December 2003

Education Policy Analysis Archives 11/45

Arizona State University

University of South Florida

EDUCATION POLICY ANALYSIS ARCHIVES

A peer-reviewed scholarly journal
Editor: Gene V Glass
College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

Volume 11 Number 45 **December 3, 2003** **ISSN 1068-2341**

Portfolios, the Pied Piper of Teacher Certification Assessments: Legal and Psychometric Issues

Judy R. Wilkerson

William Steve Lang
University of South Florida, St. Petersburg

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

Abstract

Since about 90% of schools, colleges, and departments of education are currently using portfolios of one form or another as decision-making tools for standards-based decisions regarding certification or licensure (as well as NCATE accreditation), it is appropriate to explore the legal and psychometric aspects of this assessment device. The authors demonstrate that portfolios being used in a high-stakes context are technically testing devices and therefore need to meet psychometric standards of validity, reliability, fairness, and absence of bias. These standards, along with federal law, form the cornerstone for legal challenges to high-stakes decisions when students are denied a diploma or license based on the results of the assessment. The conclusion

includes a list of requirements and caveats for using portfolios for graduation and certification decisions in a standards-based environment that help institutions reduce exposure to potential litigation.

The Portfolio: Panacea or Pandora's Box

Portfolios, both paper and electronic, have become hot topics in standards-based performance assessment. Salzman, et al. (2002) report that almost 90% of schools, colleges, and departments of education (SCDE's) use portfolios to make decisions about candidates, and almost 40% do so as a certification or licensure requirement. In our own recent study of teacher preparation programs in Florida, we found that virtually every institution in the State is using portfolios in some way to help make certification decisions (Wilkerson and Lang, 2003). In fact, portfolios seem to be viewed by many as the panacea of performance assessment. It is hard to go to national meetings without being greeted by software professionals who have designed electronic portfolio products that they claim will help SCDE's meet state and national standards for accreditation and program approval. Yet, for many educators, the jury is still out. Some have not yet reached a conclusion about whether or not to use portfolios for teacher certification. Others are reconsidering this decision having determined that the time involved for both faculty and candidates is excessive. Hence, there is a need to clarify the issues being raised nationwide.

As teacher educators, we view the standards movement as an appropriate impetus for the continuing professionalism of teaching when standards are used as a vehicle to redesign and improve teacher education curriculum and licensure (Wilkerson, et al, 2003). They provide a vehicle for professionals to articulate what they believe is important, and this is probably why there are so many sets of standards. There are the Interstate New Teacher Assessment and Support Consortium (INTASC), Specialty Professional Associations (SPA's) affiliated with NCATE, state program approval standards, state K-12 standards, institutional outcomes and conceptual frameworks. Standards also provide a vehicle for college faculty to justify their curriculum and a challenge to SCDE's to manage the assessment process. NCATE and many states require the use of multiple assessments to deal adequately with their complexity.

As measurement professionals, we are frequently asked if portfolio assessment can be used as an appropriate and safe vehicle to make summative decisions in a certification context. Are they good measurement? Our answer is this: "No, unless the contents are rigorously controlled and systematically evaluated." As Ingersoll and Scannell (2002) pointed out, portfolios are not assessments, but are instead collections of candidate artifacts that present examples of what candidates can do. The contents need to be evaluated individually as part of the candidate's overall performance record using a database format.

Without proper attention to the psychometric requirements of sound assessment, teacher educators may find themselves on a slippery slope. SCDE's have to make sure that assessment devices are created and used properly, and that costs money. Otherwise, SCDE's may make bad decisions and face legal complaints that can have severe consequences -- expensive

trials and court imposed interventions – not to mention institutional reputation. For example, Florida's Department of Education has an extensive history of assessment challenges on their web site: http://www.firn.edu/doe/sas/hsaphome.htm.

This does not mean that portfolios are bad or useless. They are excellent tools for reinforcing learning and for making formative decisions about candidate knowledge, skills, dispositions, and growth. However, when the decision is standards-based, summative, and results in initial certification, minimal competency must be established. Growth and learning are clearly important attributes of a quality teacher preparation program; however, these are not the critical assessment issues in initial certification. As important as they may be in determining if a certified teacher has achieved "master" or "accomplished" teacher status, this decision is vastly different from the one made for initial certification. In licensure, the state must ensure first and foremost that the teacher is "safe" to enter the profession and will "leave no child behind."

Looking at this issue from a different viewpoint, in medicine, society would not dream of allowing physicians to be licensed-based on their own selection of showcased successes. We recognize that many critical failures could be hidden behind their selected portfolio entries, and such failures could certainly prevent them from being "safe" practitioners. Medical licensure requires the identification and systematic assessment of a solid set of skills. Pilots, too, must pass a series of carefully constructed performance tests. We do not want to fly on an airplane where we forgot to measure whether or not the pilot could land the plane. Landing is part of minimal competence.

In portfolio assessment systems that allow candidates to choose their own artifacts, minimal competency with regard to standards is difficult to establish. There are too many "test forms" to establish either validity or reliability. When faculty fail to adequately align the artifacts selected by candidates with specific aspects of standards that define performance requirements, the range of material may preclude adequate standards-based decisions. When faculty fail to assess artifacts with solid, standards-based rubrics, it is difficult to interpret what their decisions mean and make appropriate inferences about what they know, can do, and believe.

Portfolio assessments, like all high-stakes tests, must stand the tests of validity, reliability, fairness, and absence of bias. If a candidate is denied graduation or certification based on a portfolio assessment that is not psychometrically sound, the candidate could successfully challenge the institution (and the state department that approved the program and its assessment system) in a court of law.

This article has been written to clarify the above opinions, which we recognize to be controversial. The issues are complex, technical, and inter-related. A thorough understanding of these issues requires somewhat detailed discussion of both the psychometric requirements of high-stakes testing and the legal requirements and decisions which are related to them. These are inextricably linked. If psychometric properties do not exist, the door to legal challenges from students is open. In considering the facts of the case, the courts then rely on

psychometric issues to make decisions regarding the infringement of students' legal rights. We hope that readers of this article will be better prepared to decide whether or not to use portfolios as a summative assessment and, if so, how to construct the requirements.

Since good teachers often attempt to make issues meaningful through some scenarios of what could happen, we will present a fictitious case study of Mary Beth Joanne to introduce readers to the important psychometric and legal issues discussed in this article. After our fictitious case study, we discuss the roles of SCDE's and state departments of education (DOE's) in teacher certification, the rationale for defining portfolios as a certification test, the legal challenges being posed with regard to certification and high-stakes testing, the psychometric issues affecting certification testing and portfolios, and the history of portfolios used as high-stakes tests. At the end, we will provide some caveats about portfolios as high-stakes tests, and we will conclude with some suggestions about the use of portfolios in training and high-stakes testing.

Mary Beth JoAnne Sues XYZ University

Mary Beth Joanne, nicknamed MBJ, is a fictitious student who attends XZY University, which is located in Florida where teacher education programs must certify that their graduates have demonstrated all 12 of the Florida Educator Accomplished Practices (FEAP's). The FEAP's are very similar to the INTASC Principles. Florida has added two Practices, one on ethics and one on technology, which are embedded within the INTASC Principles. XYZ University requires candidates to successfully complete an electronic portfolio showcasing their work on the FEAP's. Here are the "facts" about MBJ and XYZ:

- Mary Beth JoAnne is 35 years old, is a single mother of three, works 20 hours a week at TarMart, and has typically enrolled in 15-18 credit hours per semester. She wants to get her teaching degree as quickly as possible so she can leave TarMart. She has the required GPA (with a 3.0), has passed the certification exam, and has successfully completed all requirements of the internship except the portfolio requirement. Mary Beth JoAnne meets with the program coordinator, Jack, to challenge the result, since she has been given a "U" ("unsatisfactory") in internship. The grade of "U" will prevent her from graduating and receiving her professional teaching certificate. Jack upholds his decision. There is no further appeals process.
- XYZ candidates must have the required GPA, pass the State teacher certification exam, and successfully complete the portfolio and the final internship to graduate. If they successfully pass the state's background check, they are awarded a five-year professional certificate, renewable every five years thereafter.
- XYZ's electronic portfolio includes 12 sections one for each FEAP. At least three to five pieces of evidence are required for each Practice. The same evidence may be used for multiple practices. These requirements are properly documented in the XYZ portfolio materials, the catalog, and an advising sheet provided to students upon admission to the program.
- For each piece of evidence, candidates reflect on their work, linking it to the appropriate FEAP. The burden of proof, therefore, begins with the

candidates; faculty either concur, or do not concur, with the student's reflection decisions, based on their re-evaluation of the work. Discussion about the FEAP's and strategies to write reflection are integrated into the curriculum.
- MBJ has attempted to complete the portfolio, but it was found to be "unsatisfactory" on two separate occasions in two sections. She failed to demonstrate the State's Practice on Critical Thinking (FEAP #4) because she was unable to provide any examples of elementary student work showing that they had learned to think critically in her classroom. She also failed to demonstrate the adequate use of technology (FEAP #12) in the portfolio itself.
- There are orientations for students at the beginning of each semester to train them in the creation of their portfolios. The requirements are distributed or re-distributed at that time. Faculty also trained on scoring the portfolios. Faculty advisors help candidates select their materials and sometimes provide candidates with the opportunity to fix their errors. Course syllabi provide advice on evidence that may be used in the portfolio, linking tasks to standards.
- The portfolios are reviewed prior to internship and at the end of internship. XYZ uses a scoring rubric for the portfolios that asks faculty to determine if the candidates have demonstrated each of the FEAP's and selected indicators for those FEAP's. Inter-rater reliability has been established.
- A fully equipped computer and materials lab is available Monday through Friday, 8 am to 5 pm.

The following are some scenarios invented to show what might happen if Mary Beth JoAnne decides to sue XYZ. Of course, there are many variables that remain unknown – testimony and dispositions, expertise and predispositions of lawyers and judges, etc. These scenarios are intended as food for thought.

Scenario #1

MBJ is Hispanic; her father is from Cuba, and her last name is Gonzalez.

She files a claim under Titles VI and VII of the 1964 Civil Rights Act. The results follow and are outlined in the steps used by the courts in such cases:

- Step 1: XYZ analyzes the results of the portfolio evaluations, and a smaller percentage of Hispanics (70%) passed than non-Hispanic Caucasians (95%). The court determines that there is disparate impact on minorities (biased results) with this test.
- Step 2: The burden of proof shifts to XYZ. MBJ claims that the portfolio could not provide valid evidence of her potential to perform in the classroom (i.e., to be certified). XYZ claims that the evidence is valid because the portfolio requirements were developed in direct response to the State's requirements, and it was organized around the State's FEAP's. The court finds as follows:
    - The court upholds XYZ on the decision about critical thinking, because the task is found to be job-related. The judge's opinion notes that the State places a heavy emphasis on teachers' ability to impact K-12 learning, and this is documented in both State Statute

and State Board of Education Rule. The K-12 students' work is found to be one of the best measures of effective teaching within an internship context. There is an appropriate relationship between the requirement and the purpose, thereby establishing some evidence of validity.

- The court finds that XYZ does not meet its burden of proof, however, for several other reasons. The most significant of these is that XYZ cannot show the relationship between the creation of a teaching portfolio and what teachers actually do in the classroom, thereby failing to establish adequate evidence of validity. While research indicates that portfolios are used as appropriate vehicles for self-improvement and showcasing, and MBJ may eventually need to create a portfolio for national level certification through NBPTS, this is not a task she would do in her K-12 classroom to help children learn. In fact, many schools in Florida do not have computers. More important, the standard on technology requires that teachers use the technology within the context of instruction and the management of instruction. Therefore, this test does not meet the standards of representativeness or relevance for the 12th Accomplished Practice on Technology. It is not an authentic representation of the work to be performed by MBJ in the classroom and is, therefore, not job-related. The "business necessity" requirement for validity is not met.
- The court also finds that the entire portfolio is not valid because the use of three to five pieces of evidence has not been validated for representativeness or relevance, nor was there any attempt on the part of the institution to look at issues of proportionality. Some evidence, and some practices, may be more important than others. Some may require more or less evidence to cover the depth and importance of the practice. Furthermore, XYZ has no procedures in place to ensure that the evidence selected by each candidate will meet the requirements of representativeness, relevance, and proportionality (validity). The court finds that the inconsistency in the specific contents of the portfolios makes the validation of the test virtually impossible.
- The court also finds that the institution has not used any research-based techniques to determine the cut-score on the portfolio evaluation that could be reasonably used to differentiate between the potentially competent and incompetent teachers. There is no rational support for equally weighting the items used in each practice and there could be no such support since the items vary from candidate to candidate.
- The court finds that instructional validity is also limited, since the preponderance of work on the portfolio was extra-curricular. MBJ did not have adequate opportunity to learn the skills needed to prepare a portfolio, and she was given inadequate opportunity to remediate. These are also issues related to fairness and due process. The fact that she was able to document lack of support for, and experience in, the technological issues for building the portfolio adds weight to this claim. Finally, the court determines that it is not reasonable to require MBJ to use university labs that are only available during

weekdays when she is a working adult. This impedes her opportunity to learn and succeed.
    - The court finds that the use of different pieces of evidence by different candidates makes it impossible for adequate reliability studies to be conducted.
- Step 3: Not applicable, since MBJ prevails at Step 2. Step 3 addresses MBJ's rights to alternatives, and it is addressed below in Scenario #2.

Scenario #2

All of the contextual elements are the same; however, MBJ does not have very good lawyers. They do not make an effective case on all the aspects related to validity. Consequently, this time, XYZ prevails at Step 2. The trial moves to Step 3 and MBJ must prove that she was denied any reasonable alternatives. Remember Jack? He did not offer her any alternatives. MBJ now asserts that she should have been allowed to substitute some other technology-based work, e.g., the use of lessons infused with technology and the development of an electronic grade book.

In this scenario, MBJ prevails again. XYZ is unable to show that the alternatives would be less effective than the original requirement.

Scenario #3

All of the contextual elements are the same as in Scenario 1; however, MBJ is non-Hispanic Caucasian. Although females are a protected class, she knows that the statistics would not support a discrimination claim under Titles VI and VII. She does, however, have a due process claim under the 14th Amendment. She asserts that the bachelor's degree in elementary education is a property right of which she has been deprived without either substantive or procedural due process. The court finds the following:

- MBJ's rights to substantive due process were abridged on the same issues of content validity as described in Scenario #1 and this is sufficient for her to prevail.
- The procedural due process claim introduces new problems for XYZ. The court finds in MBJ's favor again on procedural due process because Jack's decision was not fair. MBJ was given no alternatives and no opportunities for an appeal. He just said "no." XYZ also takes no precautions against cheating and has no written policies about the assistance that faculty and peers can provide. Therefore, an unfair advantage is provided to some students who have multiple opportunities to revise their work and submit their portfolios, study with faculty who know how to use the technology and enjoy it, and receive substantive assistance from others.

The above scenarios do not address all of the things that can go wrong in a certification portfolio test. They do, however, provide some representative issues and results that may happen as the role of SCDEs continues to grow in the certification process. We will now address the contextual changes in

teacher certification that make the Mary Beth Joanne case relevant to many SCDE's.

Contextual Changes: Shifting the Burden for Competence and Certification

By the late 1990's, all states have adopted or seriously considered increased curricular and/or testing standards for minimally competent student performance in elementary and secondary schools (CCSSO, 1998). Public attention shifted to teacher competency, and a new teacher certification testing movement arose in the South in the 1970's and 1980's. The movement eventually spread to the rest of the United States. Sireci and Green (2000) identified 45 states that now require prospective elementary and secondary teachers to pass a teacher certification test as a prerequisite for employment.

Testing requirements came in a variety of forms, including both paper and pencil tests and classroom observations for teachers. In 1980, Georgia became the first state to require an on-the-job performance assessment for certification of beginning teachers. Georgia implemented a three-part assessment tool used in addition to a multiple choice certification test. This tool included a portfolio of lesson plans, an interview to discuss the portfolio, and an observation (McGinty, 1996). Florida, too, had an on-the-job performance observation system combined with a teacher certification test. This observation system was soon adopted or copied in other states, including Kentucky. It was called the Florida Performance Measurement System (FPMS).

These state assessments, both performance-based and traditional tests, met with some quick and negative results. Among the states challenged in court were Georgia, the Carolinas, Massachusetts, Texas, California, and Alabama. Despite the legal opposition, testing in one form or another has survived.

Pullin (2001) notes that one of the unusual aspects of teacher preparation has been that over the past fifty years, each of the states has delegated to public and private institutions of higher education much of the responsibility for awarding teaching credentials. States control the process of teacher education and certification through the state's mechanisms for approval of teacher education programs. Pullin (2001) asserts that once the state has approved the curriculum of an SCDE, program completion is tantamount to being certified or licensed. Perhaps this shift in responsibility for certification is the direct result of the legal challenges faced by the States in the certification testing process. The suggestion here is that this delegation of responsibility from one state or public agency to another (in the case of public institutions) includes a shift in psychometric responsibility and legal liability. As Lee and Owens (2001) note, one of the greatest challenges faced by teacher education institutions today is to provide evidence that their candidates and graduates have demonstrated the knowledge, skills, and dispositions to support the learning and well-being of all students.

In the case of Florida, where Mary Beth JoAnne resides and attends XYZ University, the state has been characterized as one of five "bellwether" states in which new trends develop and as a "high change" state with a history of

reform. In this vein, it is not surprising that the Florida Legislature provided the SCDE's with an enormous challenge that is representative of other states now or in the future. The program approval statute has as its intent the provision to SCDE's of the "freedom to innovate while being held accountable" (Chapter 1004, Florida Statutes. Thus, the responsibility for testing candidates for certification is shared in states like Florida by the DOE and the SCDE's through state-administered exams and institutional assessments, both of which constitute a form of high-stakes testing. This has caused much consternation within the SCDE's in the State, as institutions wrestle with tough questions about what kinds of assessments they can use and how they can combine them into a decision leading to graduation and certification. We have previously noted the following:

> "In Florida, in teacher education, the State uses the program approval process to hold institutions accountable. Florida is serious about this. Florida's first continued program approval standard requires that '100% of teacher candidates demonstrate each of the 12 Accomplished Practices.' No wiggle room. This high stakes requirement is causing institutions throughout the State to focus on how to operationalize the demonstration of competency for each of the Practices…The State of Florida has said to teacher preparation programs, "You must certify that your teacher candidates have learned what we require, and you must tell us how you know they learned it.'" (Wilkerson, 2000, p. 2)

Some readers may be reading this article from the perspective of preparing for NCATE accreditation in a state that does not require the institution to participate in the licensure decision. It should be noted that there is a difference between meeting national accreditation standards that assure the public of quality teacher preparation programs and issuing a certificate or license to teach that assures the quality of teachers. The major premise of this article is that the legal and psychometric standards to be applied to the assessments differ based on the requirements and mission of the agency to which the unit is responding. That does not mean that the unit assessment system cannot be the same, but the more stringent needs must prevail if the institution wants to be "safe." NCATE allows SCDE's to "work to establish the fairness, accuracy, and consistency of … assessment procedures" (NCATE, 2000, p. 21) to meet Standard 2. If the SCDE is offering a diploma that leads to teacher certification or licensure, however, the standard to be applied is significantly higher. The SCDE becomes both a **test designer and test consumer**. We are using the word ***"test***." It is now appropriate to discuss the relationship of portfolios to testing.

Portfolios as Certification "Tests"

According to the definition of "tests" in the 1999 AERA/APA/NCME *Standards for Educational and Psychological Testing*, forms of testing may include traditional multiple-choice tests, written essays, oral examinations, and more elaborate performance tasks. Hence, portfolios that are composed of written reflections (a form of an essay) and products representative of the candidate's skills, and performance, fall under a professionally acceptable definition of

“test”. At another level, in her legal analysis of testing and certification issues, Pullin (2001), too, lumps together traditional tests and alternative assessments. Finally, the use of portfolios in high-stakes testing in states such as Georgia and Vermont lend further credibility to the classification of portfolios as a “test.” Hence, even if one does not typically think about a portfolio as a test, the classification of portfolios as a test is appropriate.

Since there are many perspectives of what a portfolio is or should be, a working definition for this article is needed. This article will use the one from Herman, et al (1992) that describes portfolio assessment as a “strategy for creating a classroom assessment system that includes multiple measures taken over time. Portfolios have the advantage of containing several samples of student work assembled in a purposeful manner. Well-conceived portfolios include pieces representing both work in progress and “showpiece” samples, student reflection about their work, and evaluation criteria.” (p. 120).

A decision about whether or not someone is allowed to enter into, or remain in, a profession or occupation is what is commonly called a “high-stakes” decision. Mehrens and Popham (1992) define high-stakes tests as tests used for decisions, such as for employment, licensure, or a high school graduation. They warn that when tests are used for high-stakes decisions, they will be subject to legal scrutiny. There is a strong possibility that individuals for whom an unfavorable decision is made will bring a legal suit against the developer and/or user of the test. They go on to note, however, that existing case law suggests that if tests are constructed and used according to existing standards, they should withstand that scrutiny.

Given the definition of a portfolio as a high-stakes test that serves, at least in part, to make a certification, licensure, or graduation decision, legal and psychometric issues apply. This is also true of any assessment device used in such decisions, regardless of whether it is authored by a test company, a state agency, or an SCDE.

Herman, et al (1992) follow their definition of a portfolio with some concerns (from Arter and Spandel, 1992) that should be kept in mind when using portfolios or other comprehensive assessment systems. These concerns serve as a useful introduction to the more technically stated issues to be raised in this article. The six concerns are (p. 200):

1. How representative is the work included in the portfolio of what students really can do?
2. Do the portfolio pieces represent coached work? Independent work? Group work? Are they identified as to the amount of support students receive?
3. Do the evaluation criteria for each piece and the portfolio as a whole represent the most relevant or useful dimensions of student work?
4. How well do portfolio pieces match important instructional targets or authentic tasks?
5. Do tasks or some parts of them require extraneous abilities?
6. Is there a method for ensuring that portfolios are reviewed consistently and criteria applied accurately?

Psychometric Issues and Legal Challenges

We have raised the specter of legal challenges, and it is time to address the challenges that can be faced in any certification test, be it large-scale or small-scale, state-administered or institutionally designed and administered. Legal challenges are based upon the convergence of federal law and psychometric properties. It is difficult, if not impossible, to separate the two.

A review of the research written about legal challenges indicates that there are four basic legal issues: two challenges under the 1964 Civil Rights Act (Title VI and Title VII) and two challenges under the Fourteenth Amendment to the United States Constitution (due process and equal protection). Title VI supplements Title VII by reinforcing the prohibition against discrimination in programs or activities that receive federal funding, which includes most SCDE's through grants and financial aid. (Sireci & Green 2000; Pullin, 2001; Mehrens & Popham, 1996; McDonough & Wolf, 1987; Pascoe & Halpin, 2001).

Precedent setting cases come from a variety of employment situations, both within and outside the field of education. Many challenges introduce psychometric issues, the chief of which is validity. The applicable guidelines and standards governing the psychometric properties of the test and the decisions made using the test, whether it be in the field of education or not, are based in educational psychology and measurement as well as employment guidelines. The two most influential resources that provide operational direction for these legal decisions are the 1999 AERA/APA/NCME *Standards for Educational and Psychological Testing* and the 1978 *Uniform Guidelines on Employee Selection Procedures* (Pascoe & Halpin, 2001).

Regarding the Civil Rights Act of 1964, Titles VI and VII forbid not only intentional discrimination on the basis of race, color, or national original, but also practices that have a disparate impact on a protected class. Courts use a three-step process, in which the burden of proof shifts back and forth from the plaintiff to the defendant. We used these three steps in our analysis of the Mary Beth JoAnne case. In the first step, the plaintiff must prove discrimination. The discrimination could either be intended or coincidental, but it is clearly the responsibility of the institution to ensure that unintended discrimination (disparate impact) does not occur. This is why the results changed from scenario to scenario, dependent on MBJ's ethnic background. She was a member of a minority group that was less successful than the majority population in the first scenario.

If discrimination has occurred, the defendant (SCDE) must demonstrate that the test was valid and is necessary, and this is most often linked to the job-relatedness (or the "business necessity") of the test. It is in this second step, where the legal and psychometric issues converge (Scenario #1 of MBJ). If the defendant proves in court that the test is valid, the plaintiff has one more chance to prevail. If he/she can prove that the defendant could have used an alternative test with equivalent results, the defendant will lose (Scenario #2).

There are two basic requirements in the U.S. Constitution's 14th Amendment

that apply to this context: equal protection and due process. For a plaintiff to win under the equal protection claim, it must be shown that there was intent to discriminate. This is difficult and, therefore, rarely used. The due process provisions, however, have become relatively common. They forbid a governmental entity from depriving a person of a property or liberty interest without due process of law. (The Debra P. v. Turlington case established the diploma as a property right.) There are two kinds of these claims: substantive and procedural due process. Substantive due process requires a legitimate relationship between a requirement and the purpose. This is much easier to establish than the business necessity requirement of the Civil Rights Act. Procedural due process requires fairness in the way things are done, and these include advance notice of the requirement, an opportunity for hearings/appeals, and the conduct of fair hearings. Psychometric properties are excluded from this claim. MBJ prevailed on both types of due process in Scenario #3. (Mehrens & Popham, 1992; Sireci & Green 2000).

Thus, the linkage between legal rights and psychometric properties can occur in two places, opening the Pandora's box of validity and reliability. First it can occur within the context of step two of a discrimination claim under Titles VI and/or VII of the Civil Rights Act where there is intended discrimination or disparate impact on a protected class. Second, it can occur within the context of a lack of a legitimate relationship between a requirement (e.g., a test) and a purpose (e.g., protecting the public from unsafe teachers) that constitutes a violation of substantive due process rights as assured by the Fourteenth Amendment of the U.S. Constitution.

There are other potential legal challenges as well, but they are beyond the scope of this article. Worth mentioning in passing, however, is the potential for challenges by faculty who are asked to conduct extensive work, without remuneration, outside of their regularly assigned course-based teaching assignments (Sandmann, 1998). This is, of course, particularly problematic with portfolios completed and reviewed outside of the regular course teaching/assessing process.

Can It Happen At My Institution?

While most of the precedents discussed in the literature refer to states and traditional teacher certification tests, institutions have been challenged on the quality of educational opportunities received in their program. They have successfully used contract and negligence law theories in asserting institutional failures to provide the educational services they felt they should have received. (Mellnick & Pullin, 2000) Now that institutions have received part of the burden of certification testing, this risk is increased and can readily be combined with the challenges previously encountered at the state level.

While courts generally hold that the policy of requiring successful performance on a teacher test is reasonable public policy, they scrutinize the tests and the test administration quite carefully. This scrutiny includes validity, reliability, and fairness. Even if a test is an appropriate measure of the knowledge and skills needed by teachers, it may not be a legal test if the cut score itself is not a valid indicator of teacher competence, set using professional standards (Mellnick &

Pullin, 2000). Since psychometric issues are so critical in preparing and administering a certification-related test, a discussion of the psychometric issues follows. The primary source of the discussion is based on the requirements established in the *Standards for Educational and Psychological Testing* (APA, AERA, NCME, 1999), which, along with the EEOC Guidelines (1978) is used consistently as the "standard" in legal disputes.

It is important for faculty designing and implementing tests used in the graduation/certification decision to understand and apply these requirements. Selected issues are described below, and they are particularly targeted at the use of portfolios in high-stakes decisions.

Psychometric Issues

Test Design

Issues related to test design are addressed in Section 3 of the AERA/APA/NCME Standards (1999). A critical element noted in the Standards is the need to carefully specify the content of the test in a framework. This framework is sometimes called a table of specifications or, in the case of traditional tests, a test map or blueprint. The Standards provide specific guidance with regard to performance assessments in general and portfolios in particular. Performance assessments are defined in this section as those assessments that "require the test takers to demonstrate their abilities or skills in settings that closely resemble real-life settings" (p. 41). They may be either product-based or behavior-based.

The Standards note that performance assessments typically consist of a small number of tasks that establish the extent to which the results can be generalized to the broader domain. The use of test specifications contributes to a systematic development of tasks and helps to ensure that the critical dimensions of the domain are assessed, leading to a more comprehensive coverage of the domain than is typically achieved without the use of specifications. The Standards also suggest that both logical and empirical evidence be gathered to document the extent to which the assessment tasks and scoring criteria reflect the processes or skills specified in the domain.

With regard to portfolios, the Standards define portfolios as systematic collections of work gathered for a specific purpose. They note that those who assemble the portfolios may select their own work, if that is appropriate to the purpose. However, the following caution is provided: "The more standardized the contents and procedures of administration, the easier it is to establish comparability of portfolio-based scores. Regardless of the methods used, all performance assessments are evaluated by the same standards of technical quality as other forms of tests." (p. 42).

Validity, Sampling, and Job-Relatedness

Section 14 of the AERA/APA/NCME Standards (1999) outlines the requirements for testing in employment and credentialing, focusing on the

applicant's current skill or competence, including entry into a profession and ranging from novice to expert in a given field. It is, therefore, one of the most relevant chapters in the Standards.

The Standards explain that licensing and certification requirements are imposed by state and local governments to ensure that those licensed or certified possess essential knowledge and skills in sufficient degree to perform their work safely and effectively, thereby protecting the public from non-qualified personnel. Tests used for this purpose are intended to provide the public with a dependable mechanism for identifying practitioners who have met particular job-related standards. Standard 14.14 requires that the content domain to be covered by a credentialing test should be defined clearly and justified in terms of the importance of the content for credential-worthy performance in an occupation or profession (AERA/APA/NCME, 1999). This is the basis for making the link in substantive due process claims between the requirement and the purpose, with the purpose referring to the State's role in certification to protect the public as delegated to SCDEs, and the requirement referring to the test including portfolios.

The content domain to be covered by a licensure or certification test should be defined clearly and explained in terms of the importance of the content for competent performance in an occupation (AERA/APA/NCME, 1999). The creation of the test requires that the author develop and implement a content sampling process. Construct irrelevant variances refers to the degree to which test scores are affected by processes that are extraneous to the intended construct. Construct under-representation refers to the degree to which a test fails to capture important aspects of the construct. It implies a narrowed meaning of test scores because the test does not adequately sample some types of content (AERA/APA/NCME, 1999).

A content validation examination can be conducted to determine if a representative sample of the domain of skills needed to perform the job is covered adequately -- often referred to as job-relatedness. To content validate a test, the test writers would examine all elements of the test and try to ascertain how well the test covered the essential areas of knowledge and skill. The extent to which the content is underrepresented or irrelevant becomes a critical concern. Proportionality of the items is another important issue. In order to meet the criteria of representativeness and proportionality, the test must reflect the entire breadth of the domain and it must place the greatest emphasis on the most significant aspects within the domain. A test that sampled knowledge or behavior from part of a domain would not be representative. A test that put great weight on insignificant or marginally related aspects of a domain would be disproportionate. In recent years, these issues have been of major concern in determining the legal defensibility of employment, licensing, and certification tests (McDonough & Wolf, 1987; Sireci & Green 2000). Thus both sufficiency and relevancy are critical issues in test construction and were critical issues in the MBJ case, Scenario #1.

AERA/APA/NCME Standard 14.4 requires that all criteria used should represent important work behaviors or work outputs, on the job or in job-relevant training, as indicated by an appropriate review of information about the job. Standard

14.9 requires that when evidence of validity based on test content is a primary source of validity evidence in support of the use of a test in selection or promotion, a close link between test content and job content should be demonstrated. The rational relationship between what is measured on a certification test and what practitioners actually do on the job is usually established by conducting a thorough practice (or job) analysis. The practice analysis can be thought of as a very detailed job description, breaking down a profession into performance domains that are characterized by specific tasks. The tasks are further delineated into knowledge and skill statements that represent the essential qualities needed to perform each task (Sireci & Green 2000; Pullin, 2001; AERA/APA/NCME, 1999).

When a test attempts to sample a work behavior or to review a sample work product, then these should approximate the real-world work setting as much as possible (EEOC, 1978). Not all aspects of job performance need to be tested, but generally a test should be fairly representative of the job in question and courts may look more closely at tests which sample only a small part of the total job. ADA requires selection decisions be based upon the "essential functions" of a job (Pullin, 2001).

Although psychometricians and courts now have a disparity of opinion about what validity is, content validity remains the primary evidence used by courts when making decisions about fairness (Pascoe & Halpin, 2001). Job relevance has been an important issue in many of the employment test cases of the past twenty years. The valid use of the test is based on a clear understanding of the rational relationship between the test and the knowledge, skills, and abilities required to do the job (McDonough & Wolf, 1987). The job-relatedness standard was a major factor in the court finding for Mary Beth Joanne on the technology issue.

Lee and Owens (2001) asserted that most educational institutions and training and development companies do not conduct validity studies for two reasons -- lack of skill in conducting these studies and fear of spending the money it takes. They concluded that if only those companies who had been sued for unfair business practices had considered the alternative costs of defending themselves in court, they might have decided to learn how and conduct the needed studies.

Reliability and Measurement Error

Reliability refers to the consistency of measurements when testing procedures are repeated. It is assumed that there is a degree of stability in scores, but there also needs to be an accounting for measurement error, or score variability that is not related to the purposes of the measurement. Measurement error can come from differences in the difficulty of different test forms (e.g., different work samples in different students' portfolios); fluctuations in motivation, interest, or attention; intervention, learning, or maturation (e.g., uneven help in completing tasks and assembling portfolios).

The APA/AERA/NCME Standards (1999) specifically address the recent development of performance assessments large-scale testing and portfolios in

particular, especially those in which examinees select their own work or work cooperatively in completing the test. They note that, “Examinations of this kind raise complex issues regarding the domain represented by the test and about the generalizability of individual and group scores. Each step toward greater flexibility almost inevitably enlarges the scope and magnitude of measurement error.” (p. 26) This was the case at XYZ University.

The Standards indicate that information about measurement error is essential, whether the test is of a traditional nature or is a portfolio of work samples, or other forms of performance assessment techniques. “No test developer is exempt from this responsibility” (p. 27). Critical information to be obtained includes the sources of measurement error; summary statistics on the size of such errors; and the degree of generalizability across alternate scores, forms, and administrations. Where there is significant subjectivity, indexes of scorer consistency, often called inter-rated reliability, are also common.

It should be clear from the above that there are multiple issues related to reliability that need to be studied. Inter-rater reliability is but one of these issues. Many factors can contribute to error including rater training, rater mood or fatigue, unclear directions, number of items, variations in the types or difficulty of evidence evaluated (e.g., student selected evidence in portfolios), unequal assistance provided to candidates, cheating (those portfolios that are being sold or distributed on campus or on the Internet), and other such factors. Institutions that rely almost exclusively on an inter-rater reliability study to “handle” the psychometric requirements are in jeopardy. There may also be many other sources of error that go undetected, especially if faculty evaluators are just plain tired from reading so many portfolios or angry that they are being forced to do it just for accreditation purposes. Combined with the potential for a lack of validity if the evidence provided is either construct irrelevant or underrepresented, it may be that those high inter-rater reliability scores only indicate that raters who are tired are consistently rating highly (halo effect) the wrong stuff just to get finished.

Cut-Scores

As an SCDE or a DOE develops its tests, faculty must ask whether or not the content measured is relevant to making the decision about minimal competence, and the potential for adequate performance on the job. The portfolio, or any other assessment device, in this context, is a qualifications test, targeted at sorting those who should be allowed to teach from those who should not, based on what they will be expected to do on the job.

Designing the testing program includes deciding what areas are to be covered, whether one or a series of tests is to be used, and how multiple test scores are to be combined to reach an overall decision about whether or not the examinee is likely to engage in safe and appropriate practice. It is not only the internal aspects of the test that must be judged valid, but also the way in which the test is used to identify masters and non-masters or successes and failures. Defining the minimum level of knowledge and skill required for licensure or certification is one of the most important and difficult tasks facing those responsible for credentialing. This is accomplished by identifying and verifying a cut score or

scores on the tests and is a critical element in validity. The cut score must be high enough to protect the public, as well as the practitioner, but not so high as to be unreasonably limiting.

AERA/APA/NCME Standard 14.13 requires that when decision makers integrate information from multiple tests or integrate test and non-test information, the role played by each test in the decision process should be clearly explicated, and the use of each test or test composite should be supported by validity evidence. In some cases, an acceptable performance level is required on each test in an examination series. Standard setting procedures (e.g., Angoff) are designed to determine passing scores that distinguish those worthy of a credential from those who are not (AERA/APA/NCME, 1999; McDonough & Wolf, 1987; Sireci & Green 2000; Kane, 1994).

Section IV of the Standards also provides guidance on this issue. In the case of licensure or certification, the cut score should represent an informed judgment that those scoring below it are likely to make serious errors because of their lack of knowledge or skills. The most difficult part is weighing the relative probabilities of false positives (keeping good candidates out of the profession) and false negatives (letting poor candidates into the profession). Because this is largely a value-laden and subjective procedure, the qualifications of the judges used in standard setting are extremely important.

Fairness

The APA/AERA/NCME Standards (1999) outline four basic views of fairness: Three will be discussed because of their relevance to this article: lack of bias, equitable treatment in the testing process, and opportunity to learn.

Bias can occur when there is evidence that scores are different for identifiable subgroups of the population tested. Bias is determined by the response patterns for these groups. If a protected population (e.g., minorities, women, or handicapped) performs worse than the majority population, bias is an issue (MBJ Scenario #1). Bias may also occur as a result of the content of the test itself. The language of the material may be emotionally disturbing or offensive or may require knowledge more common to a specific group of examinees. Bias can also occur with a lack of clarity in instructions or scoring rubrics that credit responses more typical of one group than another. Another form of bias relates to the responses provided by the examinees. For example, if the examinees answer the way they think the scorers want, bias is an issue. A portfolio reflection reviewed for dispositions toward teaching, for example, could be filled with what the candidate thinks the professors want to see rather than what the candidate really believes.

Equitable treatment refers to the manner in which the test is administered. All examinees need to have comparable opportunities to demonstrate their ability, and this includes testing conditions, familiarity with format, practice materials, etc. Opportunities to succeed must be comparable. There must be equity in the resources available, and all examinees need to have meaningful opportunities to provide input to decision makers about procedural irregularities. In the case of portfolios, if one candidate has more opportunities than another to succeed

or to challenge, based on the support provided by faculty, fairness becomes an issue. These, too were issues for MBJ in Scenario #3.

Opportunity to learn requires that the institution assure that what is to be tested is fully included in the specification of what is to be taught. In the case of portfolios, then, where reflections are written after instruction is completed and are a critical component of the scoring, institutions would need to ensure that candidates had had adequate opportunities to learn how to self-assess at the level expected in the portfolio. Candidates would also have to have had adequate opportunities to produce sufficient materials in class to provide evidence of standards demonstration. Candidates also would need adequate opportunities to fix problems.

Legal Issues and Precedents

It is difficult to remain informed about current legal practice with regard to professional licensure, but it is important (Pascoe & Halpin, 2001). The courts have granted governmental authorities wide latitude as long as they have taken reasonable steps to validate the tests and the cutoff scores. Whether the plaintiffs are minorities or members of the majority population, other steps that the courts have considered include (1) providing ample prior notice before implementation of the high stakes phase; (2) allowing accommodations in the administration of the tests for the disabled; and (3) allowing retesting and, to the extent feasible, remediation (Zirkel, 2000).

Courts recently have been supportive of performance measures (Lee & Owens, 2001; Rebell, 1991). As far as teacher educators are concerned, Pullin (2001) notes that the courts have been generally reluctant to second-guess educators' judgments of educational performance based on subjective evaluations. This is of some comfort to the teacher education community. She goes on to say that in situations in which the individual stakes are not as high, such as during an educator preparation program or during a probationary period of employment, then fewer procedural protections are required. If the decision-making seems to be based upon the purely evaluative judgments of qualified professionals, courts may be reluctant to intervene. On the other hand, how can we be sure?

Lemke (2001), too, offers an opinion. She reviewed court decisions concerning the dismissal of college students from professional programs and determined that courts upheld school decisions when the institution followed its own published processes and the students' rights had been observed. This, too, provides for a high degree of comfort. If students are told what is expected of them in clear terms, colleges are safer. But Lemke also found that there is a lack of information about what the judicial system finds to be appropriate and inappropriate admissions and dismissal procedures. She looked at the decision of Connelley v. University of Vermont (1965), in which the federal district court ruled that it is within the purview of academic freedom for faculty to make decisions about students' progress. Faculty and administrators were described as uniquely qualified to make these decisions. In those days, though, certification was still the purview of the state. Lemke also reviewed eight cases of students filing against institutions. In these cases, the institutions had the right to make decisions about a student's academic fitness as long as it followed

its advertised processes. Reasons for dismissals that were upheld included the use of subjective assessments in clinical experiences, time requirements for program completion, comparison of test scores between the plaintiff and peers, GPA, and absenteeism.

Educators in Florida, though, have seen that the PK-20 system is not so safe. The groundbreaking Debra P. v. Turlington case (1979, 1981, 1983, 1984) begins to reduce the level of comfort engendered in the previous two citations. This was a diploma sanction case, bringing educators back to the issue of content validity. It is generally conceded that a state has the constitutional right to use a competency test for decisions regarding graduation. A diploma is considered a property right, and one must show some evidence of curricular/instructional validity or what is also called "opportunity to learn" or "adequacy of preparation." In this case both due process and the equal protection clauses of the 14th Amendment were found to be violated by Florida officials who were using a basic skills test for diploma denial at the high school level. In appeals, additional issues were raised about whether the test covered material that was adequately covered in Florida's classrooms, and this has become the major precedent for looking at "instructional or curricular" validity. The judge ruled that, "What is required is that the skills be included in the official curriculum and that the majority of teachers recognize them as being something they should teach." (Mehrens and Popham, 1992) In the MBJ case, XYZ required candidates to prepare their portfolios outside of their regular courses, thereby increasing their risk of challenge based on the principle of “opportunity to learn.”

The continuing shift in responsibility to SCDEs from DOEs for more and more of the burden of making certification decisions can easily result in successful claims by unhappy students who are denied their career dreams. The diploma denial challenges, combined with the challenges based on denial of a teaching certificate by a state agency provides for a natural leap to challenge diploma/certificate denial from an SCDE.

McDonough and Wolf (1988) identified five issues around which litigation against educational testing programs occurs: (1) the arbitrary and capricious development or implementation of a test or employee selection procedure, (2) the statistical and conceptual validity of a test or procedure, (3) the adverse or disproportionate impact of a testing program or selection procedure on a "protected group", (4) the relevance of a test or procedure to the identified requirements of the job (job-relatedness), and (5) the use of tests of selection procedures to violate an individual's or group's civil rights (McDonough & Wolf, 1987).

Courts have generally required evidence that the cut-score selected for a test be shown to be related to job-performance. In the Alabama case against National Evaluation Systems (NES), the test developers, the court found that the company engaged in practices "outside the realm of professionalism" and that it violated the minimum professional requirements for test development. Among the problems found were decisions in test development that resulted in test scores that were arbitrary and capricious and bore no rational relationship to teacher competence. There was a similar finding in Massachusetts against

the same company. In Groves v. Alabama Board of Education, the court found in 1991 that the arbitrary selection of a cut-score without logical or significant relationship to minimal competence as teacher had no rational basis nor professional justification. As such, it failed to meet the requirements of Title VI and was not a good faith exercise of the professional judgment. Evidence should be available that the cut-score for a test does not eliminate good teachers from eligibility for teaching jobs (Pullin, 2001).

The California Basic Education Skills Test (CBEST) was challenged in 1983 under Tile VII by the Association of Mexican-American Educators. The State won the case based on a job-relatedness study (Zirkel, 2000). In 1984, Florida lost a challenge to FPMS when the question of the validity of the decision about a teacher's certificate removal was successfully raised (Hazi, 1989). Georgia's TPAI challenge was won by the plaintiff based on due process and validity challenges (McGinty, 1996). The U.S. Department of Justice sued the State of North Carolina in 1975 under Title VII based on results on the National Teacher Examination from the Educational Testing Service. They won the claim when the court found the test to be unfair and discriminatory because a validation study had not been conducted and the passing score was arbitrary, thereby denying equal protection. A second similar claim was filed against the State of South Carolina, but in this instance the state prevailed based on a proper validation study causing the test to be deemed fair and appropriate (Pascoe & Halpin, 2001). There are many such discussions in the literature. The point is that tests, even those written by major test publishers, can be successfully challenged.

What History Tells Us About Using Portfolios as High-Stakes Tests

Before proceeding any further, it is important to underscore that the authors are not opposed to portfolios in a general sense. This article is about portfolios used in a certification testing context, particularly when there is a high degree of flexibility allowed to students in the selection of portfolio contents. There is much in the literature to support the use of portfolios as a tool for learning, particularly the reflective or self-assessment aspect. As noted earlier, they are excellent means for documenting growth, improving instruction and learning, and causing students of any age to construct meaning and value their own progress at meeting important instructional goals. For formative assessment, they can be superior assessments. For example, when Vermont implemented its K-12 statewide portfolio assessment system in 1988 as the first attempt in the U.S. to use portfolio assessment as a cornerstone of a statewide assessment, the results in these areas were clear and strong. The studies by the RAND Corporation and the Center for Research on Evaluation, Standards, and Student Testing -- CRESST (Koretz, 1994) clearly indicated that teachers thought that portfolios were helpful as informal classroom assessment tools but that they, too, were worried about their use for external assessment purposes. The majority of teachers surveyed agreed or strongly agreed that portfolios help students monitor their own progress. However, the vast majority did not believe it would be fair to evaluate students on the basis of their portfolio scores. Most felt that the state's emphasis on reliable scoring was misguided and perverted

the original purpose of portfolios as a tool for assessing an individual student's growth. Teachers were concerned about the validity of portfolios as an assessment instrument, particularly because of the large number of uncontrolled variables and the time burden both in class and outside of class. They felt they spent too much time managing and scoring portfolios and this detracted from their time to teach (Koretz, 1994).

In a subsequent study by Gearhart and Herman (1995), further support was given for the significant benefits for instructional reform being witnessed in Vermont, but the challenges were reinforced with the question about whose work was being judged when the work was composed with the support of peers, teachers, and others. They noted that to many committed to educational reform, portfolio assessment embodies a vision of assessment integrated with instruction. Advocates find that portfolios provide a richer and truer picture of students' competencies than do traditional or other performance-based assessments by challenging teachers and students to focus on meaningful outcomes. Integrated with instruction and targeted at high standards, the portfolio is seen by its advocates as the bridge between improved teaching and accountability. However, while the vision is enticing, Gearhart and Herman (1995) asked if it would work. The RAND study raised major issues about reliability; this study brought into question the validity of inferences drawn when the assessment results are compromised by questions about authorship and support. They concluded that from a measurement perspective, the validity of inferences about student competence based solely on portfolio work appeared suspect. The problem is troubling indeed for large scale assessment purposes where comparability of data is an issue.

Questions about using portfolios in high-stakes assessments have also been raised in the teacher certification arena. The Georgia Teacher Performance Assessment Instrument (TPAI), initiated in 1980, included a portfolio component and an observational component as an interview. The TPAI was initially successfully challenged by a teacher (Kitchens) for the validity of its observational component, which was found to include behaviors that were difficult to measure (e.g., enthusiasm). However, in the aftermath of the Kitchens case, the opposition to TPAI that grew was around the portfolio process, which was again found to be far too time consuming for a beginning teacher and not a valid measure of teacher performance because the portfolios were being judged on the basis of form rather than substance. The $5,000,000 "mammoth measurement tool" was laid to rest (McGinty, 1996).

At the institutional level, after the Alabama decision to terminate state testing because of racial bias, Nweke and Noland (1996) investigated the effectiveness of using performance and portfolio assessment techniques to diversify assessment in a minority teacher education program at Tuskegee University. They concluded that the observational component correlated highly with GPA but there were no statistically significant relationships between portfolios and GPA or portfolios and the performance assessment.

This is a representative, not an exhaustive, study of the use of portfolios in large and small scale assessments. Despite findings such as these, though, portfolios continue to be a major component of teacher assessment systems in SCDEs.

AACTE (American Association of Colleges of Teacher Education) conducted a survey of member institutions in fall 2001 (Salzman, 2002) on teacher education outcomes measures. The purpose of the study was to identify and describe what SCDEs are doing to meet the requirements for outcomes assessment for unit accreditation and program approval (teacher certification). They concluded that institutions are responding to more rigorous standards and to national and state mandates for accountability through multiple types of outcome measures, including portfolios. Results from the 370 responding institutions indicated that portfolios are used as an outcome measure by 319 (87.9%) of the responding institutions. Responses further indicated that 64 (20.1% of the institutions) do so in response to a state mandate while 269 (84.3%) do so as part of an institutional mandate. Portfolios were noted as required for certification or licensure by 123 (38.6%) institutions and not required by 159 (49.8%) for licensure. Data were missing from 37 (11.6%) of the respondents. Most units (305 or 95.6%) reported that the portfolio requirements were developed by the unit (Salzman, et al., 2002).

There is a school of thought that advocates strongly for portfolios. *With Portfolio in Hand*, a recent work edited by Nona Lyons (1998), contains several important chapters advocating for portfolios. Even in these chapters, the caveats exist. For example, although Moss proposes that validity issues related to assessment of teaching be rethought to allow for the benefits of portfolio assessment, she concludes with suggestions from classical theory. On the one hand, she proposes an integrative or hermeneutic approach to portfolio assessment in which raters engage in a dialogue to reach consensus about ratings, but she acknowledges that this is a time-consuming approach for which substantial empirical work is needed to explore both the possibilities and limitations. Even with this proposed new approach, Moss acknowledges the need to ensure the relevance, representativeness, and/or criticality of the performances and criteria as well as job-relatedness, social consequence studies, lack of bias, reliability, and most other aspects of psychometrics. Dollase (1998), while advocating for portfolios in teacher certification also acknowledges the severity of the issue of time in terms of the “doability” of the approach

Requirements and Caveats Regarding the Use of Portfolios as Certification Tests

To this point, arguments have been made that that the problems associated with portfolios in a high-stakes testing environment center around validity, reliability, fairness, excessive time burdens, and loss of the meaning and value of portfolios as a viable means to improve learning.

Based on this analysis of the literature, the authors have identified eight requirements for the construction of portfolios as tests used for certification in an SCDE. These will be accompanied by some caveats related to the use of portfolios for SCDE-based certification decisions added. They are listed in Table 1.

Table 1. Requirements and Caveats for Portfolio Use in Certification Testing in an SCDE

#	Requirement for Tests	Caveats for Portfolios
1	The knowledge and skills to be demonstrated in the portfolio/test 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.	If the portfolio is used as a test itself containing new or original work created outside of courses, rather than just a container of evidence of course-embedded tasks, the portfolio must stand the test that it is job-related and authentic. The SCDE should be prepared to defend how portfolio preparation as a stand-alone activity is a critical job function that teachers perform on a routine basis, similar to lesson planning, communication with students and parents, assessment, teaching critical thinking skills, etc. In the case of electronic portfolios, if the product is used to demonstrate a standard or expectation on technology that relates to using technology in the classroom, the SCDE will need to justify that that the preparation of the portfolio is equivalent to what teachers do with technology in the classroom. This may be difficult from an authenticity perspective.
2	The entire portfolio/test (assessment system) must meet the criteria of representativeness, relevance, and proportionality	If the portfolio is a container of evidence used as a summative assessment for the certification/graduation decision, the SCDE must be prepared to defend the contents of portfolios submitted by all candidates for the representativeness, relevance and proportionality of contents against the requirements of the teaching profession, e.g., the standards being assessed from national and state agencies as well as the institution itself (conceptual framework). If the portfolio is a specific piece of evidence itself, then its place within the assessment system must be included in the analysis of representativeness, relevance, and proportionality. All criteria used to evaluate the portfolio must be relevant to the job. Criteria such as neatness and organization are particularly suspect, unless they can be directly tied to the potential for poor performance in the classroom. The SCDE will need to prove that sloppy or disorganized teachers cannot be effective teachers.
3	There must be adequate procedures and written documents used to provide notice to candidates of the requirements, the appeals process, and the design (fairness) of the	The SCDE must have adequate documentation in place that tells candidates how and when to prepare the portfolio, how it will be reviewed, who is allowed to help them and how much help they can receive, the consequences of failure and the opportunities for remediation, and what their due process rights and procedures are if they wish to challenge the review results.

appeals process.

4	There must be adequate instructional opportunities provided to candidates to succeed in meeting the requirements of the portfolio/test and to remediate when performance is inadequate.	The SCDE should embed portfolio preparation, including the contents of the portfolio, into its instructional program (i.e., coursework). Any requirements outside of the instructional program could be subjected to a claim based on instructional/curricular validity. The entire faculty need to buy into, and support, portfolio preparation activities of the students and provide remedial opportunities for components that are found lacking.
5	There must be a realistic cut-score for determining if the performance is acceptable. This cut-score must differentiate between those who are competent to enter the profession and those who are not.	This is the most difficult aspect of portfolio design. The SCDE will need to identify the specific score or characteristics that sort teachers into the dichotomous categories of competent and not competent based on their portfolios.
6	Alternatives must be provided to candidates who cannot successfully complete requirements, or the SCDE must be able to demonstrate why no alternatives exist.	If the portfolio is a container of evidence, the alternatives must relate to specific pieces of evidence. The institution must ensure, however, that alternatives do not detract from the representativeness, relevancy and proportionality criteria. If the portfolio is used as evidence of a specific standard, such as reflection, then an equivalent alternative should be identified if at all possible.
7	The results of the portfolio evaluation (scoring) and the extent to which protected populations are equally or disproportionately successful must be monitored.	If the SCDE finds that a disproportionate number of protected populations (minorities, handicapped, women) do not successfully complete the portfolio assessment process, the SCDE must prepare to defend its use of the portfolio in terms of all of the above requirements 1-6 and show why no alternatives exist or are offered to the protected classes.

8	The process must be implemented and monitored to ensure reliable scoring and to provide for adequate candidate support.	Tests of reliability must be performed and samples of candidate work and faculty scoring must be reviewed on a regular basis to ensure that procedures and scoring are not “drifting” and to minimize measurement error. Raters need to be trained and updated on a regular basis. Directions need to be clear. Portfolios across candidates need to be comparable in difficulty. Rater mood and fatigue need to be carefully monitored. Safeguards against cheating need to be implemented. The sufficiency of items in the portfolio must be adequate. Records should also be kept of all exceptions made, alternatives provided, due process proceedings, and faculty/candidate training.

Do Portfolios Have a Place in Teacher Training and Certification?

Portfolios remain an excellent assessment device to support learning. Questions raised in this article relate to the use of portfolios for summative certification decisions for all or most standards combined, especially when contents vary widely. When contents are the same across students, then questions can be raised about what purposes the portfolios actually serve. Is a checklist enough to determine if all work is completed satisfactorily? If so, could some other type of tracking system be used that provides less burden on both faculty and students? If the reflective aspect is considered essential, are there other forms of reflection that might serve equally well, such as a professional development plan? The professional development plan would be a job-related task in any state where districts require teachers to develop such plans. It is a widely accepted, and research supported, view that teachers who identify their own strengths and weaknesses as well as those of their students are better practitioners than those who do not do so and, typically, teachers participate in professional development planning and activities for improvement purposes in most states and school districts.

There are also some instances in which portfolios can be used to assess specific skills that have been accepted as critical to effective teaching. These instances can help to differentiate between the competent and the incompetent teacher and are job-related. For example, a portfolio of K-12 student work, used to assess the extent to which a teacher candidate can teach students to think critically and creatively would be an appropriate “test.” This is clearly a job-related task, since the teacher is required in most states and school districts to demonstrate that children are learning.

These authors are suggesting a more limited and focused use of portfolios – portfolios to measure specific, job-related skills. By limiting the use and complexity of portfolios, the long known values of portfolio assessment can be realized without burdening faculty and students with excessive requirements that have limited use and without taking serious psychometric and legal risks.

Conclusions and Implications

The shift of responsibility from state departments of education to teacher preparation programs has increased the likelihood that SCDEs will face legal challenges when candidates are denied diplomas and certification/licensure based on the tests used in the academic program. Particularly vulnerable are the cumulative or showcase portfolios currently being required in many SCDEs as "evidence" of candidate demonstration of standards and competency. When these portfolios are used as a measure of job performance themselves, or when they are evaluated using criteria that are related in only tangential ways to authentic job tasks, or when they are not substantially related to standards required for state program approval, or when they are prepared as an extra-curricular activity, or when they contain student-selected evidence, or when they are not adequately monitored for reliability or bias, the threat of litigation increases as the SCDE's fail to pay attention to psychometrics.

New standards make psychometric qualities more important than ever to avoid challenges…more than ever before. High-stakes testing has informed an army of students and lawyers to the details of tests, so it is easier to sue. To avoid litigation, SCDEs must carefully consider the design and implementation of portfolios and should consider a heavier reliance on individual tasks that are combined in a way that leads to an appropriate decision or cut score that differentiates between candidates who are likely to be competent teachers and those who are not. The use of key course and internship-embedded tasks that measure critical skills, that are reviewed to ensure that they are representative and relevant job-related measures of the domains, that are evaluated by the faculty who assign them, that are tracked through student records (paper or electronic), and that are combined in meaningful ways to establish which candidates are likely to be competent teachers hold far better promise of satisfying the psychometrics and keeping the big and little "children" safe in both their university and K-12 classrooms.

References

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

Arter, J. and V. Spandel (Spring 1992). Úsing Portfolios of Student Work in Instructional Assessment." *Educational Measurement: Issues and Practice* 11, 1: 36-44. (cited in Herman, et al., 1992).

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.

Dollase , Richard H. (1998) When the State Mandates Portfolios: The Vermont Experience. In Lyons, N. (Ed.) (1998). *With Portfolio in Hand: Validating the New Teacher Professionalism.* Teachers College Press, New York NY.

Equal Employment Opportunity Commission, 1978. *Uniform Guidelines on Employee Selection Procedures*. Washington, D.C.

Gearhart, Maryl and Herman, Joan L. (1995). Portfolio assessment: Whose work is it? Issues in the use of classroom assignments for accountability.

Herman, Joan L.; Aschbacher, Pamela R.; Winters, Lynn (1992). *A Practical Guide to Alternative*

Assessment. Association for Supervision and Curriculum Development, Alexandria, VA.

Hazi, Helen M. (1989). Measurement versus supervisory judgment: The case of Sweeney v. Turlington, *Journal of Curriculum and Supervision*. Spring, 1989, 4(3), 211-229.

Ingersoll, Gary M. and Scannell, Dale P. (2002). *Performance-Based Teacher Certification: Creating a Comprehensive Unit Assessment System.* Fulcram Publishing, Golden, CO.

Kane, Michael (1994). Validating the performance standards associated with passing scores, Review of Educational Research, Fall 1994, 64(3), 425-461.

Koretz, Daniel (1994). The Evolution of a Portfolio Program: The Impact and Quality of the Vermont Portfolio Program in Its Second Year (1992-1993). Report from the National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA. Office of Educational Research and Improvement, Washington, D.C.

Lee, William W. and Owens, Diana L. (April 2001). Court Rulings Favor Performance Measures, *Performance Improvement*. 40(4).

Lemke, June C. (2002). Preparing the best teachers for our children. In: NO Child Left Behind: The Vital Role of Rural Schools. Annual National Conference Proceedings of the American Council on Rural Special Education (ACRES). 22nd, Reno, NV, March 7-9, 2001.

McDonough, Matthew, Jr. and Wolf, W.C., Jr. (1987). Testing teachers: Legal and psychometric considerations. *Educational Policy.*

McGinty, Dixie (1996). The demise of the Georgia Teacher Performance Assessment Instrument, *Research in the Schools,* 3(2), 41-47.

Mehrens, William A. and Popham, W. James (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education*, 5(3), 265-283.

Mellnick, Susan and Pullin, Diana (2000). Can you take dictation? Prescribing teacher quality through testing. *Journal of Teacher Education* 51(4), 262-275.

National Council for Accreditation of Teacher Education (2000). Porfessional standards for the Accreditation of schools, colleges, and departments of education. NCATE, Washington, D.C.

Moss, Pamela (1998). Rethinking validity for the assessment of teaching. In Lyons, N. (Ed.) (1998). *With Portfolio in Hand: Validating the New Teacher Professionalism.* Teachers College Press, New York NY.

Nweke, Winifred and Nolan, Juanie (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 (48th, Chicago, IL, February 21-24, 1996.

Pascoe, Donna and Halpin, Glennelle (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 (30th, Little Rock, AR., November 13-16, 2001).

Pullin, Diana C. (2001). Key questions in implementing teacher testing and licensing, Journal of Law and Education, 30(3), July 2001, 383-429.

Rebell, Michael A. (1991). Teacher performance assessment: The changing state of the law, *Journal of Personnel Evaluation in Education*. 5:227-235.

Sandman, Warren. (1998). Current Cases on Academic Freedom. Paper presented at the annual Meeting of the National Communication Association. New York.

Salzman, Stephanie A.; Denner, Peter R.; Harris, Larry B. (2002). Teacher Education Outcomes Measures: Special study survey. American Association of Colleges of Teacher Education, Washington, D.C.

Sireci, Stephen G. and Green, III, Preston, C. (2000). Legal and psychometric criteria for Evaluating Teacher Certification Tests, *Educational Measurement: Issues and Practice*. 19(1), 22-24.

Stiggins, Richard J. (2000). *Specifications for a Performance-Based Assessment System*, NCATE Web Site, on-line: http://www.ncate.org/resources/commissioned%20papers/stiggi ns.pdf

Wilkerson, Judy and Lang, William Steve (January 2003). *Analysis of Performance Assessment Survey of Teacher Preparation Institutions.* Survey analysis prepared for the Florida Department of Education, Tallahassee, Fl.

Wilkerson, Judy; Lang, William Steve; Egley, Robert; Hewitt, Margaret (January 2003). *Designing Standards-Based Tasks and Scoring Instruments to Collect and Analyze Data for Decision-Making.* Workshop presented at the annual meeting of the American Association of Colleges of Teacher Education in New Orleans, LA..

Wilkerson, Judy (2000). Program accountability for beginning teachers' subject matter knowledge and competency and how to meet the challenge: A state's perspective on program accountability for teacher education graduates' competency. Symposium paper presented at the Annual Meeting of the American Association of Colleges of Teacher Education, Chicago, IL.

Zirkel, Perry A. (June 2000) Tests on trial, *Phi Delta Kappan*, 81(10), 793-4.

About the Authors

Judy Wilkerson and **William Steve Lang** are on the faculty at the University of South Florida St. Petersburg. Both teach courses in assessment and research. His research interests include the Rasch model and performance assessment. Her interests are evaluation and accreditation standards.

Email: wilkerso@tempest.coedu.usf.edu & wslang@tempest.coedu.usf.edu

---

The World Wide Web address for the *Education Policy Analysis Archives* is
epaa.asu.edu

EPAA is published by the Education Policy Studies Laboratory, Arizona State University