordkeeping, and other procedural details, are necessary to do this. Technology can be of great help. Options include:

- Student or personnel folders
- Portfolios
- Databases

All are useful strategies, and one or more should be selected. Keeping data in grade books, electronic or otherwise, without any system for aggregation and sharing, does not help make summative decisions about teachers. Other important implementation issues are:

- Consensus around the system needs to be built and supported. Reward systems could include anything that makes faculty or other assessors more amenable to the accountability requirements.

- A maintenance program is necessary and should be created to include training of assessors, collection of scored examples showing different levels of proficiency, orientation of teachers being assessed, alternative strategies or tasks for teachers who need them, advising materials (including due process), and an appeals process.

- Formal review times to update and improve the tasks and the system should be established in advance. Identified people or committees responsible for data collection in a timely and regular fashion are also important for the valid implementation of the system.

Step 5: Ensure psychometric integrity.

There are increasing calls for ensuring the credibility of assessments, including validity, reliability, and fairness. Assessment designers should make use of the *Standards* (1999).

To summarize, the *Standards* require the use of blueprints; a focus on job-relatedness; and evidence of validity (particularly content validity), reliability, and fairness. Logical as well as empirical data should be gathered, and particular attention should be paid to the properties of portfolios if they contain student-selected works. Inter-rater reliability is important but not the sole criterion for reliability, as many other sources of error need to be considered.

The *Standards* recommend the development of a conceptual framework for ***assessment*,** and this framework should be developed concurrently with the NCATE conceptual framework in teacher preparation programs. The framework can be operationalized through a "psychometric plan." This plan, that will ensure the integrity of the system, should be developed in the beginning, so that faculty and district personnel can be comfortable that when the time comes, they will be able to provide the evidence necessary that their decisions about teachers are both truthful and trustworthy. The plan should include personnel responsible, timelines, and all the standard features of an action plan. To be confident about the psychometric integrity of the plan, at least one person who is competent in measurement should be among the personnel responsible for plan development and implementation. The plan should include the following elements:

- Purpose, context, and consequences of the system
- Construct measured (teacher job performance)
- Conceptual framework (knowledge, skills, and dispositions)
- Use, interpretation, and reporting of scores
- Test specifications, content map, and item pool
- Assessor/rating selection and training procedures
- External assessment system review personnel
- Assessment system analysis methodology (e.g., CTT or IRT)
- Methodology for gathering evidence of validity, reliability, and fairness

In terms of the last two bullets, assessment designers should determine whether they will use Classical Test Theory or Item Response Theory as their basic methodology and then frame questions and specific research strategies to address them. An example for validity, using IRT, is:

Evidence based on content:

Question #1: Does the assessment system provide adequate coverage of the Standards?

Judgmental Method: Charts showing representativeness and relevance of tasks per FEAP by indicator, created by test developers.

Empirical Method: Rasch analysis with logistic rulers of calibrated criteria (to show gaps, ceilings, or floors where unexpected).

Question #2: Are the tasks an adequate representation of the job? Is each task critical to job performance, authentic, and frequent?

> Judgmental Method: A survey of users, including university faculty and administrators, district coordinators and mentors, all of whom are considered to be expert judges: All tasks are evaluated for criticality, authenticity, and frequency.
>
> Empirical Method: Rasch analysis in which separation statistics are compared to measure range.

Applying the CAATS Model in Florida

Florida as a Typical State

For many years, Florida, like many other states, has relied predominantly on standardized tests and administrator observations to make decisions about teacher certification and employment. Here, the Florida Teacher Certification Examination and the Florida Performance Measurement System or FPMS (Florida Statutes, 1983) have long served as the test and the observational components of the assessment system. FPMS includes an instrument that requires observers to tally effective and ineffective behaviors.

In Florida, like in many other states, teacher preparation institutions and school districts are required to develop assessment systems based at least in part on a local set of standards. State standards are often modeled after the INTASC Principles, and Florida is among the states that have taken this path toward standards development. Here, the standards are called the Florida Educator Accomplished Practices or FEAPs (Florida Education Standards Commission, 1996). Florida also has its own set of content standards for teachers. So, teacher preparation institutions in Florida are held accountable for the FEAPs and the Florida content standards, as well as INTASC and SPA standards through NCATE accreditation. School districts are primarily concerned with the FEAPs. For both groups, adequate coverage of standards presents a challenge.

The Florida Alternative Certification Program (FACP) Assessment System

Overall Design: The Florida Alternative Certification Program (FACP) Assessment System (Wilkerson, et al., 2002) follows the purpose, contents, and use outlined in Example 1 in Table 1 of Step 1 of the CAATS model. Standards were analyzed, aligned, and synthesized into sub-sets of critical skills, consistent with the Florida Educator Accomplished Practices. The INTASC Principles were included in this synthesis. For the teacher preparation version of this system other national standards were incorporated into the alignment process.

The FACP system is the one described predominantly in this article. It includes a series of 42 tasks, which at the time of this writing, are being combined and sequenced into clusters. These clusters, or thematic portfolios, will be based on district feedback and recent revisions to the NCATE SPA review that now requires six to eight key assessments. The college version is somewhat longer, incorporating more standards important to traditional certification. There are more similarities than differences between the two systems, since the original 42 tasks serve as the core for the subsequent set of 60. The traditional certification system is larger because of the need to assess content, with over 25 tasks for the content standards, including both tests and performances in each content area.

There are between two and five tasks per FEAP, with the exception of the content standard. Each sub-set of tasks is based on a vision of minimal competence for the standard. A few are knowledge-based, rather than skill-based, because of the nature of the standard. In these cases, the lower levels of authenticity were found to be acceptable because of the criticality of the knowledge (e.g., theories of human development and learning).

Each task has been aligned with the locally defined critical skills, as well as the relevant FEAPs and INTASC Principles, at the indicator level, to ensure adequate representation of the domain. An example of a set of tasks and the coverage analysis (first validity study) follows:

Tasks for 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

Table 3: Sample FEAP Sample Key Indicators for Critical Thinking and Tasks

#	Criteria	Tasks
4.1	Provides opportunities for students to learn higher-order thinking skills.	4A, 4B, 4C, 4D

4.2	Identifies strategies, materials, and technologies which she/he will use to expand students' thinking abilities.	4A, 4B, 4C, 4D
4.3	Has strategies for utilizing discussions, group interactions, and writing to encourage student problem solving.	4A, 4B, 4C, 4D
4.4	Poses problems, dilemmas, and questions in lessons.	4A
4.5	Assists students in development and use of rules of evidence.	Not covered
4.6	Demonstrates and models the use of higher-order thinking abilities.	4A, 4B
4.7	Modifies and adapts lessons with increased attention to the learners' creative thinking abilities.	4B
4.8	Encourages students to develop open-ended projects and other activities that are creative and innovative.	4B, 4C
4.9	Uses technology and other appropriate tools in the learning environment.	4D

The complete versions of the Critical Thinking tasks are on-line at: http://pats.stpt.usf.edu/criticalthinking. These tasks are being modified, as are all tasks in the current system, based on initial experience with them. It is also expected that additional tasks will be added over time. The versions on-line are presented by way of example only to illustrate the format developed based on the item specifications.

The task specifications are the same as the ones outlined in the model. The complete list of the 42 tasks is also attached in Appendix A, again, by way of example – not as a finite set of tasks that will not change over time. This system is evolving based on data; the developmental model is the main focus of this article. As noted earlier, a clustered version of the tasks, to be organized in thematic portfolios, is also being developed.

Of the 42 original FACP tasks, there are 39 products and three observational assessments in the system. What needed to be assessed through short and long-term products is assessed through the written work of the teachers; what needed to be observed live is assessed through observation. Some tasks can be completed in a short period of time; some span a semester. Task by task, the method had to match the job function as defined in the conceptual framework. In the college-based version of this system, assessments are provided in courses where it is appropriate and in field-based experiences where it is necessary.

The products have tangible results, e.g., a record of accommodations, a lesson plan to teach critical thinking skills, a classroom management plan, a portfolio of K-12 student work, an assessment with mastery analyzed and follow-up suggested, a semester and unit plan, a folder of communications with parents, a case study showing results of improvement with one child. The three observations provide instruments to determine interactive communication with students, sensitivity to diverse populations, and ability to maintain a supportive learning environment.

There is a strong focus in the set of tasks on impact on K-12 learning, consistent with both NCATE accreditation requirements and the concerns expressed in the NCTAF report. Five tasks target impact specifically:

01C:	Classroom Assessment System
01D:	Case Study of a Student Needing Assistance
01E:	Demonstration of Positive Student Outcomes
04C:	Portfolio of K-12 Student Work
05B:	Documentation of Diversity Accommodations

Decisions: There is much work yet to be done on the system, including the establishment of cut-scores based on judgmental and empirical data. At this time the decision-making process requires data aggregation at multiple levels. Teachers have three opportunities to succeed on every task, and no unacceptable ratings are tolerated after three attempts. There is a heavy emphasis on assessment as learning in this system, since many of the assessments are conducted on-the-job (internship) for traditional preparation or all OJT in alternative certification (induction year). A graphic representation and table explaining the current structure follow:

Figure 1.

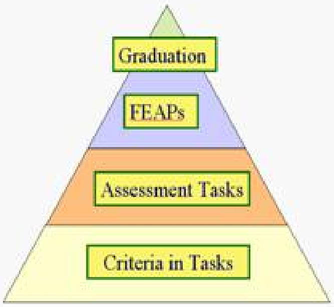


Table 4: Levels of Decisions in the FACP Assessment System

Criterion Decisions	**Task Decisions**	**Practice (FEAP) Decisions***
Target (T): The teacher's work meets all expectations.	**Demonstrated (D):** If all the criteria are rated at the "target" level or "acceptable" level, the teacher has demonstrated the Practice. **Up to** one-third of the criteria may be rated as just "acceptable". None of the criteria can be rated as "unacceptable."	**Accomplished**: The teacher has successfully demonstrated this Practice, consistently showing evidence of ability in the Practice. He/she is ready to assume full responsibility for teaching and continuing professional development with regard to this Practice. This level of achievement has been demonstrated by multiple ratings of "demonstrated" on the tasks in this system. The candidate has received not more than one rating of "partially demonstrated" and ***no*** ratings of "not demonstrated."
Acceptable (A): The teacher's work is essentially correct but has minor problems that can be addressed through counseling or advising. The assessor expects that the teacher will be able to self-correct without difficulty in subsequent attempts at similar tasks.	**Partially Demonstrated (PD):** If **more than** one-third of the criteria were rated as just "acceptable," the teacher is rated as having partially demonstrated the Standard. None of the criteria can be rated as "unacceptable."	**Competent**: The teacher has demonstrated this Practice adequately for certification. However, there are some gaps in performance that require continued monitoring, and there were two or more final decisions of "partially demonstrated" on the tasks for the Practice. These performance gaps need to be remediated over time through implementation of a professional development plan and should be monitored closely by the employing district
Unacceptable (U): The teacher's work has a major problem that needs to be corrected and re-checked. The problem is serious enough to question whether or not the teacher can be effective and should be certified.	**Not Demonstrated (ND):** One or more of the criteria were rated as "unacceptable." The teacher must fix whatever was unacceptable and earn a PD or D rating on the task to be certified.	**Not Competent**: The candidate has not demonstrated minimal competence on this Practice and, therefore, is not eligible for certification. The teacher has received a "not demonstrated" rating on one or more products or performances in the assessment system for this Practice, despite three opportunities provided for successful demonstration.

*Note: In Florida, all FEAPs must be demonstrated, hence, for the final decision on certification, all 12 FEAPs must be at the competent or minimally competent level.

Other Design Aspects Related to the CAATS Model as Applied in the FACP Assessment System

The tracking system is electronic, with a database system developed and supported in Tallahassee. Training of assessors is ongoing, and exemplar assessments are being collected and posted on the web. Training resources are also available on the web (http://www.altcertflorida.gov/).

Psychometric studies are on-going, since the data are being analyzed with the Rasch model of Item Response Theory.

Over time, a logistic ruler with all tasks and all criteria scaled will be calibrated. Initial results indicate that the model is working; however it is premature to present those results at this time since the system has recently been implemented and there is inadequate connectivity to report meaningful statistics. Using the Rasch analysis, however, the potential to identify both misfitting items and misfitting persons in the future is promising for both improving the tasks and diagnosing teacher weaknesses.

Initial Evidence of Validity of the FACP Assessment System

Content Validity Evidence Based on Alignment with Standards

By using the FEAP and INTASC principles as the foundation for the definition of minimal competence and task creation, initial evidence of content validity based on the conceptual framework was easy to establish. Using standards-based language and aligning tasks to the standards made it clear that the tasks were relevant. The coverage analysis described and demonstrated above also indicated that the tasks were representative of the domains being assessed. Two former school principals on the design team validated the criticality of each task before it was included in the system, thereby establishing job-relatedness (Wilkerson et al., 2002).

Content Validity Evidence Based on Expert Judgmental Review

A second validation (Wilkerson & Lang, 2002a) was provided through a judgmental review by representatives of a large Florida school district. This was the first stakeholder (or expert) review, in which prospective users from staff development reviewed the tasks for criticality and authenticity (job-relatedness). Modifications were made based on their feedback to individual tasks, but no tasks were eliminated or added.

Construct Validity Evidence on Job-Relatedness

Method: The third validation study (Wilkerson and Lang, 2003a) was conducted during the first year of implementation (2002-2003), after multiple districts had had an opportunity to start using the tasks. A survey of district personnel was administered, using a theoretical framework based on the suggestions of Crocker (1997). In her validity study of the assessments used by the National Board of Professional Teaching Standards (NBPTS), Crocker asked judges to rate the frequency, criticality, and realism of the performance exercises. In the Florida study, frequency and criticality were also used but authenticity was substituted for realism. These three terms were used as an operational definition of "job-relatedness," and they are consistent with the *Standards* (APA, AERA, and NCME, 1999).

The questionnaire was distributed electronically by the Florida Department of Education with scores received electronically. Each of the 42 Assessment tasks was assessed on all three criteria. Each criterion was assessed on a three-point scale as follows:

Frequency (F): How frequently are the knowledge, skills, and attitudes measured in the tasks evidenced by good teachers in the classroom? How often should they display these skills? 3 = daily or weekly, 2 = monthly, 1 = once a semester , 0 = or not at all.

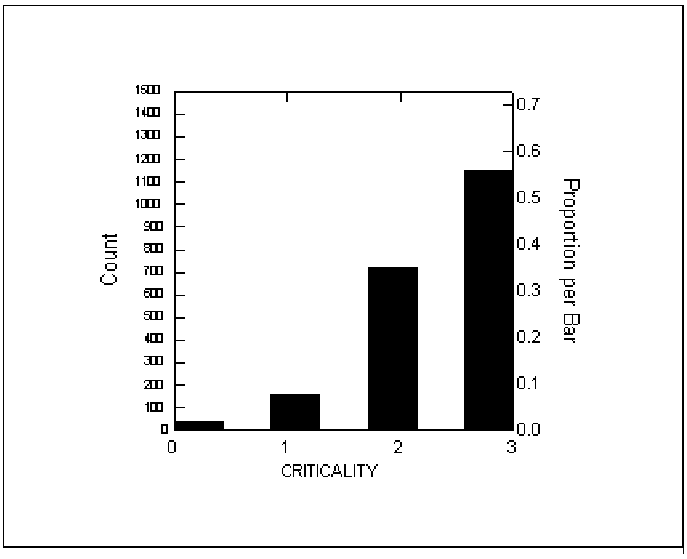
Criticality (C): How important or critical are the knowledge, skills, and attitudes measured in the tasks? 3 = critically important , 2 = very important , 1 = somewhat important, 0 = not important at all.

Authenticity (A): Determine if teachers really do the kind of work represented in the tasks (or are observed for these behaviors) -- even if they do not typically put the results of their work in a neat folder or write a report/reflection about it. 3 = highly authentic, 2 = moderately authentic, 1 = slightly authentic; 0 = not authentic at all.

Results: A total of 49 district coordinators or mentors responded after one year's experience with the FACP tasks. The following general conclusions for all tasks combined were made:

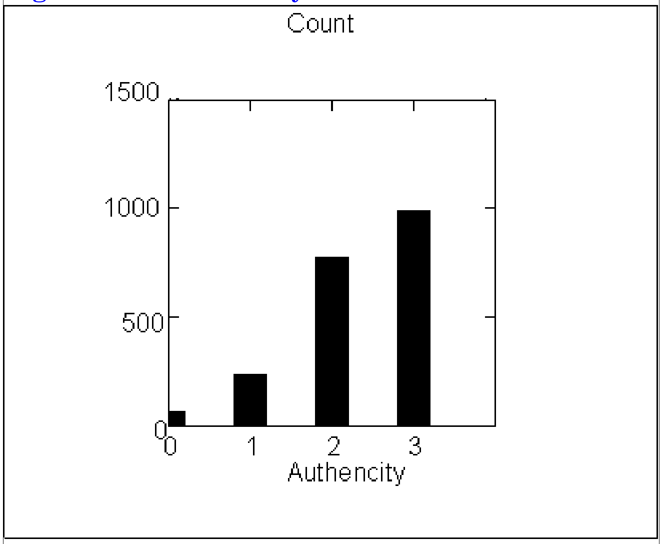
- Criticality: Across all tasks, more than half of the ratings (56%, n=1149) indicated that the tasks were critically important and another third (35%, n=717) very important. Thus, a total of 91% of the responses indicated that these tasks are critical to the performance of a competent teacher. Less than one-tenth (8%, n=157) indicated the tasks to be somewhat important, and only 2% (n=35) unimportant. This is graphically represented in Figure 2:

Figure 2. Criticality of ACP Tasks



- Authenticity: It is important to strive for authentic tasks; however, from time to time, a standard may best be assessed at the knowledge level. In these cases, the authenticity criterion was rated less high. Across all tasks, almost half of the ratings (48%, n=998) were highly authentic and slightly more than another third (37%, n=770) moderately authentic. Thus, a total of 85% of the responses indicated that these tasks are authentic tasks typically performed by a competent teacher. A little more than one-tenth (12%, n=237) of the ratings were slightly authentic, and only 3% (n=63) not authentic. Since some tasks needed to be at the knowledge level to assure representativeness of standards coverage, the lower authenticity rate was expected. This is reported graphically in Figure 3:

Figure 3. Authenticity of ACP Tasks



- Frequency: Frequency was the least important criterion assessed and also the most difficult to group. Some tasks are performed infrequently, but still may be critically important, e.g., semester planning or development of a professional development plan. It would be expected, therefore, that these tasks be rated as infrequent (a one on this scale). Most important in interpreting the results of the frequency criterion is the low incidence of zeros – or "not at all." Only 3% (n=63) of the ratings were a zero.

Similar results were determined for each of the Florida Educator Accomplished Practices, when tasks were combined by FEAP.

Construct Validity Evidence Based on Empirical Analysis

The effort in Florida depends on both judgmental and empirical analysis, and this is recommended for all institutions and districts developing an assessment system. Some references are included here for interested readers. The Rasch model of Item Response Theory has been chosen as the measurement model for this system, in part because the model is robust with regard to missing data and accommodating different item types in a test (Wright & Panchapakesan, 1969). Given the on-going nature of assessment in this system, with teachers completing the "test" over a two-year period rather than in one sitting, the robust nature of the model for missing data is extremely important for on-going diagnostic and remediation purposes. Also, the tasks are clearly using different item structures (observations, products, etc.) as measures. Another advantage of the choice of the Rasch model is the ability to detect and correct rater effects in judged assessment (Myford & Wolfe, 2003).

There are two ways to establish empirical construct validity that are useful for measures in systems such as this. One is the operationalization or functioning reality of the measures, which Trochim (2002) calls Translation Validity and consists of a blend of face validity and content validity. This approach asks the basic question of whether or not the numbers are working in different situations as expected to support the definition of the construct. Additional empirical evidence includes many descriptive analyses that use measures resulting from the tests as part of convergent and discriminant validity studies such as multitrait-multimatrix. The choice of the Rasch model for item analysis is also useful for this purpose.

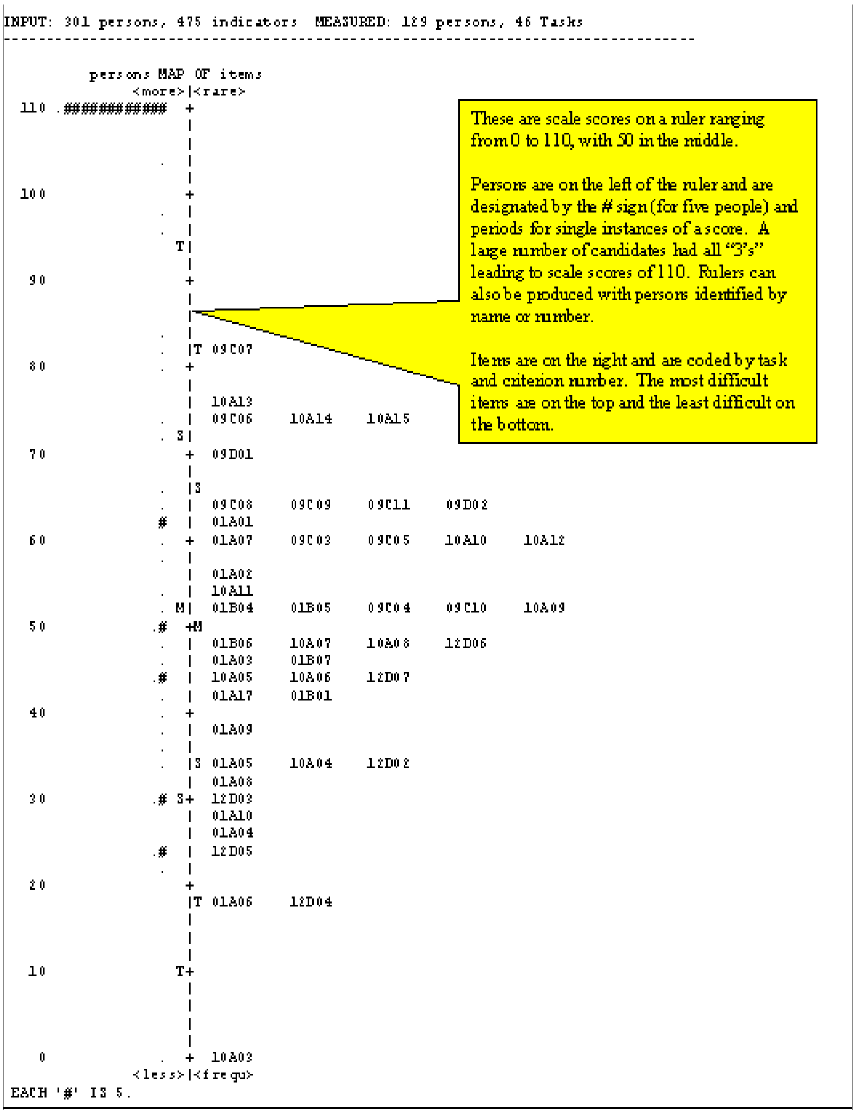
Bond & Fox (2001) stated that, "In his American Psychological Association (APA) presentation, *Construct Validity: A Forgotten Concept in Psychology?,* Overton (1999) detailed the importance of Fisher's (1994) claim that the Rasch model is an instrument of construct validation." (p. 192). Fisher (2001) later describes the internal statistical analysis of a test as necessary to establish construct validity separately from content validity. Linacre (1996) describes the comparison of Rasch and the true score models for various correlational studies that would be typical of convergent and discriminant validity studies. Linacre demonstrates the advantages of the Rasch model as opposed to a true score model for applications similar to the performance system described here.

Early results using the Rasch procedure with the Florida performance tasks support empirical evidence of construct validity. Figure 4 provides the sample logistic ruler, calibrating the items from 5 of the 42 tasks in the current performance system. Even at these early calibration stages, where sparse data remain largely unconnected, it is possible to confirm that items that were expected to be more difficult are being scaled as more difficult and items expected to be easier are being scaled as less difficult. One example is demonstrated in Figure 4 from task 01A – Unit Exam. Instructors' experience indicates that teachers have less difficulty in making ESE and ESOL accommodations on tests (criterion 5, coded 01A05) than on matching test items to instructional content (criterion 1, coded 01A01). Further, the lack of gaps in the scale of items supports the adequacy of coverage of the domains, an indication of content validity interpreted as construct validity by Trochim (2002).

The successful calibration of items onto an interval level scale (logistic ruler) is an important step for any number of future criterion-related validity studies. The complete discussion of construct evidence from Rasch analysis is beyond the purpose of the current article, but useful statistics would include coherence (Lopez, 1996), separation (Linacre, 1996), fit (Bohlig M., et al., 1998), discrimination (Engelhard, G., 1994), and principal components analysis (Linacre, 2003). In our example, all these statistics were consistent with predicted construct validity but are not reported here. For an alternate classical treatment of psychometric properties, even though we didn't find it as useful for our application, see Ingersoll & Scannell (2002).

The ruler in Figure 4 also provides evidence of a simpler and more often overlooked component of validity: operational functionality. An assessment is only as good as the ability to report information that is practical and informative to the user. Percent correct or percentile rank results accompanied by a cut score are weak for these purposes. In the example below, even those who are not statisticians can quickly see that most teachers have mastered the tasks, but that a few are lacking. Outliers among both persons and items are readily observable. Gains on the measures, prerequisite ordering of tasks, gaps and redundancy of items, specific diagnosis of person weaknesses, and the interaction of different tasks are graphically visible. A few points are demonstrated in the callout on Figure 4 to illustrate.

Figure 4: Logistic Ruler (Scaled Scores for Items and Persons) on skills based tasks



```
INPUT: 301 persons, 475 indicators  MEASURED: 129 persons, 46 Tasks
--------------------------------------------------------------------------------

         persons MAP OF items
                <more>|<rare>
 110 .###########  +
                   |
                   |
                .  |
                   |
 100               +
                .  |
                .  |
                  T|
                   |
  90               +
                   |
                   |
                .  |
                .  |T 09C07
  80            .  +
                   |
                   |  10A13
                .  |  09C06     10A14     10A15
              . S|
  70               +  09D01
                   |
                .  |S
                .  |  09C08     09C09     09C11     09D02
                #  |  01A01
  60            .  +  01A07     09C03     09C05     10A10     10A12
                .  |
                   |  01A02
                .  |  10A11
              . M|  01B04     01B05     09C04     09C10     10A09
  50           .#  +M
                .  |  01B06     10A07     10A08     12D06
                .  |  01A03     01B07
               .#  |  10A05     10A06     12D07
                .  |  01A17     01B01
  40            .  +
                .  |  01A09
                .  |
                .  |S 01A05     10A04     12D02
                   |  01A08
  30           .# S+  12D03
                   |  01A10
                   |  01A04
               .#  |  12D05
                .  |
  20               +
                   |T 01A06     12D04
                   |
                   |
                   |
  10              T+
                   |
                   |
                   |
                   |
   0            .  +  10A03
               <less>|<frequ>
 EACH '#' IS 5.
```

These are scale scores on a ruler ranging from 0 to 110, with 50 in the middle.

Persons are on the left of the ruler and are designated by the # sign (for five people) and periods for single instances of a score. A large number of candidates had all "3's" leading to scale scores of 110. Rulers can also be produced with persons identified by name or number.

Items are on the right and are coded by task and criterion number. The most difficult items are on the top and the least difficult on the bottom.

Notes:

1. Data on 301 persons and 475 items were input and measures for 129 persons and 46 persons were calibrated.
2. Reliability for the items was .83. Reliability in this model is equivalent to KR-20, Cronbach's alpha, and generalizability coefficients.
3. Real separation for items was 1.94. Separation is in root mean square error units, and it is the adjusted standard deviation divided by the root mean square error. This is easier to interpret than reliability coefficients.
4. Reliability = separation2 /(1+ separation2)

Plans for Continued Implementation and Improvement

The most obvious plan for continued implementation and improvement is the continuing development of the tasks. It is expected that not only will existing criteria be refined over time but also that additional tasks and alternative tasks will be added to the system.

Additional validity studies are always expected as part of monitoring a system and continuing to improve it. Demographic data are being entered into the tracking system to look for bias in decisions. The survey will be repeated seeking a larger response rate, once more assessors and coordinators have direct experience with all tasks. Plans are underway to conduct a convergent validity study with district-used observations based on FPMS. Once there are adequate graduates from the program to study, criterion validity using retention and rehire rates are possible. More powerful studies will be available as a consequence of having created interval level scores with IRT for measurement with any choice of quantitative methods or unit of analysis, something that is virtually impossible from a mix of non-parametric data, surveys, and unrelated demographic statistics that are currently being used in most teacher quality assessments today (Wright & Stone, 2004).

The performance task approach to assessing teaching competence holds promise for measuring teacher development at the pre-professional stage as well as on an on-going basis throughout teachers' careers. Beginning work on such a set of progressive standards has been started in Florida, since the Florida Educator Accomplished Practices have been benchmarked with sample key indicators for each of three career levels – pre-professional (prior to entry into the profession), professional (one to two years of teaching), and accomplished (master teacher status). Unfortunately, the indicators are not as clear and measurable as they could be. These indicators, however, when combined with the INTASC Principles and the National Board for Professional Teaching Standards (NBPTS) hold promise for developing a realistic, limited, and measurable set of standards allowing for the development of construct valid tasks in a comprehensive system.

The expectation is to develop a set of tasks aligned with standards that are calibrated on levels of competence (horizontal multidimensional model) throughout teachers' careers (vertical model). As tasks are added to the system and old tasks are modified, equating studies can be conducted.

A vertically moderated or calibrated model, as suggested by Lissitz and Huynh (2003), and a horizontal multidimensional model as suggested by Briggs and Wilson (2003), are under consideration.

Such a scale can be used not only in decisions related to certification, but also for career advancement and merit pay (career ladder), focused professional development plans for individual teachers, focused in-service plans for districts, and research on variety of important questions (e.g., relationship of teacher quality to class size, impact on K-12 learning by teacher ability level, etc.) This scale will open a wide range of research opportunities that can add more evidence of validity for this system. For example, we may explore the effect of interventions on task performance, predictions of retention rates based on scale scores, and profiles of teachers likely to stay in the profession based on demographics and scale scores.

Conclusions

Objective tests and certification exams, observations that tally or rate behaviors, and portfolios provide important evidence of job performance and the potential for good job performance, but they should not be used to the exclusion of comprehensive performance assessments over time. Multiple assessments, reflective of the day-to-day work of teachers, aligned with standards that define what a quality teacher knows and can do is a reasonable approach to measuring teacher competence.

This article suggests a five step design process for standards-based performance assessment of teachers that starts with an analysis of psychometric issues and continues with systematic design procedures. We have titled the model, Competency Assessment Aligned with Teacher Standards (CAATS). The model combines the requirements of teacher standards and the *Standards of Psychological and Educational Testing* of the Joint Committee of AERA, APA, and NCME (1999). While the tasks themselves will change and grow over time, the process should remain stable. Other states and individual institutions can apply this approach in identifying critical tasks for teacher assessment within their own local contexts.

Testing of the model on the Florida Alternative Certification Program Assessment System, which has been modified for use in teacher preparation programs, has yielded useful evidence of the validity of decisions that can be made using this system. This responds to the challenge offered by the National Commission for Teaching and America's Future.

The purpose-driven CAATS model and its application with this FACP assessment system provide promise for combining several important things: (1) course-embedded or on-the-job performance tasks that provide data for certification and accreditation , (2) a system that has psychometric integrity, addressing not only accreditation standards but also test design standards, (3) an alternative to using portfolios of student-selected work that is best suited for other purposes, and (4) output that can be used to develop the first level of a career ladder of teacher skills using modern measurement techniques.

Just for Kids, Inc. and the Southeast Center for Teaching Quality (2003b) researched teacher quality indicators by comparing average and high-performing schools in Texas. Among the nine factors they identified that accounted for school differences was frequent, relevant assessment. High-performing schools used systematic processes for student assessments including early, on-going data collection that allowed for targeted instruction and interventions based on precisely identified individual needs. Low performing schools did not. In fact, there is much evidence that frequent assessment by classroom teachers promotes complex achievement (Gentile, 2003).

So if systematic, targeted, and on-going assessment works for children, why are we so reluctant to use it with teachers, too? The answer probably lies in the issue of academic freedom and the difficulty of achieving faculty consensus. Strong leadership is necessary in the colleges of education today to overcome this obstacle as we face two pairs of competing paradigms: (1) faculty consensus and institutional accountability and (2) political difficulty and psychometric necessity. Realization of the power and usefulness of a valid system will, hopefully, pull the opposite views together for the ultimate improvement of all teacher preparation.

Note:

This paper is based in part on work completed for two divisions of the Florida Department of Education – the Bureau of Educator Recruitment and Professional Development (faculty development workshops for program approval) and the Bureau of Educator Certification (assessment system for Florida Alternative Certification Program).

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.

Bond, T. G. & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences,* Mahwah, NJ, Lawrence Erlbaum.

Bohlig M., Fisher, W.P. Jr., Masters, & G.N., Bond, T. (1998) Content Validity and Misfitting Items. *Rasch Measurement Transactions, 12:1,* 607

Briggs, D.C. & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4:1,* 87-100.

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.

Crocker, L. (1997). Assessing content representatives of performance assessment exercises. *Applied Measurement in Education. 10:1*, 83-95.

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.

Engelhard, G. (1994). Resolving the attenuation paradox. *Rasch Measurement Transactions, 8:3,* 379.

Fisher, W.P. (2001). Invariant thinking vs. invariant measurement. *Rasch*

*Measurement Transactions 14:*4, 778-81.

Florida Education Standards Commission (1996). *Florida Educator Accomplished Practices.* Florida Department of Education, Tallahassee, FL.: Author.

Florida Statutes sec. 231.433 (1983), *State Master Teacher Program*, Tallahassee, FL.

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

Impara, J. C. & Plake, B. S. (2000). *A comparison of cut scores using multiple standard setting methods.* Paper presented at the Large Scale Assessment Conference, Snowbird, UT.

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.*

Linacre J.M. (1996) True-Score Reliability or Rasch Statistical Validity? Rasch Measurement Transaction 9:4 p. 455-6

Linacre, J.M. (2003) A User's Guide to Winsteps Rasch-Model Computer Programs, Chicago.

Lissitz, R. and Huynh, H. (2003). Vertical equating for state assessments: Issues and solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation. 8:1.* Retrieved December 26, 2003 from http://PAREonline.net/getvn.asp?v=8&n=10.

Lopez, W.A. (1996). Communication validity and rating scales. *Rasch Measurement Transaction. 10:1*, 48.

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

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement.* 4:4, 386-421.

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

National Commission on Teaching and America's Future (1996). What matters most: Teaching for America's future. New York: Author.

National Council for Accreditation of Teacher Education (2001). *Professional Standards for the Accreditation of Schools, Colleges, and Departments of Education.* Washington, D.C.: 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.

Rudner, L. M. (2001). *Measurement Decision Theory*. Retrieved May 28, 2004: http://edres.org/mdt/home3.asp

Southeast Center for Teaching Quality (2003a). NCLB Teaching Quality Mandates: Findings and Themes from the Field, *Best Practices and Policies. 3: 4*, December. Chapel Hill, NC: Author.

Southeast Center for Teaching Quality (2003b). How do teacher learn to teach effectively? Quality indicators from quality schools. *Best Practices and Policies. 2: 7.* January. Chapel Hill, NC: Author.

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

Trochim, W. M. (2002). *The Research Methods Knowledge Base, 2nd Edition*. Available online at: http://trochim.human.cornell.edu/kb/index.htm.

Wilkerson, J.R. & Lang, W.S. (2003a). Florida Alternative Certification Program Assessment 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.

Wright, B. D. & N. A. Panchapakesan (1969). A procedure for sample free item analysis. *Educational and Psychological Measurement, 29*, 23-48.

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

Descriptors: Beginning Teachers; Assessment; Formative Evaluation; Standards; Accountability

Citation: Judy R. Wilkerson & William Steve Lang (2004). A standards-driven, task-based assessment approach for teacher credentialing with potential for college accreditation. *Practical Assessment, Research & Evaluation*, 9(12). Available online: http://PAREonline.net/getvn.asp?v=9&n=12.