DOCUMENT RESUME

ED 337 487 TM 017 310

AUTHOR Mehrens, William A.
TITLE Issues in Teacher Competency Tests.
SPONS AGENCY California Univ., Berkeley. Graduate School of Education.
PUB DATE 87
NOTE 112p.; Revision of "Validity Issues in Teacher Competency Tests" prepared for the Institute for Student Assessment and Evaluation at the University of Florida.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC05 Plus Postage.
DESCRIPTORS *Competence; Construct Validity; Elementary Secondary Education; *Licensing Examinations (Professions); Occupational Tests; *Teacher Certification; Teacher Evaluation; Teacher Improvement; Teacher Qualifications; Test Use; *Test Validity
IDENTIFIERS *Teacher Competency Testing

ABSTRACT

The types of validity evidence that are appropriate for teacher licensure tests are discussed. Included are a discussion of the current popularity of teacher competency tests and the reasons for that popularity. Licensure, certification, and employment examinations are differentiated. Because the area of interest is the minimum competency necessary to prevent harm from coming to the clients, it is argued that content validity is the type of validity evidence that is most appropriate for licensure tests. Criterion-related validity, construct validity, and "curricular validity" are also discussed. It is concluded that teacher licensure tests allow valid inferences for a delimited set of inferences. An effective teacher licensure test will not eliminate the need for subsequent teacher evaluation; it will not cure all educational ills; it will not eliminate all ineffective teachers. Nevertheless, it should ensure that individuals who are licensed have a minimal level of competence on some important subdomains of knowledge and skills relevant to their profession. A 162-item list of references is included. (Author/SLD)

ISSUES IN TEACHER COMPETENCY TESTS

William A. Mehrens

Michigan State University

Prepared for the Commission on Testing and Public Policy
Graduate School of Education
University of California, Berkeley
(This is an expansion, a revision, and an update of a paper entitled Validity Issues in Teacher Competency Tests prepared for
The Institute for Student Assessment and Evaluation,
The University of Florida.)

1987

Running head: ISSUES

Table of Contents

ISSUES IN TEACHER COMPETENCY TESTS

William A. Mehrens
Michigan State University

Abstract

This paper presents a brief introduction which includes a discussion of the current popularity of teacher competency tests and the reasons for that popularity. Following that is a section differentiating between licensure, certification and employment examinations. The main portion of the paper discusses the types of validity evidences that are appropriate for teacher licensure tests.

Because the inference of interest has to do with the minimum competency necessary to prevent harm from coming to the clients, it is argued that content validity is the type of validity evidence most appropriate for licensure tests. However, evidences of criterion-related validity, construct validity and "curricular validity" are also discussed.

It is concluded that teacher licensure tests allow valid inferences for a delimited set of inferences. An effective teacher licensure test will not eliminate the need for subsequent teacher evaluation; it will not cure all educational ills; and it will not eliminate all ineffective teachers. Nevertheless, it should ensure that those individuals who are licensed have a minimal level of competence on some important sub domains of knowledge and skills relevant to their profession. That is a step in the right irection.

ISSUES IN TEACHER COMPETENCY TESTS

> Scott wont pass in his assignment at all, he had a poem to learn and he fell tu do it (Time, 1980).

> Validity has long been one of the major deities in the pantheon of the psychometrician. It is universally praised, but the good works done in its name are remarkably few (Ebel, 1961, p.640).

It seems so simple. We do not want incompetent teachers. Tests should be able to weed out teachers with skills at a level such as that demonstrated in the first quote above. Why should we be concerned whether the deity of validity has done good works in teacher competency tests? Why should we let those psychometricians who believe in the deity criticize our attempts to upgrade the teaching profession?

But many scholars, college professors and other educators (partially overlapping domains) believe that things are not always as simple as they seem. What is "teacher competency?" How do we know whether tests really measure it? Such questions should be, and have been, asked. This paper is intended to take a close look at several issues regarding the validity of teacher competency tests. A general conclusion of the paper is that such tests are valid for a delimited set of inferences.

Introduction

In this section the stage for discussing the validity of teacher competency tests is set through the consideration of some background information. Some surveys on the current popularity of teacher competency tests are discussed as well as the motivating factors behind that popularity. A brief discussion of the traditional certification requirements and their limitations follows. A short section on defining and assessing teacher competence and a section delimiting the focus of this paper comprise the last two sub-sections of this introduction.

Current Popularity of Teacher Competency Tests

Teacher certification tests are not new. They were first officially endorsed in 1686 (Vold, 1985) and administered as early as the 18th Century (Carlson, 1985). However, they are currently enjoying a revival. This revival began in 1964 when North Carolina required prospective teachers to take the National Teacher Examinations (NTE). Because the movement is spreading so rapidly, it is difficult to be current in any reporting. Until 1977, North Carolina was the only state that required such an examination. However, by October of 1983, 28 states required (or planned to require within the year) the passing of a test as one of the requirements for an initial professional teaching certificate. Twelve additional states were contemplating such a teacher competency examination (Lehmann and Phillips, 1985). Goertz (1986) reported that by 1987, 32 states will require

> "aspiring teachers to pass a state-prescribed, standardized test before entering a teacher education program and/or before being certified to teach" (p.17).

Seventeen of the states require prospective teacher education candidates to pass a test before entering a teacher education program and 29 states have a testing requirement for certification.

Most of the certification testing is being done using the National Teachers Examinations (NTE) or the Pre-Professional Skills Test (PPST) developed by the Educational Testing Service (ETS); or tests developed by National Evaluation Systems (NES). The NES tests are custom built for each state to match the certification areas and other specific state requirements. Certification tests cover basic skills in 21 states, general knowledge in 12 states, professional knowledge in 16 states, and the teacher's specialty area in 20 states (Goertz, 1986, p. 21).

The very rapid spread of teacher testing programs is politically based and supported by the public. Gallup polls (1984) indicate that 89 percent of the public (and 63 percent of the teachers) believe that teachers "should be required to pass a state board examination to prove their knowledge in the subjects they will teach." [Gifford, 1986, suggests t at the teacher support "may have been engendered by a defensive reaction to public fears of teacher incompetence rather than by confidence in testing and in their own professionalism" (p. 252).] An Educational Research Service (ERS) poll of teachers and principals indicated that 82% of the teachers and 86% of the principals agree that new teachers should be required to pass exams in their subject areas. About 75% of both teachers and principals also believe new teachers should be tested on their knowledge of teaching methods (Newsnotes, 1984). Albert Shanker, the president of the American Federation of Teachers-AFL-CIO, has called for a national teacher examination for licensing new teachers. He has stated that

> "There must be a grueling national (but not governmental) teacher examination ... that will inspire fear and loathing among those about to be examined---exactly the way the bar and medical examinations do---and that will, by the very degree of its difficulty, challenge the best and the brightest among us to take it, to pass it, and to teacl. in our schools" (Shanker, 1985, pp 28-29).

Historically, the National Education Association had been opposed to such exams because they claimed such tests do not measure teaching ability or guarantee that those who pass will be good teachers. However, in 1985 Keith Geiger, vice president of the National Education Association, said his organization would support such exams to "upgrade the profession" (Feinberg, 1985). Mary Futrell, the current president of the NEA, recently indicated that she would press the union for an endorsement of a certification test. "We have to face reality. You can't run away from or ignore the fact that the tests are being used and will continue to be used as part of the pre-

certification requirement" (Cornell, 1985). At the 1985 convention the NEA delegates did support teacher competency tests and called

> "for 'rigorous' standards for new teachers, including 'valid and unbiased' tests. But the proposed resolution stopped short of saying there should be a single national certification exam...The NEA jettisoned language it adopted two years ago saying 'no single criterion should be used' to determine who can teach as well as the 1984 resolution which said the NEA was 'convinced that no test in existence is satisfactory for such usage'" (A.P., 1985).

The teachers unions, as of the time of this writing, do not support the testing of already certified teachers for purposes of recertification. However, a few states (Arkansas, Georgia, Texas) require such testing and a recent Gallup poll shows that 85% of the public believe experienced teachers should periodically be required to pass a statewide basic competency test in their subject areas (Gallup, 1986).

The recent Holmes (Holmes Group, 1986) and Carnegie (1986) reports on teaching both support examinations for prospective teachers. The Holmes Group report recommends that all prospective teachers, prior to certification, should pass a written test in each subject they will teach.

> "This exam should test for their understanding of the basic structure of the discipline, and tenets of a broad liberal education" (Holmes Group, 1986, p. 11).

In addition, the Holmes Group report recommends a test of reading and writing ability and a test of pedagogy. These tests "should be sufficiently difficult so that many college students could not pass" (p. 11).

The Carnegie report suggests that a National Board for Professional Teaching Standards be created to establish high standards and to certify teachers who meet those standards.

Why Teach Competency Tests?

The motivating factor behind teacher competency tests is that the public believes that some of our teacher-training institutions have admitted

low quality candidates. These institutions have granted diplomas to, and states have certified, teachers who are not minimally competent. The public believes that our colleges and our state licensing boards both have failed as gatekeepers. There is considerable evidence for those beliefs. For example, Feistritzer reported that

> "Never before in the nation's history has the caliber of those entering the teaching profession been as low as it is today" (1983,p.112).

In speaking of the results of research just completed for the National Center for Educational Information, Feistritzer was quoted as follows:

> "The certification of classroom teachers in the U.S. is a mess. There are far too many colleges where a student can show up with a high-school diploma and a checkbook and get out with a bachelor's degree in education" (U.S. News, 1984, p. 14).

She went on to say that one third to one-half of the colleges operating teacher-training programs "ought to be shut down."

The Committee for Economic Development stated that:

> "Right now too many students entering college programs leading to teaching careers are among the lowest achieving graduates of U.S. high schools. We are preparing our poorest students to be our future teachers, and this situation must be changed" (quoted from the Carnegie Task Force, 1986, p. 36).

Pugach and Raths stated that "Teacher education programs in the United States traditionally have not played a significant role in preventing unqualified persons from becoming certified to teach" (1983, p. 39) They cited several reasons why the gatekeeping function is typically absent in programs of teacher education including the threat of suit and professors' unwillingness to "play God." Further, they suggested the previous teacher shortage "has encouraged what is essentially an open admissions and graduation policy" (1983, p. 39). They reported a study by Southall (1982) which found that for student teaching grades in 34 institutions, 79% of the students given traditional letter grades received As and 18% received Bs. "People who advocate competency testing of teachers base their position in

part on the traditional reluctance of teacher educators to do their job properly" (1983, p. 39).

Weaver (1980) reported data showing that college of education students score lower on academic ability tests than do other college students. For example, in a 1975-76 study of college freshmen ACT scores, education majors were tied for seventeenth place on math scores and fourteenth on English scores among 19 fields of study which were compared. A study by the National Center for Educational Statistics (1982) reported that education majors ranked fourteenth out of 16 fields on SAT verbal scores. Only students concentrating in ethnic studies or in trade and vocational education scored lower. As Weaver suggested, "teacher education is the field that shows the least selectivity, from college-bound applicant to completion of degree, among the programs for which comparable data are available" (1980, p. 15).

Shanker and Ware reported the results of a study "which showed that education majors at Virginia's state Universities scored an average of 121 points lower on the SAT than did those who received bachelor's degrees in other fields" (1982, p.7). Because the academic quality of the students entering a program is obviously related to the academic quality of students leaving a program, data such as those mentioned in the previous paragraphs on the academic ability of students in colleges of education are of concern to many individuals.

An earlier Time Magazine article (1980) reported that twenty percent of all teachers had not mastered the basic skills they were supposed to teach. In 1978, the Dallas Independent School District gave the Wesman Personnel Classification Test (WPCT) to 535 first year teachers and a volunteer group of high school juniors and seniors. The students out-performed the teachers and more than half the teachers fell below the score considered acceptable

by the district. On a teacher competency test in Houston, job applicants scored lower than high school juniors in mathematics achievement (Benderson, 1982).

Hathaway quoted from a 1979 editorial in The Washington Post

> "In the District, public school teachers have been hired for years on the basis of their college records and interviews. Most are graduates of ... teacher's college, which in 1977 permitted two students to graduate even though they had failed basic math courses. One of the graduates could not add fractions such as 3/4 plus 1/3. Faculty members said incompetent students had been slipping through (the college) and going on to teach in the city's public schools for 10 years.
> "Something must be done now before children are made mental cripples. (The) Superintendent is considering a requirement to have new teachers pass a test of academic skills.. He should" (From Civille Right, 1979; quoted from Hathaway, 1980).

There is at least one state where some college graduates score at the chance level on the state's teacher competency test! This finding is probably generalizable to many other states. Sykes (1983) has said that teacher education has become an intellectual ghetto.

There is simply no doubt that schools of education often lack rigorous admission standards. Further, few students fail once they are admitted (See also Goertz, Ekstrom, and Coley, 1984). In addition, Feistritzer (1984) found that 100 new teacher education programs had been initiated in the previous ten years. Most of these were in small private colleges with few standards.

[To be fair, we should point out that low admission standards are not limited to colleges of education. See Burns (1985) for a biting commentary on the low standards in post-secondary education and the reasons for those low standards.]

A few educators may discount, or perhaps even support, the deplorable standards (arguing that love, patience, compassion etc. are the important criteria to be a teacher).

> "One of the favorite ploys of some teacher colleges and indeed, of some administrators, is that 'humanistic' education is more important than knowledge. What has never been explained is why the teacher who knows his subject is automatically a 'dehumanizing' factor in the classroom while the one who doesn't is somehow 'more sensitive to the needs of children.' I speak no hyperbole. I've heard those very words" (Hilldrup, 1978, p. 28).

The public and professional educators interested in reform, however, do not support low standards. They are dismayed that some teachers communicate with parents in the style quoted at the beginning of this manuscript. They are dismayed that not all elementary school teachers have mastered elementary school arithmetic. The public (and almost all educators) believe that teachers should be able to read, write, and do simple arithmetic. Most would accept the reasonable assumption that you can not teach what you do not know; that if you are to teach the basics you should know them.

> 'We cannot improve the quality of education in our schools without improving the quality of the teachers in them" (Holmes Group, 1986, p. 23).

> "Without a profession possessed of high skills, capabilities, and aspirations, any reforms will be short lived" (Carnegie Task Force, 1986, p. 2).

> "Teaching is, at its root, a learned profession. A teacher is a member of a scholarly community" (Shulman, no date, p. 12).

But are examinations necessary to establish that applicants for a teacher certificate know the basics? Why not rely on colleges of education or certification agencies? Because the traditional approaches have not worked. The problems with admissions requirements and standards within colleges of education have already been discussed briefly. Could not regulatory agencies correct these problems? They could, at least in part, but there are problems inherent in non-examination procedures. Consider the traditional approach of program approval.

Traditional certification requirements

> "Forty-six states now rely at least in part upon program approval. Under this mechanism, the state establishes guidelines for acceptable teacher education programs, evaluates programs for compliance with those guidelines, and automatically accepts the credentials of graduates of approved programs. While program approval gives the state more control over the content, quality, and philosophy of the academic preparation of prospective teachers than do course completion requirements, it tends to shift the focus of attention away from the competence of the individual applicant to the quality of the program" (Eisdorfer and Tractenberg, 1977, p. 111).

In addition to the shift of focus away from individual competence, there are several problems inherent in the use of program approval to ensure quality control. Gubser discussed many of these at length and suggested that "The program approval process has in most states proved less than satisfactory" (Gubser, 1979, p. 12). Scriven helped explain why:

> ...virtually all state certification systems (are) extremely favorable to the programs under review, in the sense that even if criticisms are raised and officially sanctioned as "considerations" they are always removed before any actual suspension of the right to credential occurs. Now it might be the case that this reflects a high level of performance by the programs, but that there should never be a failure stretches credibility. ...when we consider...that political pressure makes it impossible for a political commission made up largely of representatives of the very institutions that are accredited to actually take punitive action, we must obviously be cautious about assuming the worst" (Scriven, 1979, p. 1).

Even if program approvals were not subject to political considerations, there is no compelling reason to believe they would fulfill their purpose of protecting the interests and welfare of the public. As Freeman pointed out;

> "In general, the development of certification requirements appears to have been dictated, to a large extent, by the intuitive notions of 'what a teacher or guidance counselor needs to know' and then using available higher education categories to express the requirement. One might well make out a case that an elementary teacher should have a general knowledge of mathematics. As expressed in rules and regulations, this intuitive judgment becomes 'four hours of mathematics.'" (1977, p. 75).

It is ironic to note that some of the critics of current examinations suggest they are based on inadequate job analyses. What about the course

requirements, or the general program requirements established by certification boards? Where are the job analyses that determined "four hours of mathematics" gives elementary teachers sufficient knowledge of mathematics?

Defining and Assessing Teacher Competence

Because having a college degree from an approved program does not guarantee competence, there needs to be some additional gatekeeping function, some additional method of assessing teacher competence.

> "One major problem inherent in teacher evaluation is that there is no clear definition of what characterizes an effective teacher or constitutes effective teaching and, consequently, no definitive measures to be used for teacher evaluation" (Webb, 1983, p. 3).

Further, not all writers differentiate between the quality of the teacher, the quality of the teaching, and the outcomes of the teaching (Darling-Hammond, Wise, & Pease, 1983). Medley (1982) presented the following useful definitions of four terms that others have treated as synonyms:

Teacher competency: Any single knowledge, skill, or professional value position, the possession of which is believed to be relevant to the successful practice of teaching. Competencies refer to specific things that teachers know, do, or believe but not to the effects of these attributes on others.

Teacher competence: The repertoire of competencies a teacher possesses. Overall competence is a matter of the degree to which a teacher has mastered a set of individual competencies, some of which are more critical to a judgment of overall competence than others.

Teacher performance: What the teacher does on the job rather than what she or he can do. Teacher performance is specific to the job situation; it depends on the competence of the teacher, the context in which the teacher works, and the teacher's ability to apply his or her competencies at any given point in time.

Teacher effectiveness: The effect that the teacher's performance has on pupils. teacher effectiveness depends not only on competence and performance, but also on the responses pupils make. Just as competence cannot predict performance under different situations, teacher performance cannot predict outcomes under different situations.

Generally, the definitions of the competency tests designed for teachers are much like the definition Medley used for teacher competency. For example, the Alabama Board stated their test was "to measure the specific competencies which are considered necessary to successfully teach" (Alabama State Board of Education, 1980).

Considered and necessary are the two key words in that statement. "Considered" suggests, correctly, that the decision is a professional judgment and "necessary" suggests that the competency is not sufficient. Critics of the Alabama test tried to suggest that the phrase "to successfully teach" implied that the test should have criterion-related validity and that successful is a matter of degree that can be measured along a continuum among those who are qualified. Experts for the defense felt more comfortable with interpreting successful as above a minimum cut score, and that the "considered necessary" portion of the definition indicated the Board did not expect to find criterion-related validity.

Shulman (no date) suggests that the kinds of tests given in most states trivialize teaching, ignoring its complexities and diminishing its demands. In two very fine papers (no date, 1986) he points out the complexities of teaching, the extensive knowledge base for teachers, and the importance of understanding the content of the subject matter.

> "The teacher need not only understand <u>that</u> something is so; the teacher must further understand <u>why</u> it is so, on what grounds its warrant can be asserted, and under what circumstances our belief in its justification can be weakened or even denied" (Shulman, 1986, p. 9).

This is a legitimate expectation for high quality teachers. Unfortunately, it may be a bit unrealistic as a standard for initial certification.

Focus of this paper

This paper is limited to the assessment of teacher competencies. That is, with the approaches that measure a set of those single things each of which Medley refers to as a "teacher competency." These approaches, of course, do not attempt to measure the total repertoire of competencies a teacher possesses so following Medley's definitions, they are not measures of "teacher competence." While one might consider using teacher performance or teacher effectiveness as criteria for teacher competency tests, the measurement of those two characteristics will not be discussed except somewhat indirectly in this paper.

One could test for teacher competencies at various times and for various purposes. Testing for basic skills prior to entry into teacher education programs is one method of assuring some teacher competencies. If only students who have mastered the basic skills are allowed into teacher programs, we obviously do not need to worry about teachers not having those competencies. One could also test for teacher competencies prior to graduation from a teacher education program. At this point one could test for basic skills, knowledge of the subject matter the individual wishes to teach, and/or knowledge of pedagogy. While such assessment programs have merit, this paper is limited to the issues of teacher competency tests that are given for certification (licensure) decisions.

The various purposes of teacher competency tests can also be subdivided into formative evaluation purposes and summative evaluative purposes. There is debate in the literature about whether these two purposes can be accomplished within the same evaluation system. This paper will be limited to the summative role of the assessment although it is not meant to suggest that an evaluation could not contain some information that could be considered of diagnostic value for those who wish to use the data as a basis for improvement.

Thus, this paper is limited to examining the validity issues of competency tests used for assessment by licensing agencies. Tests that colleges might wish to use for either entrance or exit purposes are not considered. Tests used for employment purposes are not considered. Further, measures of teacher performance or measures of teacher effectiveness (except for the role they may play in evaluating the validity of the teacher competency tests) are not considered.

Licensure, Certification and Employment Exams

The terms licensure and certification have been used interchangeably by some individuals in education and it is not always clear to educators how employment exams differ from the other two types. But both the legal and psychological professions have made distinctions among the three terms. Thus, some definitions and explanations are in order.

The United States Department of Health, Education, and Welfare defined licensure as follows:

> Licensure: The process by which an agency of government grants permission to persons to engage in a given profession or occupation by certifying that those licensed have attained the minimal degree of competency necessary to ensure that the public health, safety and welfare will be reasonably well protected (1971, p. 7).

The same agency defines certification as follows:

> Certification: The process by which a nongovernmental agency or association grants recognition to an individual who has met certain predetermined qualifications specified by that agency or association. Such qualifications may include graduation from an accredited or approved training program, acceptable performance on a qualifying examination, and/or completion of some specified amount or type of work experience (1971, p.7).

One of the major distinctions in the two definitions is whether or not the agency is governmental or nongovernmental. Because, historically the "certification" of teachers has been done typically by a governmental agency, what the public has typically called teacher certification

requirements are actually licensure requirements. Given the call for a National Board for Professional Teaching Standards by the Carnegie Task Force (1986) and a national exam by the Holmes Group Report (1986) in the future there may well be a national, nongovernmental examination properly called a certification examination.

A second distinction is that licensing is a mandatory program designed to protect the public from incompetents. It is a selecting-out process. Licensure procedures are to determine whether or not individuals have minimal competence. Certification is typically voluntary and grants special status to the individuals. It is a selecting-in process. Certification typically goes beyond the minimum requirements. (The type of examinations Shulman and Shanker advocate would not appear to be minimal.) Hecht stated that: "I believe teaching certification to be a misnomer, ... because it is a legal requirement to begin teaching, to protect the public from incompetent teachers, and signifies no special standing within the profession" (1979, p. 17). However, as Shimberg pointed out, the traditional distinction of minimum competence for licensure and well-beyond the minimum competence for certification has become blurred (1984). The report of the National Commission for Health Certifying Agencies states that

> "These perceived differences obscure the common underpinning of these two regulatory mechanisms--namely, agreement that the public has the right to services from qualified practitioners..." (NCHCA, 1980, p. 4).

Thus, although there are distinctions in the definitions of the two words and these distinctions would suggest both different purposes as well as different properties of the examinations, the use of the phrase "teacher certification" probably is not too misleading. However such programs as will be discussed in this paper are, in fact, state licensure programs. Their purpose is to protect the public from incompetents.

Employment tests ususally have a quite different purpose from licensure tests. Employment tests typically are intended to help identify those applicants for a job who are likely to be the most successful. Whereas licensing exams are designed to further the states' interests, employment exams are intended to further the employers' interests.

The AERA/APA/NCME Standards for Educational and Psychological Testing clarify the differences in a succinct manner:

> "For licensure or certification the focus of test standards is on levels of knowledge and skills necessary to assure the public that a person is competent to practice, whereas an employer may use tests in order to maximize productivity" (AERA/APA /NCME, 1985, p. 63).

[1]

In addition the Standards state that

> "Whereas employment tests may measure appropriately an individual's aptitude to learn a specific job, people who take licensure or certification tests have usually completed training and are seeking to be deemed qualified for a broad field, rather than for a specific job. This distinction has important implications for the content to be covered in licensing or certification tests" (1985, p. 64).

Because employment and licensure examinations serve different purposes, they may well be constructed somewhat differently. Whether or not the examinations differ, because we make different inferences from the scores of examinations used for employment and licensure, the kinds of validity evidence gathered to support their uses should differ. Although the focus of this paper is on licensure examinations, many people confuse validity requirements of the two types of examinations so the different requirements will be discussed in more detail at various points in the rest of this paper.

Prior to discussing further the issues of validity in teacher certification tests a short overview of validity in general is presented.

Validity: Some General Notions

The AERA/APA/NCME Standards state that validity

> "refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences" (1985, p. 9).

Although, as the Standards point out, validity is a unitary concept, evidence may be accumulated in many ways. Traditionally, psychometricians have categorized the various types of validity evidence into content-related, criterion related, and construct-related evidence of validity although "rigorous distinctions between the categories are not possible" (p. 9).

Construct-related validity evidence "focuses primarily on the test score as a measure of the psychological characteristic of interest...Such characteristics are referred to as constructs because they are theoretical constructions about the nature of human behavior" (p. 9). "In general, content-related evidence demonstrates the degree to which the sample of items, tasks, or questions on a test are representative of some defined universe or domain of content" (p. 10). "Criterion-related evidence demonstrates that test scores are systematically related to one or more outcome criteria" (p. 11).

The lack of a rigorous distinction among the categories of validity evidence is especially true between the catergories of content and construct validity evidence. As Tenopyr suggested, "the confusion between content validity and construct validity runs rampant in psychology today" (1977, p. 47). The distinction she preferred is that content validity deals with inferences about test construction whereas construct validity involves inferences about tests scores. Others such as Guion (1977) and Messick (1975) would agree with her. Although Cronbach (1980) referenced Guion, Messick, and Tenopyr as if he agreed with them, he worded the point quite

differently. As he stated, "content validity is established only in test construction, by specifying a domain of tasks and sampling rigorously. The inference back to the domain can then be purely deductive" (Cronbach,1980, p. 105, emphasis added). This wording holds more appeal to me. We do make deductive inferences from the score on the test, which is a sample of items from the domain, to the domain. The defense of this inference from a score on a sample to a score on a domain is contingent on the test construction process which includes domain specification and item sampling.

The Uniform Guidelines also make reference to the "borderline between content validity and construct validity" although that document suggests that, at the extremes, the two types of validity evidence are quite "easy to understand" (EEOC, 1978, 38292).

Ebel, in commenting on some of the difficulties in test validation suggested that test specialists are the source of some of the problem. Quoting Bishop Berkeley he suggested that "We first raise a dust, and then complain that we cannot see" (1977, p. 55). He preferred to call the process that we typically call content validity a content reliability or job sample reliability. But in so doing, he was not suggesting we should gather any other type of "validity evidence" when the process of measurement itself defines the thing to be measured. He suggested in those cases that "the question of validity need never arise" (p. 57).

The point of all of this is that measurement specialists do not all use the terms the same way. This is unfortunate, but not incapacitating. In this paper the words will be used in what might be called the "traditional" sense. If some type of evidence described under content validity evidence seems more to some reader like construct validity evidence or vice versa, that reader is surely capable of handling the internal translation he/she must engage in to comprehend the discussion.

The terms curricular validity and instructional validity are being used increasingly in the educational measurement literature. While many would suggest that these terms are not categories of validity (and they are not in the index of the new Standards), they do have some relevant meaning. "Curricular validity relates to the question of the degree to which the test content is covered in the curriculum materials. Instructional validity is a more restrictive term and relates to the degree to which the test content is actually taught" (Mehrens & Lehmann, 1987, p. 83). Instructional validity is important if one wishes to make inferences about instructional effectiveness (i.e. did individuals learn what they had been taught). Curricular validity is important for minimal competency tests required for high school graduation. It is generally considered irrelevant in judging the quality of licensure examinations. Reasons for this will be discussed later.

Some individuals have suggested that curricular validity should be a sub-category of content validity. In fact, an appellate court ruling on the Debra P. case stated that "an important component of content validity is curricular validity" (Debra P., 1981, p. 6770). In general, measurement specialists have not adapted the court's use of the term. As mentioned, the new Standards do not even index the term. Yalow and Popham (1983) argued that instructional/curricular validity issues are really issues regarding the adequacy-of-preparation for a test. As they stated, "adequacy-of-preparation is not a component of content validity. Not only is it not a component of content validity, it is not a form of validity at all" (p. 12).

Inferences From Teacher Competency Tests

As mentioned, validity refers to the inferences made from test scores. Before discussing the kinds of validity evidence needed for teacher competency tests it is necessary to consider what inferences we wish to make from the scores. It seems important to distinguish the inferences the test

builders and test users wish to make from the inferences that others may draw (or claim you cannot draw) from the scores. The builders and users of tests have a responsiblity to gather evidence (or use logic) to support their particular inferences. In the process of doing this they may use logic or evidence to rule out the plausibility of some potentially competing inferences. However, they do not have any responsibility to gather evidence to support (or refute) all inferences others may make (or claim can not be made) from the test scores. This point needs to be stressed because a common method of attacking the use of tests is to state that there is no evidence that the scores predict some variable that the users/builders never intended the scores to predict. For example, some educators attack teacher competency tests used for licensure purposes because the passing of such tests does not guarantee one will be a good teacher. As Mehrens and Lehmann pointed out,

> "That, of course, is true but totally irrelevant. (One wonders if such an argument is not evidence for a need for a minimum competency test in logic!). The tests are not designed to be predictive among the competent or to ensure that all certified teachers will be good teachers" (1984, p. 582).

This procedure of attacking a test because its scores do not measure something they were not intended to measure has been recognized for decades. In 1946, Rulon stated that

> "Validity is usually described as the extent to which a test measures what it is purported to measure. This is an unsatisfactory and not a very useful concept of validity, because under it the validity of a test may be altered completely by arbitrarily changing its 'purport'" (1946, p. 290).

Some individuals have been known to criticize tests of teacher subject matter or pedagogical knowledge because they do not measure love, warmth, compassion or some other characteristic just as, a few years ago, some individuals criticized intelligence tests because they did not measure motivation. It should not take too much sophistication in measurement to recognize that a test designed to measure one variable should not be

criticized for not measuring another! Wood made this point over 45 years ago:

> "To abandon examinations of intelligence, general culture, and professional information because they do not also measure personality, moral character, interest in children, and other important factors that determine teaching ability, would be as illogical as to abandon the use of the clinical thermometer and stethoscope because they do not measure a thousand other important diagnostic factors...The validity of the examinations should be judged by the accuracy with which they measure not the total complex of teaching ability, but those parts which they are designed to measure..." (1940, 278-279).

Of course, if test builders/users do not wish others to impute incorrect uses from the scores, they have a responsibility to make clear just what inferences they wish to draw, and the evidence or logic supportive of those inferences. Kane (1984) suggested that there are two common interpretations of the scores on licensure examinations.

> "First, they can be interpreted as providing predictions of an examinee's future professional performance. Second, they can be interpreted as providing evidence of an examinee's present competence on specific abilities that are needed for practice" (1984, p. 2).

Kane, and almost all others who write in the professional literature regarding licensure examinations, argue that the second of these two interpretations is the more appropriate one. Of course, if one uses a teacher competency test as an employment examination, there usually is an implied inference of the first type.

Content Validity Evidence For Teacher Competency Tests

Teacher competency test scores are potentially useful for licensure/certification, and/or employment purposes. The 1985 Standards, with its separate chapters on employment testing and licensure/certification, make it quite clear that different types of validity evidence should be gathered for the two types of uses. Measurement leaders in the field of licensure generally agree with the position taken in the Standards that content validity is the primary concern for licensure

tests. Examples of just two such quotes follow, but there exist many others making the same basic point.

> "The content validity strategy is the one that appears to lend itself best to licensing and certification tests" (Shimberg, 1982, p. 62).

> "The appropriate type of validity to consider in evaluating licensure examinations is content validity" (Vertiz, 1985, p. 97).

However, the content validity evidence should differ for licensure and employment purposes. As pointed out earlier, and as stated in the Standards, for licensure tests the

> "focus of test standards is on levels of knowledge and skills necessary to assure the public that a person is competent to practice, whereas an employer may use tests in order to maximize productivity" (AERA/APA/NCME, 1985, p. 63).

Further, as pointed out, employment tests may measure aptitude to learn a specific job whereas licensure is usually to determine current qualifications for a broad field rather than a specific job. This has implications for the content to be covered (AERA/APA/NCME, 1985, p. 64).

Another distinction is that while an employment test should cover the totality of the knowledges, skills, and abilities (KSAs) desirable on the job, the content domain of a licensure test should be limited to the "knowledge and skills necessary to protect the public." (AERA/APA/NCME, 1985, p. 64). Note that "abilities" was left out of this quote.

> "It is also emphasized that skills that may be important to success but which are not necessary for competent performance and therefore are not needed to protect the public are appropriately excluded from consideration in a licensing examination" (Linn, 1984, p. 9).

Kane made the same point suggesting that "Specification of test content in terms of critical abilities does not require an exhaustive listing of the abilities required for practice" (Kane, 1984, p. 7). There is at least some legal precedent to suggest that a licensure examination need not evaluate the full range of skills desirable to practice a profession (Eisdorfer

& Tractenberg, 1977 p. 119).

It may seem that because licensure examinations do not need to test the totality of the KSAs which would be desirable on the job that the content validity requirements of licensure tests are easier to fulfill than the content validity requirements of employment tests. However, for licensure tests Yalow & Collins argued that the domain tested should be "necessary, not merely desirable" (1985, pp. 5-6). They suggested that "a focus on the more common parlance of 'job-relevance' is too lenient a standard to impose" (p. 9). This stance has some merit. Note that the quote from the Standards given earlier suggests that the focus should be on necessary knowledge and skills to assure that the person is competent to practice. It is somewhat debatable whether the Uniform Guidelines' apply to licensure examinations (to be discussed further at a later point in the paper). Nevertheless, it should be pointed out that the Uniform Guidelines' technical standards for content validity studies do allow for a non-representative sample of KSAs if it "is a necessary prerequisite to successful job performance" (EEOC, 1978, p. 38302). I agree with others who interpret 'successful' as used in the quote here to mean adequate, rather than exemplary performance (see Yallow & Collins, 1985).

The problem with the "necessary" requirement is that very, very few specific competencies are probably absolutely necessary to adequately practice any profession, yet if one person has twice as many very important competencies as another person it is certainly prudent to believe that the public is safer with the first person than with the second. Further, if one only tested for necessary skills, it would follow that the cut score should be set at 100% [or whatever other percentage one may arrive at through those "counting backwards from 100%" procedures that Glass (1978) talked about]. Nevertheless, one can anticipate that expert witnesses for the plaintiffs in teacher competency test law suits will hold very high

standards for the "necessity" of the knowledges and skills that form the domain for the test construction process. This suggests that one should use the word "necessary" when asking educators about the relevance of the knowledges and skills.

The necessary requirement is probably least debatable in the subject matter tests of teacher competency. A reasonable argument is that one cannot teach what one does not know. Therefore it is necessary to know the subject you are to teach. Galambos (1984), suggested that this assumption has been accepted as self-evident by legislators. Critics of licensure examinations also will be likely to accept this assumption as self-evident at the general, abstract level. But even in subject matter tests there will be questions that ask about specific knowledge that is not absolutely essential. For example, every reasonable person would probably agree that an American History teacher should have some knowledge of American History in order to teach it. However, a specific question that taps a specific portion of the overall domain may test for knowledge that not all would consider absolutely essential. This could be true even though the question matches a fairly specific relevant objective. What needs to be made clear in these situations is that the test samples the domain, and that a single inference is made about the knowledge of the domain rather than a set of inferences about the knowledge of specific questions (or specific objectives). If a test is composed of questions, all of which measure relevant objectives within a relevant domain, then it is reasonable to infer that a person with a high test score over that domain has the minimum necessary knowledge to teach the domain, and to infer that a person with a low test score over the domain does not have the necessary knowledge. These could be reasonable inferences even though one might not believe that the knowledge tapped by any single question was <u>absolutely</u> necessary.

As mentioned in the previous paragraph if knowledge of every single specific piece of information tapped by the questions were absolutely necessary, then the cut score should be set at 100%.

The necessity to have knowledge regarding classroom management, assessment techniques, or developmental psychology is probably less "self-evident" than the necessity to know the subject matter. The same is true for knowledge of basic skills. It is probably least self-evident that a test over general knowledge measures necessary knowledge. Is it necessary for a person to be well educated in a general sense in order to be an adequate teacher?

Tests over pedagogy, basic skills, or general knowledge are almost certain to contain questions testing specific knowledge that is not absolutely essential. For example, most of us would probably agree that teachers should know something about how to measure the knowledge of their students. A test over that sub-domain of pedagogy would be considered relevant. We could all probably agree at the abstract level that a teacher could know so little about that sub-domain that he/she should not be licensed to teach. That indeed, giving a license to teach to someone who knew very little about measurement techniques could well result in harm to individual pupils. To protect the public from that potential harm one might well decide to build a test covering measurement knowledge. Questions matching objectives within that sub-domain might help contribute to a correct inference about whether prospective teachers knew the minimum amount necessary about the sub-domain to be licensed even if each specific question, standing alone, could not be defended as measuring absolutely essential knowledge. Obviously the same point could be made for the basic skills. For example, we would probably all agree that teachers should have some skill in spelling. We could probably all agree at an abstract level that there exists a level of spelling proficiency so low that people with

only that level of proficiency should not be licensed. We might be able to make correct inferences about the inadequacy of necessary spelling skills from a spelling test even though we could not defend the absolute necessity of being able to spell any single word in the test.

Making an inference about the general adequacy of necessary knowledge from a test sampling a domain without making any assumptions about the necessity of each specific piece of knowledge tapped by each question should not be something about which the measurement community would disagree. However, if the critics play word games with the judges, test proponents need to be prepared to make the above point as clearly and forcefully as possible.

Content Validity Established Through Test Construction

Content validity is established only in test construction (Cronbach, 1980, p. 105). Thus, it is essential that those who wish to argue the validity of teacher competency tests through content validity evidence must follow appropriate test construction procedures. Yalow & Collins suggested that there are three choice points for test developers to help assure content valid teacher competency tests: "(1) the isolation of eligible skills for assessment, (2) the selection of the skills to be assessed, and (3) the operationalization of selected skills as test items" (1985, p. 4).

The steps used in the development of the tests for Georgia, for example, included (1) topic outline selection and review, (2) objective writing, review, and revision, (3) job analysis survey, (4) approval of selected objectives, (5) item writing, review, and revisions, (6) field testing, and (7) content validity and minimum cutoff determination (see Georgia Department of Education, 1985, p.7). Florida enumerates their steps as follows:

A. Planning

(1) Identification and validation of the essential teacher competencies
(2) General planning for examination development
(3) Development of subskills for each competency
(4) Development of test and item specifications

B. Writing and Validating Test Items

(1) Create the test items
(2) Pilot test the items
(3) Conduct a review of the items

C. Field testing the certification examination items

D. Setting cutting scores

E. Preparing for test assembly, administration, and scoring. (Florida Department of Education, 1981, pp. 1-13).

While the wording of the steps listed in the examples given above differs slightly, the major points of concern in establishing the content validity appear to be (1) developing an original list of competencies, (2) doing some type of job analysis, (3) specifying the domain for the test, (4) writing and validating the items, and (5) obtaining an overall judgment of the content validity of the test. These five steps will be discussed in some detail below plus the additional sixth step of communicating the domain to the test takers and the general public. The setting of cut scores will be discussed briefly in another section of this paper.

Developing an Original List of Competencies

Whoever is involved in developing the original list of competencies must understand the purpose(s) to which the exam will be put. As has been discussed, licensure and employment examinations have somewhat different purposes and therefore the competencies tested may not be identical (although certainly there would be considerable overlap).

The most general starting point for developing the list of competencies

is to appoint a relevant committee to do the task. This committee should be composed of experts within the field. For teacher competency exams these experts may be practicing K-12 teachers, supervisors, university professors, and/or state department personnel. The members of the committee should have the necessary expertise and the committee should have credibility with the appropriate constituents. It is probably useful to have a variety of perspectives represented on the committee. In fact, Yalow and Collins suggested that

> "it is critical to have input from individuals who have a variety of perspectives regarding the on-the-job demands that educators face" (1985, p. 4).

The starting point for the committee should be a thorough review of the relevant literature (Burns, 1985). What kind of knowledges and skills should be considered critical? Kane made the following point:

> "The American College Dictionary defines a profession as a 'vocation requiring knowledge of some department of learning or science.' Presumably many of the critical abilities will be drawn from the department of learning or science associated with the profession. (1984, p.6).

This is precisely what the various states (and other test constructors) are doing when they look at teacher-effectiveness (process-product) research. Thus, any new test of teacher competencies should include a thorough review of the teaching competencies tested in other states, the scope and content outlines from state departments of education, and the literature on teaching effectiveness. Note that this is not the same thing as trying to establish the "curricular validity" of an examination. The purpose of going to the literature is to find out what is critical, not to find out what is being taught in any particular curriculum.

One additional literature source that may be helpful in formulating task statements is the literature reporting how teachers spend their time in the classroom. In a report of an ETS job analysis Rosenfeld, Thornton & Skurnik (1986) reference two studies in Great Britain and seven studies in

the United States that include classroom observations. They state that "the categories used by these researchers to code teacher activities and the results of their investigations were useful leads . . ." (1986, p. II-2).

Of course, the literature review would be somewhat different for examinations in pedagogy than for examinations in subject matter fields. As mentioned earlier, for subject matter fields, an assumption considered self-evident is that one can not teach what one does not know. Therefore, it is critical that teachers know the content they are to be certified to teach. To determine this content, a search of the curricular materials in the appropriate grade levels for which certification will be given is appropriate. However, it is <u>not</u> being suggested that teachers only need subject matter knowledge at the level they are teaching (see Shulman, 1986).

Another approach occasionally taken in addition to the literature review is to interview teachers. This was done by the ETS team in their job analysis (Rosenfeld, Thornton, Skurnik, 1986) and is currently being done by the University of South Florida in their construction of licensure examinations for the state. This step is probably differentially useful depending upon the recent teaching experience of the committee developing the list of competencies.

Doing the Job Analysis

Professional standards, logic, and legal precedent all stress the importance of job relevance or job relatedness in both employment and licensure exams. However, as mentioned, employment and licensure exams do not measure exactly the same domain, so the determination of job relevance may not be carried out in exactly the same way. Two commonly accepted methods of determining job-relatedness are through document review and group discussion. These two methods should be employed by the committee developing the list of original competencies discussed in the previous

section. According to the Principles for the Validation and Use of Personnel Selection Procedures (APA, 1980) this process of using the pooled judgment of experts is a recognized approach to determining job relatedness. It is occasionally labeled a job analysis, but other writers prefer to refer to it more generally as a method of documenting job relatedness.

Another common approach to job analysis is observation. But,

> "some jobs, including many in the white collar occupations, do not lend themselves readily to analysis by observation. Employees in such jobs frequently can describe their work fairly readily" (U.S. Civil Service Commission, 1973, p. 6).

Most experts feel this quote is particularly appropriate to the job of teaching, especially for licensure exams where the critical job elements need to be included as opposed to the total domain of job elements. While a few educational measurement experts would wish the job analyses to include observations, they appear to be in the minority. Two problems exist with the observational approach. One, for an observer to know the frequency of a skill use, he/she would have to observe for a long time. This seems somewhat unfeasible. Further, the observer would have to be very observant. One technique to enhance the accuracy of observations is to compile a list of all behaviors to watch for in advance. Of course, if one could identify all the skills that were important prior to doing the observation, the only thing that would be observed would be frequency of using the skill.

The frequency with which elementary teachers use a knowledge such as dividing fractions by fractions would be almost impossible to determine through observations unless one had a very large group of observers observing throughout the total school year in all of the elementary grades. This is just not a reasonable approach. Most experts would rather make an inference regarding frequency from surveying a group of teachers rather than relying on observers' tally marks from a limited amount of observation.

Of course the non-necessity of doing observations for every job analysis is not to deny the potential usefulness, mentioned earlier, of using the published literature describing how teachers use their time as leads in developing the original list of competencies.

What appears to be the most common and feasible approach for doing the job analysis is through a survey of the people in the profession. This is true of job analyses for doctors, lawyers, psychologists, and teachers to name just a few of the professions. This job analysis survey is valuable in confirming or adding to the judgments of experts (Pecheone, Tomala, & Forgione, 1986). The survey instrument itself can vary in the specifics of the wording, and there are a number of variations in the sampling process.

Almost invariably the surveys ask respondents to rate the importance and/or frequency of use of a set of competencies gathered by a panel and based, in part, on a literature review. Job analyses for employment exams typically place heavy emphasis on frequency data (Williamson, 1979 as referenced in Kane, 1984). For licensure exams it is common also to gather data regarding the importance or criticality of the job element with respect to the purpose of protecting the public. Williamson (1979 as reported in Kane, 1984) stated that 32% of a physician's time at work is spent on activities other than patient care. One could reasonably argue that the competency of the physician on those activities is not critical to protecting the public. In the ETS job analysis a "time spent" rating scale was included in the original draft of the inventory but was eliminated based on advice from an external advisory committee and the project staff.

As Kane suggested "Given that the purpose of licensure is to protect the public, the 'harmful if missed' category would seem to be especially important for licensure examinations" (1984, p. 12). Kane went on to point out that a strong logical case can be made for the linkage between knowledge in a profession and effectiveness in the profession. He stated that

> "If there are several approaches to some issue of professional practice and the evidence does not consistently favor one approach, it would still be reasonable to require that candidates for licensure know enough about the various approaches to recognize their potential benefits and limitations. Given the purpose of licensure, it is especially important that practitioners be aware of any dangers inherent in various interventions" (1984, p. 14).

He suggested that job analyses are useful in providing information about the kinds of situations encountered in practice,

> "...while the department of learning is a more reliable source of information about how these situations should be handled. In weighting various critical abilities, both empirical job analyses and the department of learning have major roles to play" (p. 14).

The particular wording of the questions in the job analyses have varied somewhat. As mentioned earlier, for licensure exams some would argue that it is important to find out whether or not a skill is essential. A fairly common procedure in job analyses is to ask questions regarding both the importance of the skill and the frequency with which it is used. Again, with licensure examinations, one could make the case that the frequency is less important.

Not much research has been done on who should be sampled by the survey. Generally, the sampling has been done from the domain of practicing teachers in the state who are licensed in the field for which the test is designed. As Yalow and Collins suggested,

> "A large-scale survey in which practicing educators are asked to confirm/disconfirm the wisdom of any advisory panel's recommendations may add greater credibility to the decisions finally made" (1985, p. 5).

To my knowledge, no research has been done comparing the results obtained from a cross section of teachers with those obtained from teacher supervisors; nor has any research been done comparing how 'superior' teachers respond as compared to 'incompetent' teachers. Of course a problem in doing such research is the measure of the criterion variable. However, it is a reasonable assumption that 'superior' teachers would be more alert

to the demands of the job both in terms of teaching skills and content knowledge needed. Thus, if one used a cross section of teachers, a more conservative estimate of the job requirements would be obtained than if one used only superior teachers. [Levine, et al. (1981), suggested there are few differences in information obtained from superior and other performers in a variety of non-instructional job settings.]

For some speciality areas, there may be very few licensed teachers in a state so that the survey may well be given to only a few teachers. While a small number of respondents may be grist for those who want to find fault with the job analysis, one could hardly ask for more than to survey the whole population of relevant teachers.

If one wished to check the consistency of the survey data due to sampling error, one could divide the participants into two half-samples. This was done in a least one state (Echternacht, 1985). Basically, the evidence suggested that there was considerable consistency across the half-panels. "...if the panel members on one panel determined that a question was job relevant, the panel members on the other panel would almost always agree" (p. [illegible].15).

Obviously, most surveys done to determine job relevance are not done at the item level (exceptions would be for those surveys performed to "validate" existing tests). Surveys are done prior to final determination of the appropriate domain and the table of specifications for the test. All this, of course, is accomplished prior to building items for a test. Nevertheless, some critics have contended that the survey portion of the job analysis should be at the item level. The argument goes something like this: Just because an objective may be determined to be job relevant, it does not follow that an item purporting to test that objective is also job relevant. That is a theoretical possibility given certain flaws that may occur in the item writing procedures. Nevertheless, the determination of

the test's domain, which is what the job analysis helps do, simply is not done at the item level. One does need to have item review procedures to assess the item validity and these will be discussed later. These procedures are not reasonably considered a part of the job analysis.

One question that has not been addressed adequately is the degree to which a job analysis conducted in one state is generalizable to another state. If the job of a teacher in one state is similar to the job of a teacher in another state, perhaps even the same test can be used across states. Of course, this is done by those many states that use the NTE exam. But the states do typically have separate validation studies. Galambos made the following points about this issue.

> "The development of customized state tests usually results from the desire to make the tests as acceptable as possible to teachers within a state. Designing test items against objectives that are developed by teachers within a state is a means of 'selling' the test on the basis that it represents the curriculum established for that state. Yet, there are some myths associated with this philosophy. First, it is questionable whether the curriculum in certain basic subjects such as English and mathematics is really different from state to state. ...Secondly, even where a state contracts with a test development firm for a customized test, that firm usually has a bank of test items that is used for its clients, so that even a customized test represents a medley of nationally used items" (1984, p. 7).

While I have not looked at enough tests to support or refute the second point, it does seem to me that in many cases the variance in teachers' jobs across states is probably no larger than the variance in teachers' jobs within a state.

ETS (Rosenfeld, Thornton, & Skurnik, 1986), completed a job analysis of teachers in three different states. Basically a job analysis questionnaire was developed which asked teachers to rate the importance of 83 teaching tasks and 59 knowledge areas. Based on a principal components analysis of the results the researchers classified the tasks into six factors representing important core functions:

1. managing and influencing student behavior;
2. clerical, administrative, and other professional functions;
3. assessing, grading, and recording student learning progress and evaluating instructional effectiveness;
4. planning the lessons, selecting the materials, and previewing the instructional programs;
5. implementing the planned instructional programs using a variety of approaches;
6. identifying students with individual or similar instructional needs and teaching them accordingly

Although this job analysis was not done on a nationally representative sample, it lends some support to the notion that each state should not feel the need to do a unique job analysis. (This statement is not an endorsement of the NTE. A state may well decide to build their own test and not use the NTE. The point being made here has solely to do with the generalizability of the job analysis.) Of course, it is possible that the feeling of ownership and uniqueness that may come from a state job analysis more than compensates for the cost accrued even if the job analysis is no better than one it could adopt.

Examples of State Job Analyses of Teachers.

In at least four states (Georgia, Alabama, South Carolina, and Oklahoma) a sample of educators were asked to rate a set of objectives in terms of the amount of time spent teaching or using the objective, and the extent to which the objective was essential in their field. The objectives were developed by panels of content experts in much the manner previously discussed. In Florida, a five percent random sample of all certified educational personnel employed in Florida were sent a survey of 48 competencies. They were asked to rate them in terms of their importance to the field. In Connecticut a state-wide survey of a random sample of 2743 Connecticut teachers and administrators was conducted to determine their views regarding previously identified competencies. Respondents were asked

to answer two questions for each of 85 competencies: (1) the importance of each competency as a measure of teacher effectiveness, and (2) whether or not the behavior described was directly observable by the evaluator. The second question was important for that state because the actual assessment was to be through observational techniques.

Other states have conducted similar job analyses. A reasonable summary statement is that these various job analyses have been conducted with a high degree of professionalism and are, in fact, more thorough than the job analyses conducted for most state licensure examinations in other fields.

Determining the Domain Specifications

As Elliot and Nelson pointed out: "There is little to guide the developer of teacher licensing tests in making the huge leap from job analysis to domain specifications" (1984, p. 9). This should not surprise us. Experts in the field of achievement testing have for years been unable to reach complete accord on how explicitly the content domain needs to be defined or what algothrims one might set up to weight the sub-categories of the domain or to sample within the sub-categories.

Kane (1982) made an important distinction between import and validity. As he suggested, "import is a qualitative concept, which emphasizes the scope and significance of the inferences that can be drawn from a measurement" (p. 150).

> If import is ignored, it is easy to generate measurements with a high degree of validity by defining the universe of generalization narrowly enough so that the inferences to the universe scores involve generalizations over few facets. Such inferences are likely to be very dependable. ...However, the issue of import cannot be ignored, and trade-offs between validity and import must be made... The researcher who interprets observations narrowly draws more accurate inferences but also says less about the world than the researcher who interprets observations broadly. The choice between narrow but dependable interpretations and broader, but less dependable, interpretations is a choice of strategy. The

> continuum of available options has strict operationalism at one end and construct validity at the other end" (pp. 151-152).

The problem is that the critics of any given teacher competency test argue that one should define the domain very precisely, but then criticize the test because they claim that the inferences are not of large enough import. That seems unfair and contradictory to me. At any rate, determining the trade off between import and validity is obviously a judgmental task, and as Cronbach suggested, "the defense must be prepared to show that the domain is relevant and that weight is properly distributed over it" (1980, p. 105).

Three general points need to be made about the domain of licensure tests. (1) the domain should be fairly broad because a certificate is not for a specific job but for a general kind of job, (2) the domain does not have to cover the total set of tasks determined by the job analysis, and related to that, (3) a licensure test does not need to have, and probably should not have, sub-category weights that are proportionate to the amount of time one spends on that subcategory on the job.

What one should do is cover the domain of critical knowledges and skills. The weighting of the area should be based on the degree of criticality which in turn is based on both frequency and impact. One should be particularly alert to the "harmful if missed" category for licensure examinations (Kane, 1984).

As mentioned earlier, a fairly common procedure in conducting the job analysis survey is to ask questions both about the amount of time teaching or using an objective, as well as the essentiality of the objective. Typically these data are combined in some fashion to determine a single value of "importance" for each objective. The algorithm used to combine the two pieces of information does not matter a great deal because there is good evidence that the correlation between the responses to the two questions is

quite high. While it is probably most common to weight the responses to the questions equally, I would prefer to weight the essentiality question greater (for reasons discussed earlier). At a practical level it just does not matter much. For example, one unpublished study investigated the intercorrelations among three formulas for combining information. Job analysis information was collected for three different scales: A) Have you taught directly or utilized this objective during this school year or the past school year: If answered affirmatively, two more questions were asked: B) How much time was spent teaching or using this objective? (5 point scale); and C) How essential is it that this objective be included in the curriculum of my entire teaching field or the content of my instructional support field? (5 point scale). Values were computed separately for each participant using the three formulas: $\sqrt{B^2 + C^2}$, ABC^2 and ABC. These values were averaged across participants. The correlation of the objectives between the first two formulas was .93, between the first and third it was .91 and between the second and third it was .996. (Lahey, 1985).

Once one has some sort of ordering of the objectives, it is both appropriate, and common practice, to use that information along with any sub-domain information to select a proportional number of important objectives within each sub-domain. It generally would be considered acceptable practice to give the panel of experts some flexibility in choosing objectives rather than forcing them to use some unflexible algorithum based on the ratings (see Millman, 1986).

There is some disagreement about whether there should be some minimum cut score based on the job analysis ratings before an objective can be included in the final pool of objectives. Linn & Miller (1986), for example, advocate using absolute rather than relative ratings. However, I believe if the committee developing the original pool of objectives did

their job properly all of the objectives are likely relevant and their relative ratings are more informative than their absolute ratings.

Writing and Validating the Items

The most commonly used item format for licensure examinations is the multiple-choice item (Shimberg, 1982). This seems quite appropriate because the purpose of most licensure tests is to see whether or not the applicants have the necessary knowledge. Almost all authors of measurement texts have advocated the use of multiple-choice items (see for example, Bloom, Madaus, & Hastings, 1981; Ebel, 1979; Gronlund, 1985; Hills, 1981; Hopkins & Stanley, 1981; Mehrens & Lehmann, 1984; Nitko, 1983; Sax, 1980). There is a wide body of literature demonstrating that multiple-choice items can measure knowledge. This is stressed here because some critics have suggested this format to be inappropriate. Pottinger (1979) argued against such tests because they do not do an adequate job of protecting the public. That is, they let too many incompetent people get certified. This may be true. Research generally has shown that short answer questions are more difficult than multiple-choice questions. This is particularly true of questions requiring solutions to problems. Apparently generating a solution is more difficult than choosing one. However, the correlation across people between the two types of tests is typically quite high. Further, the cut score procedure is based on the multiple choice items so supposedly the individuals determining the cut score have taken item format into account.

Other critics have argued that multiple-choice tests keep competent people out. Such critics seem to base their criticism on the notion that some people know a lot of material but are unable to demonstrate it on multiple-choice tests. The available evidence certainly suggests that you can not be admitted to, or graduate from, a reputable college without having the limited skill necessary to respond to such items. Logic plus previously

available evidence of the validity of tests using multiple-choice items suggests that one can adopt this format without having to gain independent evidence of the validity of such a format for this particular type of situation.

The writing of multiple-choice items is basically no different for the purposes of teacher certification tests than for any other test given to educated adults. Item writing guidelines abound (see above referenced texts) and most test constructors attempt to follow most of these guidelines even though research evidence suggests violating many of the guidelines apparently either has no negative impact on test quality or results in poorer students gaining the most from the poorly written items (Board & Whitney, 1972; McMorris et al., 1972). This last point is mentioned because critics occasionally seem to imply that poorly written items handicap the low scoring students more than the high scoring students whereas the reverse appears to be true.

Some professionals prefer item writers to work from what are commonly called item specifications (Popham, 1984). Several states that have built their own tests have used such a procedure (e.g. Florida). Other states and companies contracting with states have had their item writers work directly from objectives. There is no particular reason to prefer either approach although Millman (1986, p. 3-8) suggests, and I would concur, that more testing experts are in the latter camp. If the job analysis survey was based on some statements of the competencies desired (perhaps as statements of objectives), then translating these into item specifications prior to writing items in no way guarantees that the items will be more valid measures of the original competencies than if the items were written directly from the statements of competencies. It is true, of course, that well written item specifications tend to ensure that the items match the item specifications, but there may well have been some slippage between the

statement of competency and the item specification. This slippage could well be greater than that between the statement of competency and the item written directly from it. Almost all popular measurement texts (such as those referenced a few paragraphs back) do not advocate including item specifications as a stage in test construction. The new Standards (AERA/APA/NCME, 1985) do not mention item specifications in the index, nor as far as I could determine, anywhere throughout the book. (See Popham, 1984; and Roid, 1984 for positions advocating item specifications.)

Whether or not items are written from item specifications, it is necessary to have the items reviewed. Specific procedures for the item reviews have varied somewhat across states, but the general intent in all cases is to determine the adequacy of the items as measures of the objectives (or statements of competencies). Hambleton (1984) has an excellent overview of some of the methods of judging item validity. He suggested that when item generation rules (i.e. writing items from detailed item specifications following specific algorithms) are used this is an "a priori approach to item-objective congruence; the approach itself assures essentially that the items are valid indicators of the domain" (p. 207). This may be somewhat optimistic. It depends both on what one considers the "domain" to be and the format of the job analysis survey questions. As suggested earlier, such approaches may ensure that the items match the item specifications, but they do not ensure that the items cover the domain of competencies unless someone has made the judgment (or defined by fiat) that the item specifications are the domain. Most test construction experts would probably recommend that a post hoc judgment be made about whether items even match the item specification. It is not true that all item writers will be able to write items that match either very detailed, or quite broad item specifications.

Hambleton was quite correct when he suggested that

> "When domain specifications...are utilized..., the domain definitions are never precise enough to assume, a priori, that the items are valid. ...Thus the degree of item-objective congruence in a context independent of the process by which the items were generated must be determined" (1984, pp. 207-208).

Hambleton suggested two general methods for judging items: using empirical techniques and collecting judgments from content specialists. He and most other measurement specialists prefer the second approach. He described several possible judgmental procedures. One of these is a procedure developed by Rovinelli and Hambleton (1977) that results in an index of item-objective congruence. This procedure requires each judge to determine whether or not <u>each</u> test item reflects the content of every domain specification (e.g. objective). Thus, even with only 10 objectives and 5 items per objective one would have to make 500 judgments. As Hambleton (1984) suggested, this procedure is very time consuming to implement. Also, it will be a waste of time if the objectives are quite heterogeneous.

A second approach mentioned by Hambleton is to have content specialists rate the item-objective match. A third approach would be to have the judges match the test items with the objectives. In all the procedures mentioned, one could check the expertise or care of the judges by including some "marker" or "lemon" items which did not match the objectives to see if the judges identified these bad items. Hambleton reported that in one study it was found necessary to eliminate one reviewer (out of 20) because that reviewer detected only 2 of 19 bad items. As Hambleton pointed out:

> "An interesting question that needs to be answered concerns whether or not content specialists should be informed about the existence of "marker items" in the pool of items they are reviewing. How would the information impact on their ratings? On the surface it appears that they might be more attentive, but on the other hand, many reviewers may be reluctant to participate if they themselves are being judged" (1984, p. 212).

While I like the notion of marker items, to my knowledge most reviewers have not used them. I would not consider their absence as an indication that the item review was inadequate. If only one out of 20 reviewers turns out to be incompetent or careless, that suggests there are plenty of reviewers who do spot bad items.

An approach developed by Nassif (1978) and commonly used by NES is frequently called a dichotomous judgment model. In this procedure, each member of a panel of content experts indicates for each item whether or not the item is accurate, congruent with the objective, significant, and lacking in bias. For an item to be considered valid it must pass all four criteria. To "pass" the judges'results are compared to the binomial distribution to determine the probability, due to chance alone, of obtaining "x" valid responses for an item from a total of "N" raters. In essense, this means that for an item to pass almost all the raters would have to indicate that the item is valid on all four criteria.

Another example of the item review process was the one used by Florida. First, a review panel keyed the items; traced them back (blind) to the subskill and content categories; and then rated the items for appropriateness. Secondly, three separate reviews of the items were conducted: for content, bias, and technical quality. The content reviews were conducted by the content specialists; the bias reviews were conducted by minority persons, women, and experts trained in linguistics; and the technical review panel included both measurement and language arts experts.

Some measurement experts would prefer the approach of using separate groups of experts to make the separate judgments. Others believe that what evidence exists suggests a panel of content experts is sufficient to do all the tasks. In fact, there is some anecdotal evidence to suggest that minorities select fewer items as being biased than do non minorities (Ruch, 1984). Berk (1984, p. 100) suggested that the panels be composed of

individuals representative of the appropriate subpopulations (e.g. males, females, blacks, whites, Hispanics). Tittle (1982) suggested using "at least two representatives from each group as expert judges" (p. 55), although she suggested that further research was needed with respect to the use of expert judges.

There is also some disagreement as to whether or not the judges should be meeting as a group and forming a consensus, meeting as a group and having the opportunity for discussion but voting independently, or making totally independent judgments. Each method has some potential disadvantages. The first two may suffer from social psychological factors. An assertive, strong willed person may end up "controlling" the vote. The third approach may suffer due to the lack of opportunity to discuss with others, which may stimulate one's thinking.

Overall judgment of content validity

As mentioned at the outset, content validity is established only in test construction. Thus, the judgment of the adequacy of the content validity should be based on a judgment of the adequacy of the construction process. In licensure exams, the purpose is to make inferences about whether or not the test takers have sufficient knowledge and skills to protect the public from harm. If the original list of competencies has been developed by experts, if the job analysis (or survey) is accomplished appropriately, if the test specifications have been developed from the results of the first two steps, and if the items have been written and validated in a satisfactory manner, then the test will have appropriate content validity. It will be assessing those competencies that experts in the field thought necessary for beginning professionals to have in order to protect the public. Even if all the steps were not executed perfectly, the use of multiple review groups on multiple occasions should provide "enough

safeguards against the inclusion of some out right invalid topic or objective" (Millman, 1986, p. 3-7).

States that adopt the various NTE tests frequently make an overall judgment as to the content validity of that test in a different fashion than that described here. Typically a thorough review of the test construction process is not made. Rather an analysis of the items within the various NTE tests is made. The approach used is to survey a group of individuals (frequently called the job relevance panel). These individuals are asked to make judgments about the degree to which the knowledge or skills tested are relevant to competent performance as a beginning practitioner. The states set some cut-off on the degree of relevance ratings to arrive at a decision regarding whether the total test has sufficient content relevance to administer in the state. Cross reports that 35 studies involving one or more of the NTE tests have been completed in 35 states. He concludes that "the results of past studies tend to support the use of the Core Battery and most of the area tests for initial certification" (1985, p. 9). That is a value judgment with which not all would agree. Many of the studies only required a majority of those surveyed to vote that an item was valid in order for it to be so counted. Further, many items could be considered non valid and the state could still use the test. Gifford concludes that the NTE Core Battery is lacking in content validity. His conclusion is based, in part on the fact that "while as many as 38 percent of the test items have been identified as invalid for a given state, the NTE has not been modified accordingly, but continues to be administered as originally designed" (Guifford, 1986, p. 260).

Once the test is constructed, or reviewed by some procedure such as that described for the NTE, it seems inappropriate for measurement experts to second guess the decisions of the content experts involved regarding whether or not the test or the specific items are measuring appropriate

content. Of course measurement experts will continue to argue about the standards regarding what percent of the respondents need to regard an item or objective as valid because those standards, like all standards, are arbitrary.

Communicating the Domain to the Public

"Licensure is a public function, subject to public scrutiny" (Kane, 1984, p. 24). Both the individuals applying for licensure and the general public have a right to know the general content of a licensure examination. No one debates this. However, there is some debate about just what the public is to be told. Generally, the survey of objectives (job analysis) results in a greater number of objectives being rated as essential than it is possible to test in any given test. Thus, the test itself must sample the objectives from the total domain of objectives.

In my opinion, the situation in licensure tests is the same as for any other test where there is a sampling of objectives. One wishes to make an inference from the objectives sampled to the total set of objectives judged relevant. In order to do so, one must communicate the total set of objectives rather than the subset which are sampled for the test. Of course, if the test objectives are broad enough, or the test is long enough, so that the total set of objectives are tested, then there is nothing wrong with communicating to the public the specific objectives tested because the inference does not go beyond those particular objectives.

Not all would agree with my analysis. An expert witness at one trial testified that he found it misleading to communicate a larger set of objectives to candidates than are acutally being tested. Perhaps the measurement issue revolves around the meaning of a "criterion-referenced" test (CRT). Perhaps some feel that one can only infer to the objectives

specifically sampled and that an inference to the domain from which the objectives were sampled is not appropriate. In any event, the purpose of a licensure exam is to protect the public and the inference made is that a candidate does, or does not, have sufficient competence on all the knowledge relevant for that protection. If that domain is reasonably large, as it is almost sure to be in most professions, it will be necessary for the test to sample the domain.

Criterion-Related Validity Evidence

As mentioned earlier, some critics attack teacher competency tests because the passing of such a test does not guarantee that one will be a good teacher. A true, but totally irrelevant point. As Vold suggests, the promise of teacher exams "is not so much that they can identify competent teachers, but they do seem capable of weeding out incompetent ones" (1985, p. 5). Johnson & Prom-Jackson point out that the cognitive abilities of teachers "constitute a necessary but not nearly sufficient condition" [for teacher success] (1986, p. 279). Certainly, no one who knows anything about validity and testing would suggest a test score can offer any guarantee. But should not such tests have some predictive validity? A few writers would argue yes. Hecht, for example, although admitting that predictive validity studies in licensure tests are rare indeed, suggested that "predictive criterion-related validation studies would be the type most closely fitting the expressed purpose of licensure exams" (1979, p. 21). However her opinion is certainly not the common view held by most psychometrists. Shimberg stated the more commonly held position quite nicely:

> "What Hecht overlooks, however, is a difference between the purpose of a test intended for use in an employment situation and one intended for use in licensing. In the employment setting, employers frequently want to rank applicants so that they can select those with the greatest likelihood of success. They may

> define success in terms of sales, production, supervisor ratings, or some other work-related criterion. Whatever the criterion, they want assurance that those who score high on the test will also perform well in terms of the selected criterion.
>
> The licensing board has an entirely different goal in mind. Since the primary purpose of licensing is to protect the public from incompetents, boards need tests that can help them to differentiate in a reliable way those applicants who are able to demonstrate a minimum level of competence and those who cannot, those who at least demonstrate that they possess basic knowledge, skills, and/or abilities that can be used to safely serve the public, as contrasted with those who might pose a threat to the public if they are allowed to practice. Boards have no responsibility and, indeed, no authority to rank applicants in order of merit or to predict which applicants will perform best on the job...
>
> Those who believe that it is the purpose of licensing boards to predict job success might think so, but to follow their lead would drastically change the nature and purpose of licensing. It is doubtful that many legislators would agree that predicting job success should be a function of licensing boards" (Shimberg, 1982, p. 60).

Tenopyr took a position somewhat similar to the one taken by Hecht.

> "A particular problem extant in employment psychology today is that of the licensing or certification test. Those who construct such tests appear to treat them as pure achievement tests and argue that a licensing test only assures prospective employers or the public that a person has the necessary knowledge and skills to practice in a given profession or trade. However, the assurance of minimum skills is merely an aspect of prediction. It is predicted that those not possessing the minimum skills will do a poorer job of professional practice than those who do possess those skills" (1977, p. 49).

Tenopyr's more precise statement would probably receive more support at a theoretical level than would Hecht's. Kane (1984), in arguing against any reason to expect a correlation coefficient based on data from passing candidates admitted that a measure of agreement between the pass/fail dichotomy on the licensure examination and a competent/incompetent dichotomy in subsequent practice would have some relevance. However, an index

> "that would address this issue cannot be estimated without having criterion scores for those who fail the examination as well as for those who pass. Attempts to collect such data might be considered unethical (and probably illegal) in many professions" (1984, p. 5).

Even if such data were gathered, a lack of a relationship could well be due to our inability to detect those practitioners who are incompetent and causing harm to the public. Linn (1984), Kane (1982, 1984), Shimberg (1982, 1984), and others all have argued that it is both unfeasible and inappropriate to expect criterion related validity of a licensure examination. Rosen (1986) also states that the predictive model is inappropriate for licensure examinations. Part of his explanation for his stance is because he looks at licensure as a public reassurance.

> "When a licensure is denied, the statement being made is: We do not have sufficient reason to abondon the assumption of incompetence. Abandoning the assumption of incompetence is not at all the same as predicting competence" (1986, p. 11).

The Standards state that

> "Investigations of criterion-related validity are more problematic in the context of licensure or certification than in many employment settings. Not all those certified or licensed are necessarily hired; those hired are likely to be in a variety of job assignments with many different employers, and some may be self-employed. These factors often make traditional studies that gather criterion-related evidence of validity infeasible" (AERA/APA/NCME, 1985, p. 63).

One of the major practical problems in criterion-related validity is that there is no clear definition of what it means to be an effective teacher (Webb, 1983). This certainly complicates the criterion problem. Stark and Lowther (1984), for example, listed six different conceptualizations of teaching and give as examples 10 different criteria for teacher evaluation. Ebel (1961), in a widely referenced article, discussed some of the general problems in criterion-related validity studies. To quote in part:

> "Even in those rare instances where criterion measures have been painstakingly devised, the validity of the test is not determined unless the validity of the criterion has been established. This requires a criterion for the other criterion, and so on ad infinitum. We can pursue such an infinite regress until we are weary without finding a self-sufficient foundation for a claim that the test is valid" (1961, 642).

Kane made the following points about the criterion problem for licensure exams.

> "The usefulness of predictive validity for licensure examinations is limited greatly by the fact that criteria of proven validity are not available for licensure examinations. The development and validation of a criterion measure of professional performance presents fundamental conceptual problems as well as great practical difficulties. In part because practice requires a high level of professional judgment for effective performance, the distinction between good practice and poor practice is not clear-cut in most cases...and the development of general measures of the quality of practice that are comprehensive, reliable, and valid is probably not possible for most professions" (1984, pp. 3-4).

As Cronbach stated: "When a test fails to predict a rating, it is hard to say whether this is the fault of the test or of the rating" (1970, p. 127). It is probably safe to suggest that if teachers had good supervisors' ratings on teaching effectiveness but could not pass a test on the content they were supposed to be teaching, most reasonable people would doubt the ratings. That may not be quite as likely if the test were covering pedagogy. (It is interesting in this general regard to reflect on what the differences might be in public perception if an M.D. or an attorney practiced "successfully" but had not passed the prerequisite licensure examination and/or not received the prerequisite professional training. In those cases where someone has been caught practicing medicine without a license the general reaction of the public is not that such instances indicate that the person practicing was competent and that licensure of M.D.s is not needed. Rather, they typically interpret the situation as an instance of an incompetent not getting caught sooner. If such an interpretation is not as likely in education, it may be because too many educators argue incorrectly that the licensure examinations are unnecessary to ensure competence in teachers. However, to argue in the abstract against the validity of professional education tests is to argue against the validity of the profession.)

It is important to point out once again, that validity has to do with the inferences one wishes to draw from a score. Few, if any, of the advocates for licensure tests in general, or for teacher licensure tests in particular, wish to infer that the scores will predict degree of success on the job. Consider Ebel's comment:

> "Never, while I was at the Educational Testing Service, did I hear any of the administrators of that organization or the directors of the National Teacher Examination program claim that the test would predict success in the classroom. What we did claim was that it would indicate how much the applicant knew about the job of teaching. We claimed that it was a necessary, but certainly not a sufficient condition for effective classroom performance. We defended this claim on logical grounds. We believed it could not be defended empirically, and did not need to be. That is, none of us believed that a correlation between ratings of classroom effectiveness and NTE scores could shed more than a feeble and uncertain light on how well the test was doing the job it was intended to do. None of us doubted that knowing how to do a job usually facilitates doing it" (1977, p. 60).

In another article, Ebel made the following point:

> "Often the test itself is as good a criterion of competence to teach as we are likely to get. In such a situation, it makes little sense to demand that the validity of the test be demonstrated unless, of course, the intent is not to validate but to discourage its use" (Ebel, 1975, pp. 26-27, emphasis added).

Licensure tests are not designed to predict degrees of success among those licensed. It is generally conceded that criterion-related validity studies for licensure tests are unfeasible. Many individuals would rather trust the test scores than the criterion measures if a criterion validity study were done and the test failed to predict.

It does not follow from all of the above statements that it is inappropriate to attempt to find out what, if any, correlates of teacher licensure tests exist. (A cake can be good without icing, but most people agree that icing on the cake is an added plus.)

While correlational data are somewhat sparse, they are consonant with the logical inference that knowledge about teaching and the subject matter being taught (comptence) should be related to both performance and

effectiveness in teaching. For example, a review of the validity of the old NTE found that correlations between undergraduate GPA and scores on the Weighted Common Examinations Total (WCET) ranged from .23 to .74 with a median of .55 (Quirk, et al., 1973). More recent research reviewed by McPhee and Kerr (1985), produced similar results. Hardly anyone would claim that amount of knowledge acquired as an undergraduate was totally irrelevant to success as a teacher. Although the WCET scores did not correlate very highly with supervisor ratings, that was generally interpreted as due to a defect in the rating scales. However, as the authors pointed out:

> "Perhaps more important than revising principal and pupil rating scales is to conduct systematic studies of the relationship between the NTE scores of teachers and average residual achievement gain scores of pupils in their classes" (Quirk, et al., 1973, p. 109).

Webster (1984) has performed just such a study using both a general aptitude test: the Wesman Personnel Classification Test (WPCT), and the NTE Common Exam. As Webster pointed out, the WPCT was not designed to identify persons who would make excellent teachers.

> "It was assumed however, that persons who scored very low on the WPCT would be expected to encounter more-than-average difficulty in a profession that depends so much on one's ability to communicate. In short, it seemed logical that successful teachers should be minimally competent in acquiring, remembering, and transmitting knowledge" (Webster, 1984, p. 4).

Using a class average residualized composite score (CARCS), Webster found a correlation of .47 between CARCS and the NTE Common, .47 between CARCS and WPCT-Verbal, .37 between CARCS and WPCT-Math, and .48 between CARCS and WPCT-Total. All correlations were significant at $p \leq .01$.

Piper & O'Sullivan (1981) had university supervisors rate elementary education majors on a Performance Evaluation Instrument designed to measure classroom competencies. They found that scores on that instrument were significantly correlated (.43) to NTE Common Examination scores. Coleman et al. (1966) found that the verbal ability of teachers was the single most

important characteristic of teachers in accounting for student outcomes. Other research shows that teacher competency tests are related to admission tests. For example, Ayres (1983) reported a correlation of .88 between SAT scores and the NTE common examination score. McPhee and Kerr (1985) reviewed other studies with similar findings.

Validity Generalization

Validity generalization studies also support the notion of the validity of teacher competency tests. A recent United States Employment Service report provided evidence that suggests that "cognitive ability is a valid predictor of job performance for all jobs" (USES, 1983, p. 14). Hunter (1983), in a causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings found a correlation of .53 between cognitive ability and performance; a correlation of .61 between cognitive ability and job knowledge; and a correlation of .67 between job knowledge and job performance. He summarized his analysis in part as follows:

> "The data in the 14 empirical studies reviewed here confirm the routine assumptions of psychologists in the employment area over the last 50 years. There is a high correlation between cognitive ability and job performance that is in part the result of the direct impact of ability differences on performance but that is even more the result of indirect causal impact due to the high correlation between ability and job knowledge and the high relevance of job knowledge to job performance" (Hunter, 1983, p.265).

He explained some of this as follows:

> "Ability should be related to performance in two ways. First, to the extent that the job calls for reasoning, planning, or memory, speed and smoothness of performance will depend on cognitive ability. Secondly, ability determines the extent to which the person masters the knowledge required for efficient and excellent performance. Ability is especially important if the job requires adjustment to novel circumstances or changes in behavior due to changing job requirements" (1983, p. 257).

This view seems to fit the teaching situation.

Thus, some direct empirical evidence showing that teacher competency and ability tests are related to student performance, some evidence that teacher competency tests are correlated with ability, some evidence that both ability tests and job knowledge are generally related to job performance, and the notions of validity generalization should make us feel somewhat comfortable regarding the relationship between teacher competency tests and success on the job. Of course, as the Standards state:

> "The extent to which predictive or concurrent evidence of validity generalization can be used as criterion-related evidence in new situations is in large measure a function of the accumulated research" (AERA/APA/NCME, 1985, p. 12).

Many of us would want more evidence gathered on these relationships for more teacher licensure tests in more settings before we would feel totally comfortable in generalizing from the predictive validity in one setting to the predictive validity in another setting. But some predictive validity evidence is available. And remember, the Standards do not require licensure tests to have evidence of predictive validity anyway!

Construct Validity Evidence

Construct-related validity evidence

> "focuses primarily on the test score as a measure of the psychological characteristic of interest...Such characteristics are referred to as constructs because they are theoretical constructions about the nature of human behavior" (AERA/APA/NCME, 1985, p. 9).

While some measurement experts believe that all validity is construct validity, other measurement experts worry some about this because theoretical often implies hypothetical (Ebel, 1974). In speaking specifically about educational and employment testing Ebel suggested the following:

> "Most of what we teach in educational institutions, and most of what we test for in employee selection are knowledges, skills, and abilities. These can all be defined operationally. They are not hypothetical constructs. Ability to type, to spell, to weld, to solve problems with algebra, calculus or computers; these are not the kind of latent traits Cronbach and Meehl had in mind. We would speak more sensibly, I think, if we did not call them constructs.
> Why do we continue to talk about construct validation as if it were something we all understood and have found useful? Has any educational or employment test ever been shown to possess construct validity?...It should be of no real concern, at the present stage of its development, to those of us engaged in achievement or job testing" (Ebel, 1977, p. 61).

Many measurement experts are concerned about any implied necessity for construct validation because it is viewed "as an ill defined and unending process" (Linn, 1984, p. 7). The Standards do not require construct validity evidence for licensure tests. However, they do state that

> "Standard 11.2: Any construct interpretations of tests used for licensure and certification should be made explicit, and the evidence and logical analyses supporting these interpretations should be reported. (Primary)" (AERA/APA/NCME, 1985, p. 64, emphasis added).

The problem is that a critic may infer a construct the builder/user did not want implied and then criticize the builder/user for not making it explicit! Those measurement experts who think all validity is construct validity would probably suggest that the very term "teacher competency" implies a construct, although the definition by Medley given earlier in this paper would not necessarily lead to such a conclusion. According to Medley's definition,

> "competencies refer to specific things that teachers know, do, or believe but not to the effects of these attributes on others. ...Teacher competence is the repertoire of competencies a teacher possesses" (Medley, 1982).

This does not seem like a theoretical construct.

Builders/users of teacher competency tests are in somewhat of a dilemma with respect to referring to evidence they gather as construct validity evidence. If they do so, then the critics say "Ah-ha, you do admit

competency is a construct." Then, because construct validation is somewhat an ill defined and unending process the critics can attack whatever evidence is gathered as being inadequate. The very choice of wording may eventually cause builders/users legal grief. For example, Kane, when talking about licensure tests commonly used the phrase "critical abilities." Ebel, in the quote earlier used the same word. Although, to Ebel, the word "abilities" does not imply a construct, its usage allows some individuals to infer a construct. The writers of the Standards, apparently very alert to this issue, wisely only used the terms knowledges and skills in referring to licensure tests, leaving abilities out of the commonly used KSA terminology.

If a builder/user wished to gather evidence called "construct validity" evidence (in spite of the illogical but real legal dangers of so doing) how should it be done? If one is going to suggest that a test measures a construct, it would appear necessary to define the construct. Kane suggested that as one example of construct validity the construct at issue, "professional competence, is defined in terms of the network of theoretical and empirical relationships incorporated in the department of learning" (1984, p. 8). However, in another article he pointed out that

> "the validation of measurements that are interpreted as dispositions does not depend on theory. Measurements of a disposition are valid to the extent that they provide accurate estimates of universe scores. The existence of laws or theories involving a dispositional attribute has no direct bearing on the validity of measurement of the attribute....This point of view is generally consistent with the interpretation of measurement is science...Campbell...concluded that 'measurement is essential to the discovery of laws' but he did not use the laws to evaluate measurement procedures" (Kane, 1982, p. 151).

This latter view suggests that the validity evidence that a certain dispositional attribute has been measured (construct validity?) is not dependent upon evidence of a nomological net. Given the state of theory construction in education that is a good thing!

There is wide agreement that "Evidence identified usually with the criterion-related or content-related categories... is relevant also to the construct-related category" (AERA/APA/NCME, 1985, p. 9). Thus, test development procedures, test formating, administration conditions, reading/language level of the test, and internal consistency estimates are all relevant data for inferring the measurement of a construct (AERA/APA/NCME, 1985, p. 10).

Because all such data and procedures are typically well documented for teacher licensure tests there exists considerable evidence one could call "construct validity" evidence. The comment to Standard 11.2 quoted earlier suggests that

> "Good performance on a certification examination should not require more reading ability, for example, than is necessary in the occupation. The job analysis procedures used in establishing the content-related validity of a test can also contribute to the construct interpretation. One may show, for example, that qualified experts helped to define the job, identify the knowledge and skills required for competent performance, and determine the appropriate level of complexity at which these knowledges and skills should be assessed" (AERA/APA/NCME, 1985, pp. 64-65).

Certainly it readily can be inferred that minimally competent professional educators (to keep up in their professional literature, read the principals' memos, and indeed read the material they assign their students) need to be able to read at a level higher than that required of multiple-choice tests. The job-analysis, content validity evidence discussed earlier in this paper is usually available for well developed licensure tests. There are some criterion related validity studies done locally and/or that can be generalized from other studies, and internal reliability estimates typically suggest that only one construct is being measured by a test. Possible sources of error such as college graduates not being able to take multiple choice tests or not being motivated for a licensure examination can be ruled out thus eliminating competing hypotheses for what it is the test is measuring (e.g. test taking skill or motivation).

Given certain assumptions about the construct(s) being measured, other procedures could be used such as Guttman's scalogram analysis, factor analysis, experimental studies, and the multitrait-multimethod approach (Hambleton, 1984).

If all the above mentioned procedures are acceptable for establishing construct validity, then builders/users of teacher competency tests can and do provide construct validity evidence. What cannot be provided very easily, is evidence that a teacher competency test measures some broad, general theoretical notion such as the worth of a teacher. As has been pointed out by a variety of writers (see, for example, Darling-Hammond, Wise, and Pease; 1983), the evaluation of teaching in any generic sense depends on one's conceptions. The Medley distinctions made early in the paper between teacher competence, teacher performance, and teacher effectiveness must be kept in mind. While teacher competence may be related to teacher performance and effectiveness, licensure tests measure the former, not the latter two. Builders/users should not imply they measure the latter two, and they should not be held responsible for any evidence (or lack of evidence) by those who inappropriately wish to make such inferences.

Curricular Validity

In general, experts on licensure examinations do not discuss what some educators refer to as curricular or instructional validity. Licensure tests are designed to protect the public and the appropriate judgment of validity should be based on whether or not the tests cover the knowledges and skills that those licensed should possess. For the purpose of the licensure decision, it is irrelevant and inappropriate to consider curricular validity in judging the quality of the test.

The confusion that exists among some people regarding curricular validity in educational licensure probably arose for two reasons: (1)

confusing the situation in the Debra P. case with licensure decisions (see Rebell, 1986a), and (2) forgetting the original purpose of the NTE and the reason for the NTE Guidelines. The Debra P. case related to whether it was legal to deprive a high school student of a diploma based on a minimum competency test. An appellate court ruled that it would be considered unfair to withhold a diploma from those who did not learn unless, through the curriculum/instruction, they had been given an opportunity to learn the material. (For those of you not aware of the case, the state won because it demonstrated that the test did have curricular validity.) Of course that is all irrelevant to the quality of a licensure examination.

> "The criterion of 'job-related' validity is different from 'instructional' validity as argued in the Debra v. Turlington (1981) case. These two perspectives are opposite in outlook or goal direction. From the licensing examination perspective, job-related validity looks to the future or practice-related competence, whereas instructional validity looks at the relationship of the examination with past instruction/training...a licensing agency that addresses itself to instructional validity instead of job-related validity would be considered somewhat irrelevant to the societal concerns and problems at stake today" (D'Costa, 1985, p. 2).

The Standards (AERA/APA/NCME, 1985) implicitly recognize the legitimacy of the distinction between the two uses. Although they do not use the term curricular validity, they do address the notion in Chapter 8: Educational Testing and Psychological Testing in the Schools. Chapter 11: Professional and Occupational Licensure and Certification, makes no mention of such a standard.

The NTE Guidelines state that: "The primary function of NTE tests is to provide objective, standardized measures of the knowledge and skills developed in academic programs... (ETS, 1983b, p. 2). Given that primary function, the guidelines for the proper use of the NTE stated that one component for conducting a validation of the NTE tests for certification is "an assessment of the appropriateness of the tests' content, given relevant teacher-training curricula ... "(p. 9). They also suggested that the

certifying agency should:

> "Validate the tests to determine that they measure a representative sample of the knowledge and skills required for certification of beginning teachers..." (p. 8).

The published NTE Guidelines quote the federal district court ruling in the South Carolina case that the tests are

> "a fair measure of the knowledge which teacher education programs in the state seek to impart....there is ample evidence in the record of the content validity of the NTE....The NTE have been demonstrated to provide a useful measure of the extent to which prospective teachers have mastered the content of their teacher training programs" (p. 21).

That decision seemed by many to be reasonable. The tests were fair, and they did what the Guidelines stated was the primary purpose of the test--to provide measures of the knowledges and skills developed in academic programs. What that has to do with the quality of the test as a licensure examination is hard to determine. One could argue that because the acad mic programs are good programs, covering appropriate knowledges and skills, then a test measuring those knowledges and skills would be a good test. But one of the whole purposes behind licensure examinations is that the public does not wish to depend upon the quality of the educational/training programs. As was discussed earlier in this paper, some researchers feel that as many as 50% of the colleges of education should be shut down. It would make little sense to build a licensure examination based on the curriculum of an inadequate college! Roth provided a brilliant summary of the South Carolina case.

> "In the United States v. South Carolina case, the Plaintiffs presented only one alternative, graduation from an approved teacher training program, to the use of the NTE for certification purposes. The trial Court did not feel that the alternative would achieve the State's purpose in certifying minimally competent teachers as well as the use of the NTE. The Court in support of this finding made two points. One, evidence demonstrated that the teacher training programs varied in admission requirements, academic standards, and grading practices. Two, evidence demonstrated that the State approves only general subject matter areas covered by the programs, not the actual course content of the programs. Both of these points

> would seem to weigh negatively on the Court's position that validation against the teacher training programs was sufficiently reflective of actual knowledge needed for the teaching positions. Here the Court would seem to be admitting that the twenty-five teacher training programs were in fact different and therefore not all would be to the same degree reflective of knowledge needed to competently perform the job. The Court, however, while finding the teaching programs themselves an inadequate measure of teacher competency saw no inconsistency in finding test validation against those same teacher programs acceptable" (1984, p. 4).

Roth went on to argue that the validity question for licensure examinations is job relevance, not training program relevance. As was discussed under the section on content validity, this is the commonly--almost universally--accepted position.[2]

Most states using the NTE have "validated" them both for their match to the colleges' curricula and to the requirements of the job. Several states have referred to their studies involving a match between the test content and the curricula content as content review and have considered it a part of the validation process (Garvue, et al., 1983 for Louisiana; Echternacht, 1985 for Maryland; Hankins & Hancock, 1984 for Mississippi; ETS, 1983c for New York; and Hall, 1984 for New Mexico). Other states have referred to the job relevance question as content validation and have referred to the curricular match question as instructional validity (Cross, 1984 for Virginia), as opportunity to learn (Echternacht, 1983 for Delaware), or as adequacy-of-preparation (IOX Associates, 1983 for Kentucky). Roth (1984, for Arkansas) did not gather any evidence on adequacy of preparation.

There is certainly nothing wrong with doing a study to determine whether or not students have been given the opportunity to adequately learn what is in a licensure examination. What would be wrong would be to leave questions on essential knowledges and skills out of an exam (or not score them) because they were not in some curriculum. Further, the general

directions concerning job relevance as worded by many states validating the NTE tests had to do with whether the questions asked on the NTE were relevant. This is not a good way for finding out whether or not the tests left out some objectives whose mastery was essential for the protection of the public. The point here is not that the NTE tests are inadequate as licensure examinations. If, in the original construction, adequate attention was paid to the types of considerations covered under the content validity section of this paper, then the NTE tests are, no doubt, adequate. However, the primary purpose of the tests as stated by ETS was not to protect the public, but to measure the knowledges and skills developed in academic programs. The "validation" studies by the states do not address whether "harmful if missed" objectives have been assessed.

It seems reasonable to conclude that the methods used by the states for validating the NTE tests (and setting their cut scores) minimize the chances for false rejects and increase the chances for false acceptances as regards protecting the public. If a prospective teacher has not learned an adequate amount of what is both in the curriculum and considered relevant, then the person probably does not have a sufficient amount of the essential knowledges and skills to be licensed. However, a person could have mastered the specific knowledges and skills validated and tested and still not have some other essential but nontested knowledges and skills. (Of course the use of any test decreases the number of false acceptances from what one would obtain if no licensure test were required.)

A reasonable argument can be made that if the State Department of Education has oversight responsibilities over both the program approval of colleges of education and the content of licensure examinations, there should be a relationship between them. That relationship will, no doubt, exist in most states for most objectives. It obviously has existed to some

degree for all those states which have approved (validated) the NTE tests. But if it does not exist, and if the licensure examination has appropriate content validity as described in an earlier section, the indictment is against the program, not the licensure examination. In Florida, to ensure a relationship, the competencies in their licensure test "have been adopted by the Board of Education as curricular requirements for teacher education programs in Florida's colleges and universities" (State of Florida, 1982, p. 1). Other states' curricular guides for colleges of education also show strong relationships to the licensure examinations. The relationship of course holds only for tests over knowledge of the profession of education. It can not and should not occur for licensure exams that cover basic skills such as reading, writing, and basic mathematics. These should not be taught as part of the curriculum of a professional school. The competencies in subject matter such as that taught in secondary schools should also not be covered by the colleges (departments) of education although they, perhaps, have some responsibility for assuring that graduates have competent prerequisite skills in those areas as well as necessary subject-matter college course work somewhere in the university (college).

As mentioned in an earlier section of this paper, there are several problems in the use of program approval to assure quality control. If colleges graduate individuals who have not been given the opportunity to learn the necessary knowledges and skills required in the profession to protect the public, what should we do? We might consider closing down those colleges or bringing about additional pressure for them to do a better job. A state might even consider giving an inadequately prepared student free remediation (assuming the inadequate preparation is the institution's fault, not the individual's fault). What the state must <u>not</u> do is to give an inadequately prepared graduate a license to teach!

At times it has been suggested that a licensure test is an inappropriate measure for assisting in evaluating the professional curricula of colleges if the tests have not been built based upon the college curricula or instructional objectives. That notion is based on a grevious confusion between curricular and instructional evaluation. If one is evaluating the efficacy of the instruction then it is important for the test to match the instructional objectives. However, if we wish to determine whether or not a college is teaching (and/or the students are learning) the material deemed crucial for professionals to know, then the test must be based on that material--not the material that happens to be taught. Given the frequently quoted statement from Cronbach it would seem this confusion should have ended twenty years ago:

> "In course evaluation, we need not be much concerned about making measuring instruments fit the curriculum...An ideal evaluation might include measures of all the types of proficiency that might reasonably be desired in the area in question, not just the selected outcomes to which this curriculum directs substantial attention" (Cronbach, 1963, p. 680).

Finally, in this section, it seems worthwhile to point out how demeaning it is to the profession to suggest that teachers should not be expected to be independent learners.

> "While law students are seen as individually responsible for acquring a considerable body of skills and knowledge prior to entry into a true profession characterized by autonomous decision making, the common perception is that teacher candidates are products molded by their schools ... . Consequently, while lawyers and law schools are accorded high status, teachers and schools of education are not" (Guifford, 1986, p. 254).

The Cut Score -- An Aspect of Validity

The purpose of this section is not to review all the many methods of setting a cut score. They have been reviewed in detail elsewhere (see, for example, Berk, 1986; Jaeger, 1986; Livingston and Zieky, 1982). There is considerable debate about what method is "best." In discussing the various

drafts of the Standards, Linn stated that while earlier drafts of the Standards contained a standard dealing with the cut score of licensure tests eventually

> "it was concluded that there was not a sufficient degree of concensus on this issue within the area of certification and licensure testing to justify a specific standard on cut scores within this chapter" (Linn, 1984, p. 12).

Avoiding debates over specific cut score methodologies, there are still some cut score issues worthy of discussion. They basically center around the issues of supply and demand, the costs of false rejects and false acceptances, and the public perception of the cut scores.

Generally, writers in the field of licensure examinations have suggested that supply and demand considerations are not relevant to the cut score decision. There has been particular concern that licensure not be used by those already licensed as a way to regulate supply and thereby economically benefit themselves. Consider the following quotes:

"Since a major purpose of licensing is to prevent the unqualified from practicing, it follows that licensing should, by definition, be exclusionary: it should exclude from practice those who do not meet a predetermined standard. Those who do meet the standard should be licensed and allowed to practice. But licensing should not be used as a way to regulate the supply of practitioners for the economic benefit of those in a given occupational group" (Shimberg, 1982, p. 35).

"The process of determining a cut score for licensure and certification examinations is different from that in employee and student selection. ...There is not an explicit limit or the number of people that can be considered qualified. Cut scores associated with selection or classification uses of tests, on the other hand, are influenced by supply and demand..." (AERA/APA/NCME, 1985, pp. 63-64).

"Except in situations where a licensing board is misusing its licensing powers for monopolistic purposes, there is no fixed number of licenses that may be issued. If all applicants are qualified, all should be licensed. If none are qualified none should be licensed. The fact that no jobs exist should not, in theory, determine the passing rate" (Shimberg, 1984, p.3).

All of the above quotes state quite firmly that supply and demand are irrelevant. They do not specifically address the issue of the costs of false acceptances and false rejections. Pottinger addresses that concern as well as several others.

> "Licenses are often restricted to those whose test scores are higher than minimal levels required for competent performance. This is especially true when cut-off scores are determined by (1) manpower supply and demand in the profession, (2) the desire to minimize false-positive measurement errors, (3) the desire to 'upgrade' the profession, or (4) other 'arbitrary' decisions about who should be allowed to enter the professions.
>
> Such occurrences discriminate unfairly against those who are competent but are selected out of occupational opportunities by those who believe in the simple equation: Higher test scores mean better job performance. The tacit assumption that superior abilities in all measured skills or characteristics are desirable for performance is highly questionable" (1979, p. 41).

At a theoretical level, there is much in all the above quotes with which to agree. The purpose of a licensure examination is to protect the public from incompetents. It is not, like an employment examination, designed to predict levels of productivity among those who pass the test. Indeed it has been argued that licensure examinations should not be required to have predictive validity partly because of their purpose and design (and partly because of criterion measurement problems). However, there are degrees of competence or incompetence. There are degrees of danger to the public. Further, tests are never designed perfectly. Many licensure tests, in fact, have many of the same characteristics as employment tests. Schmidt and Hunter suggest the following about employment tests:

> The problem is that there is no real dividing line between the qualified and the unqualified. Employee productivity is on a <u>continuum</u> from very high to very low, and the relation between ability test scores and employee job performance and output is almost invariably linear. ... No matter where it is set, a higher cutoff score will yield more productive employees, and a lower score will yield less productive employees. ...it means that if the test is valid, all cutoff scores are 'valid' by definition. The concept of 'validating'a cutoff score on a valid test is therefore not meaningful" (1981, p.1130).

In theory, employment tests should be positively correlated with the criterion true score above the cut score and, in theory, licensure tests need not be. However, in actual practice, questions get placed on a licensure test because they are judged to measure essential knowledge or skills. Then, some group of people determine that only a certain percent of these need to be answered correctly in order for a person to be licensed. Surely the higher the percent of these essential items that an individual gets correct, the less danger to the public. Surely, the higher percent of items correct, the more competent the person. In fact, to be totally competent, a person would have to score 100% on a test. Most would agree that competence is a matter of degree rather than kind and there is no single point on the continum that separates the competent from the incompetent (see Jaeger, 1986, p. 195). Due to the minimum level of many teacher competency tests and the criterion problems, one should not expect to find a great deal of predictive validity with any observed criterion. Nevertheless, logic suggests that knowing more essential skills is better than knowing less; and if one had a measure of the true criterion, one might well expect to find a positive slope for the regression line of true criterion on test score.

Assuming positive slope, a reasonable position to take is that supply and demand, costs of false rejects and false acceptances, and desire to upgrade the profession <u>should</u> all be related to the cut-off score. In fact, supply and demand concerns are logically related to costs of false rejects and false acceptances. If a person were quite ill, but no licensed M.D. were available, that person would probably prefer going to a non-licensed graduate of a medical college than going to someone with no medical training whatsoever. Particularly if the graduate was a false positive who failed the examination! If there were generally a shortage of doctors, it might make

some sense to lower the qualifications a bit. If there were generally a surplus, it might make sense to raise the standards.

Whatever the merits of the views expressed above, it is obvious that, in practice, the cut scores on tests for the licensure of teachers have not been placed high as a way to regulate the supply of practitioners for the economic benefit of current teachers. In fact, there is evidence to suggest that

> "qualifying scores may simply represent minimal levels of proficiency that are politically acceptable and that do not threaten to reduce the supply of teachers" (Gifford, 1986, p. 253).

However we currently have a shortage of teachers in some areas and many are predicting a general shortage of teachers in the near future. Some have suggested that teacher competency exams have exacerbated the problem. Should we lower the cut scores to bring supply and demand into better balance? Some would suggest we should:

> "Recognition of teacher supply and demand problems is certainly part of the proper exercise of protecting the public. . . . Obviously, having no teacher in a classroom is less preferable than a teacher who has some knowledge" (Boyd & Coody, 1986, Part II, p. 26).

Rebell & Katzive also support the relevance of supply and demand in setting the cut score and reference two recent court cases where there were rulings specifically citing supply and demand as a consideration in setting a cut score (1986, p. 65).

Given the already perceived low standards for entering the teaching profession, others would not wish to lower standards to alleviate shortages.

> "The standards for entering teachers must be raised. ... The time-honored response to teacher shortages is to lower standards for entry into the profession. But the only way to make sure the country gets the kind of teachers it needs is to raise them to levels never met before" (Carnegie Task Force, 1986, p. 35).

> "If we allow the teacher shortage to become an excuse for staffing classrooms with anything less than the most competent, best trained, and fully certified teachers, public education in the United States could be headed for a real downward spiral. I

> am proposing, instead, a controversial but educationally honest method of dealing with teacher shortages: leave the classrooms vacant, rather than fill them with lower-quality substitutes" (Watts, 1986, p. 723).

Sykes discusses the tradeoff between standards and amount of services provided as follows:

> "For the most part, the elimination of low quality services is reckoned a benefit of standard-setting, but there may be hidden social costs. Consider, for example, this tradeoff: three persons out of ten have access to high quality service, while the rest receive no service or low quality service: or eight of ten receive middling service. Raising standards to enter professional practice may improve the quality of individual service but reduce access to that service, while excluding lower quality service providers from the market, who might be willing to work in poorly served locales" (1986, pp. 6-7).

There is at least some tentative evidence to suggest that the public does not wish to lower standards in education to relieve shortages. In a Gallup poll the following questions was asked with results as reported:

> If your local schools needed teachers in science, math, technical subjects, and vocational subjects, would you favor or oppose those proposals?
>
> Increasing the number of scholarships to college students who agree to enter training programs in these subjects?
>
> | | |
> |---|---|
> | Favor | 83% |
> | Oppose | 11% |
> | Don't Know | 6% |
>
> Relaxing teacher education and certification plans so more people could qualify to teach these subjects?
>
> | | |
> |---|---|
> | Favor | 18% |
> | Oppose | 74% |
> | Don't know | 8% |
>
> (Gallup, 1986, p. 55).

Although there is disagreement about the supply/demand issue, it is obvious that the placement of the cut scores <u>has</u> been influenced by people's beliefs about the relative costs of false rejects and false acceptances. States, in general, have gone through some procedure (such as Angoff's) to get some judgmental standard regarding what a <u>minimally</u> qualified candidate should know in order to be licensed. They have then <u>reduced</u> this

score by anywhere from one to three standard errors of measurement! For example, Virginia reduced the cut score by

> "two standard errors below the derived standards in order to minimize the probability of misclassifying an individual as 'incompetent' solely as a result of measurement or sampling error" (Cross, 1984, p. 15).

Several states (e.g. Alabama, Mississippi, and Louisiana) have actually set their cut score three standard errors of measurement below the standard obtained from their cut-score study. This means that there is a 50% chance of licensing an examinee whose true score is 3 standard errors below the judged standard while there is less than a probability of .0014 that a person whose true score is equal to the standard will not be licensed! Given that a person has repeated opportunities to take the test, there is virtually no chance that a person whose true score was above the judged standard would not be licensed. However, after three attempts, 87.5% of those whose true score was 3 standard errors below the judged standard would pass the test. Obviously, the only legitimate rationale for this approach is that false rejects are considered much more expensive than false acceptances. Neither the public nor I would believe this given a sufficiently large pool of applicants.

Some measurement experts testify as if this dropping of the cut score were not minimizing the chances of false rejects! Consider the following dialogue from a deposition:[4]

> Attorney for the State: Is it correct that using or dropping three standard deviations from the raw cut score, if I might call it that, the actual score of a person--
> Expert for the plaintiffs: The score that this committee comes up with.
> Attorney: Yes. Dropping three standard deviations from that should give you a 99 percent confidence level...
> Expert: Well, that's their argument, but basically it just sets a new cut score, it just changes the cut score, that's all it does. It takes a cut score of 89 and makes it 80, that's all. And that doesn't speak to the problem of the people who failed by one or two items or three items on a test that has defective items, it just doesn't speak to that issue.

The public should rightly be concerned about the profession's apparent concern for false rejects and its lack of concern for false acceptances. An expert witness for the state in the same case testified as follows:

> "...it's really a question of rights. Do you worry about a misjustice done to an individual candidate or do you worry about a misjustice that's done to the children who are being taught. And there is no right answer to that. I think the State has a compelling interest to worry about the rights of the children. And everything has been done to protect the individual. ...The parents are the ones who are being taken over the coals on this. Not only is that standard not the standard gotten by the counting of the items that they should know, but it's been lowered not one, not two, but three standard errors. And in addition to that, on some occasions it's been lowered further by the State Board of Education. And in addition to that---in addition to that, the candidate can try again. So everything is in the protection of the candidate, at the risk of the children. And what kind of protection is the State giving to the parents and the children when they allow that kind of standard-setting?

Busch & Jaeger suggest that

> "all standards are fallible, therefore standards should be set so as to minimize errors that penalize individual examinees" (1986, p. 17).

But that seems like a non-sequitor to me. We need to consider the purpose behind the movement for teacher competency exams. The public believes that some current teachers are not competent enough. They would like to see procedures implemented to reduce the supply of incompetents. The public is concerned with false positives not false negatives and the purpose of licensure is to protect the public. Would the public be impressed with the standards set for educational licensure exams? Would they be impressed that we have, in several states, intentionally set the cut score 3 standard errors of measurement lower than the standard recommended by a qualified panel of experts? If the general public took our professional exams (the pedagogy exams, not the subject matter exams) would they be impressed at how much we expect our professionals to know? How impressed would they be at the cut scores for those basic skills exams used in some states?

Have measurement experts, who have advising the policy makers that set the cut scores, made clear the implications of reducing the cut score by some function of the standard error regarding the proportion of false positives and false negatives? If those who have the authority to make the decisions wish to reduce the false negative error rate to essentially zero and to increase the false positive error rate, fine. They might, because of their fear of law suits from individuals who fail the tests. Busch and Jaeger may, unfortunately, be correct when they suggest that: "It is likely that the courts will view favorably, a standard-setting procedure in which the rights of the individual examinee receive greater deference (1986, p. 17). However, to be faithful to their charge to protect the public the policy makers ought to be more concerned with false positives who teach than with law suits from those who fail. They also should consider "if they are willing to risk the quality of education and the lawsuits by parents whose children were assigned to teachers scoring 3 standard errors below the minimum standard" (Mehrens, 1986, p. 10).

Finally, I have a suggestion for those who are concerned that our cut scores are too high. If we consider education a profession, if we believe in standards, have pride, and have a competitive spirit we could try the following. Give the bar examination and the medical licensure exams to the general public. Determine how many standard deviations the cut score is above the mean performance of the public. Give our pedagogy exams to the public. Set our cut scores the same number of standard deviation units above the public mean as the average that exists for the other two exams. (If we are not competitive, maybe we should set it at the lower of the two.)

In summary, I believe this whole issue of whether cut scores on licensure tests should be influenced by supply/demand, relative costs of misclassifications, and desire to upgrade a profession is deserving of more consideration.

Reporting Results

Under the section on content validity it was mentioned that one should communicate the domain of the licensure test to the public. If the objectives actually on the test are only a sample of the total set of objectives in the domain, and if one wished to make inferences to the competency of teachers in the total domain, it would impede the accuracy of the inference to communicate the specific objectives sampled by the test. The broader issue of communicating the results of the tests is discussed in this section.

The ETS Standards for Quality and Fairness state in their Score Interpretation Procedural Guidelines that the testing organization should "provide score interpretation information for all score recipients in terms that are understandable and useful to each category of recipient" (ETS, 1983a, p. 18). As Shannon (1986) suggests, that guideline is somewhat vague. What is meant by "score recipient," "categories," and what criteria should be used to determine what is "useful?" Vorwerk and Gorth (1986) submit that the examination results should be reported to four parties: individual examinees; the colleges and universities the certification applicants attended; the state which must determine whether certification should be granted; and finally, the public should receive aggregate results.

The categories for the score recipients for licensure tests are "pass" or "fail." However it is generally considered wise to report out using a continuous score scale along with the scaled passing score for failing candidates. Some experts suggest the actual score is not useful for passing candidates (see Shannon, 1986, p. 36). Such scores could lead to inappropriate ranking.

With respect to what is useful information, there is considerable discussion about the necessity or value of reporting subscores. If they are

reported, there is considerable discussion about what the format of the subscore reporting should be like.

In general, licensure tests are not primarily designed to be diagnostic. They are designed to categorize individuals into two groups: those sufficiently competent and those not. Because of that, they have been (or should have been) designed to maximize the reliability and interpretability of the total test scores (see Shannon, 1986, p. 7). At the same time, most tests have content outlines that permit the breakdown of the scores into sub-test scores. There is a natural press to wish to use subtest results to guide both those who have not passed and wish to retake the test as well as those who have responsibility for the training/education of subsequent candidates. Thus sub scores are frequently reported.

Shannon (1986) discusses at length some distinctions between CRTs (he puts licensure tests in this category) and diagnostic tests. Although the two types of tests are different he points out that CRTs are often used for diagnostic purposes to provide examinees with specific information. He stresses the limitations of this:

> "Although CRT subtest scores may provide examinees with a general indication of subject matter strengths and weaknesses, they tend to be ineffective at revealing causes underlying failure (e.g. deficiencies in instruction). Subtest scores might indicate which broad skill areas should be emphasized in preparing for retesting, but would not indicate specific skill failures or suggest learning strategies" (p. 7).

As Millman, (1986, p. 3-38) points out, the AERA/APA/NCME Standards do not require subtest score reporting because such subscores are not used in the making the licensure decision. The key Standard is in the chapter on licensure and certification:

> "Standard 11.4: Test takers who fail a test should, upon request, be told their score and the minimum score required to pass the test. Test takers hould be given information on their performance in parts of the test for which separate scores or reports are produced and used in the deicision process" (AERA/APA/NCME, p. 65) (emphasis added).

If subscores are reported, Standard 2.1 may apply.

> "Standard 2.1: For each total score, subscore, or combination of scores that is reported, estimates of relevant reliabilities and standard errors of measurement should be provided in adequate detail to enable the test user to judge whether scores are sufficiently accurate for the intended use of the test (Primary)" (AERA/APA/NCME, 1985, p. 20) (emphasis added).

Note the emphasis added to the above quote. It seems possible to argue that the reporting of the subscore reliabilities is not necessary because the intended use of the test is for making licensure decisions. But if that is so, why report the subtest scores in the first place. Is there not an implication that they will be used for something? Probably. Thus, I take the position that if the subscores are reported, their reliabilities ought also to be reported. The danger is that these subscore reliabilities will be smaller than someone's arbitrary cut off score for reliabilities. One would hope that in any court battles over licensure tests judges could recognize that a test may have low sub-score reliabilities and yet be quite useful for its primary purpose--determining the competency of the applicants.

Gifford (1986, p. 266) takes the strong position that licensing tests should not be used as diagnostic tools because of the low reliability of the subscores. Gabrys, however, suggests that the second goal of teacher comeptency testing programs "is to provide diagnostic information about candidates' strengths and weaknesses to the candidates and to the teacher training institutions" (1986, p. 85). In actual practice, most states report subscore information to the recipients. The best metric to use for the subscores is beyond the scope of this paper. See Millman (1986) and Shannon (1986) for some thoughts on that issue .

Although not directly tied to reporting, it should be mentioned that it is fairly common for states to produce study guides for applicants. See Weaver (1986) for a brief discussion of such guides and their effects.

Legal Issues

This section contains a brief overview of a few of the legal issues that pertain to the validity of licensure examinations. Two points must be kept in mind while reading this section. It is certainly not based on a thorough review of the literature nor was it written by an attorney.

Licensure examinations are considered different from employment examinations by the legal profession just as they are by psychologists. The legal profession does not have any more accord with respect to the legal issues than measurement experts do with respect to the measurement issues. However, there is some general, if not universal concensus about some matters. There appear to be three general theories of how to attack licensure exams: (1) under Title VII of the Civil Rights Act of 1964, (2) under the Constitution of the United States, and (3) under anti-trust law (Pyburn, 1984). Only the first two will be discussed here.

There has been considerable discussion in the literature regarding whether Title VII applies to licensure laws. If Title VII does apply, plaintiffs would prefer to use that approach in their attacks on licensure tests. That is so because in Title VII cases, as opposed to Constitution based cases, the plaintiffs do not have to prove discriminatory intent if there is a disparate impact on a protected group. Further, if Title VII does not apply, then the Uniform Guidelines on Employee Selection Procedures would also not apply although they "have been given great weight by the courts in Equal Protection as well as Title VII cases" (Eisdorfer & Tractenberg, 1977, p. 121).

The Federal Agencies that developed the Uniform Guidelines speak to this issue as follows:

> "Whenever an employer, labor organization, or employment agency is required by law to restrict recruitment for any occupation to those applicants who have met licensing or certification requirements, the licensing or certifying authority, to the extent it may be covered by Federal equal employment law, will be

> considered the user with respect to those licensing and certification requirements" (EEOC et al., 1978, p. 38308, emphasis added).

This suggests that applicability is dependent upon the interpretations given to Federal equal employment law. Shimberg pointed out in 1981 that

> "Although the U.S. Supreme Court has not ruled on the question there seems to be a clear trend in the few federal circuit court of appeals cases that have considered the issue. These decisions suggest that Title VII does not apply to the licensing activities of state agencies" (Shimberg, 1981, p. 1145).

Shimberg footnoted opinions in Tyler v. Vickery and Richardson v. McFadden to support his statement. Pyburn (1984, pp. 4 & 5) flatly stated that Title VII does not apply to licensing tests quoting from two cases:

> "Title VII does not apply by its terms...because the Board of Bar Examiners is neither an 'employer', an 'employment agency', nor a 'labor organization' within the meaning of the statute." Tyler v. Vickery, 517F. 2d 1089, 1096 (5th Cir. 1975), cert. denied, 426 U.S. 940 (1976).

> "[The] principles of [Title VII] test validation do not apply to professional licensing examinations." Woodard v. Virginia Board of Bar Examiners, 420 F. Supp. 211, 18 FEP 836, 838 (E.D. Va. 1976), aff'd per curiam, 598 F. 2d 1345 (4th Cir. 1979).

At least several other writers share this opinion regarding the non-applicability of Title VII and the Uniform Guidelines (e.g. Vertiz, 1985, Werner, 1985 and Herbsleb, Sales and Overcast, 1985).

Rebell (1986b, p. 60) suggested that there was an unresolved technical issue regarding the applicability of Title VII to licensing agencies. Freeman & Hess also believe the matter "unresolved" (1985, p. 6).

> "...the unsettled question of the status of teacher certification--occupational license or employment criteria--suggests that the applicability of Title VII and its stringent requirements is not as remote as in the case of other kinds of occupational licensing. To the extent that teacher certification is viewed as an employment criterion and not as analogous to other occupational licensing, all certification requirements, whether or not they take the form of examinations, become subject to Title VII provisions" (1985, p. 24).

Whatever various courts may decide regarding the applicability of Title VII and the Uniform Guidelines, there is some movement under way to revise the current guidelines.

> "Many employment testing experts say the guidelines are technically outdated and make costly demands on employers that are not justified by the latest research" (Cordes, 1985, p. 1).

If the updating occurs, it may make little difference regarding validity requirements whether or not Title VII applies to licensure tests.

The constitutional attacks are based on the Equal Protection and Due Process clauses of the Fourteenth Amendment. With respect to validity the courts have generally ruled that requiring a test for licensure is not a violation of either clause if there is a "rational connection" between the test and the job. This connection is frequently demonstrated through an appropriate job analysis as discussed in an earlier section (see Pyburn, 1984). Herbsleb, Sales and Overcast state that their analysis shows "the constitutional standard for rationality is so lenient that we need not delve into the more technical points; they are simply irrelevant to the legal issues" (1985, p. 1169). They believe that:

> "In order to satisfy the criterion of rationality, it is sufficient if (a) the test is designed by knowledgeable and experienced members of the field specifically for the use to be made of it, (b) the test content bears a plausible (not necessarily a demonstrated) relationship to the knowledge and skill required in professional practice, (c) the method of scoring and the cutoff score represent the reasoned judgment of qualified persons as to the minimal level of skill that should be required of a practitioner, and (d) no demonstrably arbitrary procedures are used in implementing it" (1985, p. 1170).

Social Considerations

The disparate impact of teacher competency tests on minorities is of social concern as well as legal concern. The evidence suggests that blacks do fail teacher competency tests at a higher rate than whites. For example, in California the pass rate is 76% for whites, 39% for Hispanics, and 26%

for blacks. In Georgia, the percents of passes on the first attempt are 87% for whites and 34% for blacks. In Oklahoma, they are 79% for whites, 58% for Hispanics and 48% for blacks (Goertz, Ekstrom, & Coley, 1984). Galambos summarized the data and the problem as follows:

> "In state after state, the results show failure rates among black candidates as high as two-thirds, while white applicants fail in the 10 to 30 percent range...There is no doubt that if such failure rates continue, minority representation in the teaching force will decline" (1984, p. 8).

It should be pointed out that due to the opportunity to retake the exams the black pass rate does rise subtantially from the numbers presented here (see Solomon, 1986). Nevertheless, projections indicate that black teachers will constitute only about 5% of the teaching force by 1990 whereas the school age population will be more than 30% miniority (Baratz, 1986). While probably no one would suggest the decline of black teachers is solely, or even primarily, due to teacher licensure testing most who have studied the issue believe testing has had an impact.

Many people believe much data suggests that a major reason for the difference in the pass rates has to do with the adequacy of pre-college education. Blacks, as a group, may not receive as high a quality public school education. In discussing the quality of the black high school pool from which teachers must come Baratz reports that while 53% of the nation's white eleventh graders can read at a level she would consider adequate for college work only 20% of the black students were at that level. Thomas & Tyler (1984), reported for example, that at Alabama State University (an historically black college) more than 89% of the freshman had not taken a college preparatory program in high school.

Evidence strongly suggests that at least part of the disparate impact of the tests may be due to the quality of the academic programs at predominantly black colleges. Ayres (1983) found that after controlling for

SAT scores blacks attending predominantly white institutions averaged 25 points higher on the NTE Commons than blacks attending predominantly black institutions. Whites attending predominantly black colleges scored 37 points lower on the NTE than comparable whites in predominantly white institutions. He concluded that "Uncontrolled, precollege differences among students may account for some of these differences in NTE performance, but the analysis suggests that the universities themselves are the more important influence" (1983, p. 291).

Hilldrup (1978) reported that one year in South Carolina only three percent of the seniors in the state's six predominantly black colleges passed the NTE. This study did not control for aptitude and there is no way to separate out whether the low pass rate was due to the low quality of the programs or the low quality of the entering students.

Whatever the reason for the disparate pass rates,

> "...The bottom line of the problem centers on what will have the greater negative impact on children in the schools: the lack of role models on minority children if black representation among teachers declines, or the possibility that teachers with less than the minimum qualifications will teach in the nation's schools" (Galambos, 1984, p. 9).

Reasonably enough, there are different opinions regarding the bottom line. Guifford, a black educator, states the case for continued testing as follows:

> "We must always be mindful that the effectiveness of our school systems will not be found in the statistics on the racial composition of our teaching staffs but rather in the statistics reflecting the mastery of basic skills in reading, writing, and arithmetic by all of our students... . There is no equity in absense of excellence. If we are to meet our moral and legal responsibilities to both the potential teachers in our population and to their future students, we must continue to employ valid, job-related written examinations of potential teachers' basic skills" (Gifford, 1985, p. 62).

Raspberry, a black columnist who frequently speaks and writes about educational issues, suggests the following:

> "There's a lot we don't know about educating our children, particularly disadvantaged children. That's a failure of information, which is bad enough.
>
> But we know a lot more than we are willing to act on. That is a failure of guts, which is worse...
>
> We know that a lot of our teachers aren't as good as they ought to be. But we--and here I mean specifically the civil rights leadership--balk at insisting that incompetent teachers be weeded out, particularly if they are minorities. We'd rather feel sorry for them, as victims of society, than hold them to standards that would improve the quality of the schools for our children. ...
>
> We can have well-educated children or ignorant teachers. We cannot have both" (Raspberry, 1983).

Probably the most important point to be stressed in this section is that one can do more than "merely lament the inevitable" disparate pass rates. As Gifford (1986) points out: "All of the knowledge and skills that are tested in competency examinations are learnable" (p. 264). Futrell, President of the National Education Association has stated that:

> "I've heard some say that pre-service testing may hurt women and minorities. ... As a black woman, I don't buy that. As a matter of fact, I resent it. If we set clear and demanding expectations and then help all potential teachers reach those expectations, we can have both quality and equality" (Quoted from Rebell, 1986a, p. 398).

Holmes warns us to "not buy the convential wisdom that asserts that raised standards cannot be met by blacks and other minority groups" (1986, p. 346).

Spencer reports that in 1978-79 only 5% of the students at Grambling State University who took the NTE passed it. Although at first Grambling and other black colleges in the state resisted the NTE requirement.

> "In time, the need to serve its students well caused Grambling's College of Education officials to move from resistance, and focus instead on improving the teacher-education program . ... Results have shown this course of action to be the wisest and most productive possible" (Spencer, 1986, p. 297).

Indeed, in 1983-84, 86% of the Grambling students passed the NTE.

Hackley (1985) describes a program at the University of Arkansas, Pine Bluff, which resulted in the pass rate on the NTE at that college to increase from 42% in 1983 to 73% in 1984. Solomon (1986) reports on a

predominantly black institution that had an initial pass rate of 39% but a cummulative pass rate of 78% in an exam for English and speech teachers. The solution to the problem of incompetent teachers be they black or white is to work at increasing their competence, not allowing them to teach in spite of their incompetence due to sympathy, guilt, or some perversion of the notion of justice.

How Valid Must A Test Be? Idealistic Vs. Realistic Standards

Two general questions have been debated by measurement experts regarding the validity of teacher competency examinations: How valid should the tests be, and How valid are they?

Some people would prefer not to give examinations unless they are the best they can possibly be. This view is taken even by people who admit that the state has an interest in protecting the public. Consider the following deposition interchange between an expert for the plaintiffs and the attorney for the state:

> Attorney: In this type of test is the most important concern the potential for harm to the individual who takes the test?
> Expert Witness: No; ... has an interest also in protecting schoolchildren from incompetent people, there is that interest.
> A: Is that an important interest?
> E: It certainly is.
> A: Is it a legitimate interest?
> E: It certainly is.
> A: How do you relatively weigh the two competing interests?
> E: Well, if ... chooses to do this then it seems to me that it is incumbent upon ... to do it right so that the instrument and the product that is used to make those decisions is the best possible thing that they could have developed to make those kinds of decisions.
> A: The best possible thing?
> E: Yes
> ...
> A: Is that the best possible test without regard to the financial ability of the study?
> E: If you can't afford to do it right don't do it.
> A: So budget constraints should not enter into considerations of how to develop the test?
> E: Not if I have the potential to harm individual people's

> careers and these people have gone through state-approved programs for four years.

Compare that testimony with that offered by an expert witness for the defense:

> Witness: I think one needs to keep in mind that it is, I believe, inappropriate to not use data which could facilitate decision-making simply because it is not the best possible data that you could ideally imagine having...
>
> ...
>
> Attorney for the plaintiffs: My question is simply this. At what point should we expect the instrument to be valid for the purpose for which it is used?
>
> Witness: Well, first of all, there are degrees of validity....And my belief and my testimony is that prior to using data, whatever the degree of validity, one has to consider the relative costs of the mistakes one makes without that data compared to the relative costs of the mistakes when one uses that additional data.
>
> ...
>
> I think there should be reasonable beliefs that with the use of these data, in addition with the data that have traditionally been used to make the decision, one will improve the decision-making process.

I believe the second witness is more in accord with the mainstream of psychometric thinking on this matter. If one never used tests unless they were "the best possible thing" one would never use tests. The crucial question is whether or not test data improve the decision making over and above the decisions that would be made without the test data. I would hope that the psychometric community could agree to agree on the question although they may well differ on the answer to it.

Of course, when competency tests are used as an additional criterion (not the sole one) to those criteria already used, the result is to decrease the number of false acceptances from the number previously made and to increase the number of false rejections. Thus it is the relative costs of these two errors that must be considered. Reasonable people can disagree with respect to those relative costs. But we need to keep in mind that the whole purpose of licensure (whether or not one uses test data) is to protect the public (i.e. to decrease the number of false acceptances into the

profession). Further, no matter what the relative costs of the two kinds of errors, using tests with even a little bit of validity will, with an appropriate cut score, result in a decrease in the total costs unless one takes the position that false rejections are infinitely more expensive than false acceptances.

Another way of looking at how valid teacher competency tests should be is to compare them to other licensure examinations. By and large, other licensure tests leave much to be desired. Shimberg (1985) reported on a study he completed with Esser and Kruger which found serious shortcomings in many board-developed licensure tests.

> "Few of the tests that they studied were based on an up-to-date job analysis, and rarely was there evidence of a test plan or specifications to govern test content. Many relied on essay and short-answer questions for which even board members could not agree on acceptable answers. Where performance tests were used, test administration conditions were frequently not standardized, explicit rating criteria were not available, and raters were untrained" (Shimberg, 1985, p. 9).

Werner provided us with the following insights.

> "Too frequently, test program development proceeds from a picture of occupational practice which is outdated, imbalanced with respect to various practice specialty areas, skewed toward matters of only academic interest, or insufficiently representative of practices which have the greatest potential for public harm....
>
> And in California, we amaze barber applicants each year by asking them to specify the average number of hairs on the human head while we neglect to assess meaningfully their knowledge of potentially harmful cosmetic chemicals" (Werner, 1982, pp. 7-8).

Finally, consider some quotes from a 1978 article on the Examination for Professional Practice in Psychology which was first released in 1964.

> "Test development for the AASPB 'National Examination' has always leaned heavily upon the voluntary participation of qualified psychologists throughout the APA. Items are contributed by psychologists recognized in their specialty area....There is no way to pretest new items or to establish norms in advance of publication. ..The item classification scheme, or list of content area categories, and the distribution of items among those areas are necessarily somewhat arbitrary. ...Further studies are contemplated comparing test scores with academic background, supervised experience, and certain evidences of satisfactory or

> unsatisfactory performance in the profession. The commitment of AASPB to a program of ongoing, thorough, validity study could hardly be stronger" (Carlson, 1978, pp. 491-492, emphasis added).

In fact, the commitment of AASPB was so strong that in 1980 they decided to "...initiate a research program to ascertain whether there might be a more objective, empirically-based method for determining examination content" (Rosenfeld, Shimberg, & Thornton, 1983, p. 1-2). In 1983 the results of the formal job analysis were published, 18 years after the test was first given for licensure purposes!

In preparing to write this paper I reviewed portions of the construction/procedures for the teacher competency tests used in at least 15 different states. Without exception the care in the test construction process (which determines the content validity) and/or the care in validating the questions (e.g. for the adoption of the NTE examinations) plus the reporting of those procedures exceeded what has typically been done in other licensure examinations.

Given the truth f the following points:

1. Current non-test licensure procedures are not based on job analyses and have many other flaws;

2. The public perceives that too many incompetents have been licensed to teach;

3. The purpose of licensure is to protect the public (not guarantee everyone--competent or not--a license);

4. That the addition of a test as one more criterion for licensure can decrease the number of false acceptances and can not increase the number of false acceptances; and

5. That current teacher competency tests have more validity evidence than many other licensure examinations;

It is easy to see why so many individuals are in favor of teacher competency tests. Further, it is hard not to feel proud of the job done by the measurement profession with respect to their response to the demand for better teacher licensure procedures.

FURTHER POSSIBLE RESERACH ON VALIDITY ISSUES

It should be clear to even the most causal reader that I believe current teacher competency tests, in general, are providing us with data that facilitates our current decision making processes with respect to teacher licensure. (This is not an endorsement of all such tests. I have not seen all such tests.) In simple laymen's langauge, the tests are valid enough to be used for the purpose for which they are designed. That does not mean that more research on validity issues would not be useful. One can certainly imagine studies that could bolster the validity claims of existing teacher competency tests, just as one could imagine validity studies that could bolster the validity claims of the alternate ways we have typically used in the past to make licensure decisions. Unfortunately, there is in our society a dual standard with respect to validity evidence. We expect more such evidence when the data used to facilitate decision making eminates from tests than when it eminates from alternate sources. Thus, an important preamble to this short section is that current tests provide inferences that are valid enough to justify current test use, and that the validity evidence is far stronger than the validity evidence we have for the other data used for licensure such as "three hours of mathematics," or "thirty credits of methods courses."

It is probably fair to say that all of the five steps discussed in this paper in the development of content valid teacher competency tests could benefit from further research. First, what procedures impact the development of the original list of competencies? Do different committees, or different instructions or time lines given to the committees result in different lists of competencies? How should we "validate" the competencies the committees produce? Must we accept them on faith? If not, what external criteria would we use? Of course there needs to be further

research on what Kane (1984) called the department of learning or science associated with the profession.

While I have not looked in detail at all the job analyses which have been done, it is reasonable to conclude that we would also profit from more research in that area. Is a survey of teachers really the best way to conduct a job analysis? Would we get different results if we were to send in teams of observers to observe for hundreds of hours? If so, how should we decide which one of the approaches leads to better data? Can teachers really rate competencies in terms of their essentiality? Would there be any better way to determine how essential a competency is? What impact would changes in the directions to teachers have on their judgments regarding the necessity of the competencies? Would a description of a competent teacher attached to the survey impact the results? If the surveys more strongly encouraged teachers to suggest new competencies, would the domains be less likely to exclude those harmful if missed competencies? All of these questions are researchable. At the current time we do not have sufficient evidence regarding how much the domains might change across different job analyses strategies. Of course none of this research would empirically answer the question of which strategy produces the "best" domain of necessary competencies. That would require criterion related validity research.

As mentioned in an earlier section, current standards in the profession do not demand that licensure tests have criterion related validity evidence. Such tests are not designed to predict degrees of relative performance, they are designed to measure necessary competencies. Of course there is an implied "prediction" that individuals <u>not</u> having the basic competencies will be more likely to harm the public (students) than those who do have the minimum competencies. This is the basis for claiming the

competencies measured in a licensure examination are necessary. How should one support the claim of necessity? If there were no practical design problems and no criterion measurement problems then one could employ a criterion related validity study. Such studies may prove useful in spite of design and measurement problems. However, several things should be kept in mind. The ideal criterion is degree of harm to the children. This is not the same as teacher performance. It is a subset of what Medley termed teacher effectiver s--the effect that the teacher's performance has on pupils. Obviously not all effects can be labeled harmful. Careful consideration would need to be given to what effects are to be considered harmful and what the operational definitions of those effects should be. One might believe, as I do, that it is harmful to students to be exposed to teachers who use incorrect grammar, spell words incorrectly on the board, and/or write poorly worded notes to parents. One might believe, as I do, that it is harmful to students to be exposed to teachers who do not know the specific subject matter they are teaching, or who know it so superficially that they cannot tie it together with previously learned or to be learned material, or who do not know the most efficacious methods of teaching the material to students from a variety of backgrounds. One might believe, as I do, that it is harmful to students to be exposed to teachers who do not know how to assess the learning of their pupils, who do not know how to organize learning materials, determine appropriate objectives, maintain classroom control, or recognize the advantages of aperiodic reinforcement over continuous reinforcement. The problem is to define and measure the harmful effects, and to show that the teacher performance, which at least theoretically can be measured, lead to the harmful effects. To me, harm has been inflicted if the student learns less than the optimal amount due to a teacher's lack of knowledge about basic communication skills, the subject matter taught, or

the pedagogy of teaching. To measure that harm in a research study would indeed be difficult. Others, of course, may have a much more limited definition of harmful.

Certainly the determination of what is meant by harmful and necessary could profit from further considerations. The constitutive definitions should preceed operational definitions. Once operational definitions were obtained surely one could, at least theoretically, conduct empirical studies to determine whether lack of "necessary" knowledge resulted in "harmful" effects on children. If not, the standard for necessary could be lowered and a new study conducted. As a graduate of a university known as the dustbowl of empiricism I cannot in good conscience argue against the potential value of such studies. Nevertheless, there are countless reasons why such studies may not have enough power to show a relationship between lack of teacher knowledge and harm to the student even if the relationship actually exists. Frankly, if a study fails to reject the hypothesis that a lack of "necessary" knowledge does not harm pupils the public may well doubt the results, and so might I. Our acceptance or rejection of the empirical results would of course vary depending on our subjective notions of how low or high the standards for "necessary" knowledge had been set!

Other correlational studies could also be conducted. We could continue to investigate whether the knowledge displayed on current tests correlate with a lot of other different measures. The correlations could be based on the actual scores on the test and/or the dicotomous decisions made from the scores. I am inclined to believe these studies would be useful. However, these studies should probably <u>not</u> be carried out by the licensure agencies. The reason is that someone may misinterpret the intent of these studies and argue that the agency is using the tests to predict degree of some other variable. Research scholars interested in the relationships between teacher

knowledge of basic skills, subject matter, and/or pedagogy and other variables should be conducting these studies. With years of research and considerable luck we might be able to establish a nomological net among a variety of relevant variables. I would not be inclined to view these studies as telling me any more about the validity of the test as a measure of teacher competence than the validity of the measure of the other variable. For example, if a teacher competency test does not correlate with grades in practice teaching, or scores obtained from some teacher performance scale, that low relationship does not tell me either that the test does not measure competency or that a grade or score on a performance scale does not measure performance. (Recall that performance, according to Medley, depends on teacher competence, the work context, and the teacher's ability to apply his or her competencies at a given point in time.) If I felt some other variable was so logically related to teacher competence that a low measure of relationship was an indicant that one of the measures lacked validity, I might well suspect the other measure. Certainly, on the face of it, one should place as much confidence in an achievement test as a measure of competencies as in a performance scale as a measure of performance.

The point of this brief and very general section on possible further research on the validity issues is that of course we should continue to research how to define and measure teacher competencies and to investigate their correlates. This research will be fraught with difficulties, but potentially valuable to the profession and to the general welfare of the public. While this research is being conducted we should continue to use the best data we have available to determine who should be licensed.

Conclusion

Some measurement experts fear that the public expects too much of testing with respect to solving the problems of education (Madaus, 1985). That is probably true. We should try to temper their expectations with realism. In so doing, we should admit the limitations of tests, but not downgrade their potential (actualized in many cases) for good. Given all the criticism of testing during the past decade or so, it is understandable that some of us may now be concerned about the public's hopes and expectations for the measurement profession. Years ago Atkinson discussed what happens when fear of failure is stronger than the motive to succeed. People with such fears choose either very easy tasks--so as to be sure and succeed, or they choose very difficult tasks thus providing themselves with an acceptable excuse for failure. As measurement experts we should neither stand too far back nor get too close before we toss the ring. We should not claim to be measuring teacher effectiveness but we should be willing to measure teacher competence.

An effective teacher licensure test will not eliminate the need for subsequent teacher evaluation; it will not cure all educational ills; it will not eliminate all ineffective teachers nor even all incompetent teachers. It should ensure that those individuals who are licensed have a minimal level of competence on some important sub domains of knowledge and skills relevant to their profession. That is a step in the right direction. We should not be afraid to take that step.

Footnotes

1 The Standards For Educational and Psychological Testing is a single book and purists may wish to follow reference to it with a singular verb. However, when shortening the reference to just Standards the plural form sounds more appropriate and will be used throughout the paper. The defense, in addition to the sound, is that there are a set of standards within the single book.

2 I would have preferred that Roth not have used the Uniform Guidelines as support for his position. There is much other literature, as well as basic logic, available to support his position and many individuals do not feel the Uniform Guidelines apply to licensure tests, a subject discussed later in this paper.

3 Exstrom & Goertz argue that accountability for student failure is often misplaced:

> "Although instruction in basic skills and subject matter areas is usually not provided in the schools of education, basic skills and subject matter specialty tests are used to evaluate the teacher education programs. Teacher education departments are held responsible for education students' knowledge of these areas while non-education departments actually providing the instruction have little or no incentive to improve their teaching in ways that will improve teacher quality" (1985, p. 9).

4 Specific names and references are not given for quotes from depositions. They serve no useful purpose. In general, expert witnesses' responses to attorneys are not as articulate as their writings and many may prefer not to be identified.

REFERENCES

AERA, APA, NCME. (1985). Standards for educational and psychological testing. Washington, D.C.: American Psychological Association.

Alabama State Board of Education. (1980, January 8). Minutes. Montgomery, AL. Author.

American Psychological Association. (1980). Principles for the validation and use of personnel selection procedures. Washington, D.C.: Author.

Associated Press. (1985,July 4). Teachers support dismissal resolution. Lansing State Journal, 3A.

Ayres, Q. W. (1983). Student achievement at predominantly white and predominantly black universities. American Educational Research Journal, 20, 2, 291-304.

Baratz, J.C. (January, 1986). Black participation in the teacher pool. Task force on teaching as a profession. Carnegie Forum on Education and the Economy. New York: Carnegie Corporation.

Benderson, A. (1982). The teachers in America. Focus, 10. 1-5.

Berk, R. A. (1984). Conducting the item analysis. In R.A. Berk (Ed.), A guide to criterion-references test construction (pp. 97-143). Baltimore: The Johns Hopkins University Press.

Berk, A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. Review of Educational Research, 56, 1, 137-172.

Bloom, B. S., Madaus, G. F., & Hastings, J. T. (1981). Evaluation to improve learning. New York: McGraw-Hill Book Company.

Board, C. & Whitney, D. R. (1972). The effect of selected poor item-writing practices on test difficulty, reliability, and validity. Journal of Educational Measurement, 9, 3, 225-233.

Boyd, D.R. & Coody, C.S. (1986). Defendant's Post-Trial Memorandum. Margaret T. Allen, et al. and Board of Trustees for Alabama State University and Eria P. Smith v. Alabama State Board of Education et al., Civil Action No. 81-697-N.

Burns, R. L. (1985). Guidelines for developing and using licensure tests. In J. C. Fortune & Associates (Eds.), Understanding testing in occupational licensing (pp. 15-44). San Francisco: Jossey Bass.

Busch, J.C. & Jaeger, R. M. (1986). The implications of legal issues for setting standards on minimum competency tests for teachers. University of North Carolina at Greensboro.

Carlson, H. S. (1978). The AASPB Story: The beginnings and first 16 years of the American Association of State Psychology Boards, 1961-1977. American Psychologist, 33, 5, 486-495.

Carlson, R.E. (1985). The impact on preparation institutions of competency tests for educators. Presented as part of a symposium entitled: The assessment boomerang returns: Competency tests for educators. American Educational Research Association, Chicago, IL.

Carnegie Task Force. (1986). A Nation Prepared: Teachers for the 21st Century. The report of the task force on teaching as a profession. Carnegie Forum on Education and the Economy. New York: Carnegie Corporation.

Coleman, J. S. et al. (1966). Equality of educational opportunity. Washington, D. C.: U.S. Department of Health, Education, and Welfare, Office of Education.

Connell, Christopher. (1985, June 24). NEA may endorse ce ification test. Lansing State Journal, p. 7A.

Cronbach, L. J. (1963). Course improvement through evaluation. Teacher's College Record, 64, 672-683.

Cronbach, L. J. (1970). Essentials of Psychological Testing (3rd ed.). New York: Harper and Row.

Cronbach, L. J. (1980). Validity on parole: How can we go straight? In W. B. Schrader (Ed.), Measuring Achievement: Progress over a decade. New Directions for testing and measurement (pp. 99-108). #5. San Francisco: Josey Bass.

Cross, L. H. (1984). Validation study of the National Teacher Examinations for certification of entry-level teachers in the Commonwealth of Virginia. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA.

Cross, L. H. (1985). Validation of the NTE tests for certification decisions. Educational Measurement: Issues and Practice, 4, 3, 7-9.

Cordes, C. (1985). Review may relax job testing rules. APA Monitor, 16, 5, pp. 1, 20 & 22.

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. Review of Educational Research, 53, 3, pp. 285-328.

Debra P. v. Turlington (1981). 644 F. 2d 397, 5th Cir.

D'Costa, A. G. (1985) Documenting the job-relevance of certification and licensure examinations using job analysis. Paper presented at the annual meeting of The American Educational Research Association, Chicago, IL.

Ebel, R. L. (1961). Must all tests be valid? American Psychologist, 16, 640-647.

Ebel, R. L. (1974). And still the dryads linger. American Psychologist, 29, 7, 485-492.

Ebel, R. L. (1975). The use of tests in educational licensing, employment, and promotion. Education and Urban Society, 8, 1, 19-32.

Ebel, R. L. (1977). Comments on some problems of employment testing. Personnel Psychology, 30, 55-63.

Ebel, R. L. (1979). Essentials of Educational Measurement (3rd ed.). Englewood Cliffs, N. J.: Prentice Hall.

Echternacht, G. (1983) The validity and passing standard for the pre-professional skills tests as a certification test for Delaware Teachers. Princeton, N. J.: Educational Testing Service.

Echternacht, G. (1985). Report of a study of selected NTE tests for the State of Maryland. Princeton, N. J.: Educational Testing Service.

Educational Testing Service. (1983a). Educational Testing Service Standards for Quality and Fairness. Princeton, New Jersey: Educational Testing Service.

Educational Testing Service. (1983b). NTE programs: Guidelines for proper use of NTE tests. Princeton, N. J.: Author.

Educational Testing Service. (1983c). Executive summary report on a study of the three NTE core battery tests by the State of New York. Princeton, N. J.: Author.

Eisdorfer, S. & Tractenberg, P. (1977). The role of the courts and teacher certification. In W. R. Hazard, et al, (Eds.), Legal Issues in teacher preparation and certification (pp 109-150). Washington, D. C.: ERIC Clearinghouse on Teacher Education.

Ekstrom, R. B. & Goertz, M. E. (1985). The teacher supply pipeline: The view from four states. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Elliot, S. M. & Nelson, J. (1984). Blueprinting teacher licensing tests: Developing domain specifications from job analysis results. Paper presented at the annual meeting of the National Council on Measurement in Education. New Orleans, LA.

Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor and Department of Justice. (1978). Uniform guidelines on employee selection procedures. Federal Register, August 25, 43, (166), 38290-38315.

Feinberg, Lawrence. (1985, March 13). NEA, panel support exam for new teachers. Washington Post, p. A8.

Feistritzer, C.M. (1983). The condition of teaching. Princeton, N.J.: The Carnegie Foundation for the advancement of teaching.

Feistritzer, C.M. (1984). The making of a teacher: A report on teacher education and certification. Washington, D. C.: National Center for Education Information.

Florida Department of Education. (1981). The Florida Teacher Certification Examination Bulletin IV: The Technical Manual. Tallahassee, FL.: Department of State.

Freeman, L. D. (1977). State interest, certification, and teacher education program approval. In W. A. Hazard, et al. (Eds.), Legal issues in teacher preparation and certification (pp 67-108). Washington D. C.: ERIC Clearinghouse on Teacher Education.

Freeman, L. D. & Hess, R. (1985). Testing teachers and the law. Paper given at the annual meeting of the American Educational Research Association, Chicago, IL.

Gabrys, R. (1986). The use of state-mandated competencies in building the teacher certification testing program. In W.P. Gorth & M.L. Chernoff (Eds.), Testing For Teacher Certification (pp. 75-88). Hillsdale, N.J.: Lawrence Erlbaum.

Galambos, E. C. (1984). Testing teachers for certification and recertification. Paper presented at a Hearing of the National Commission on Excellence in Teacher Education, Atlanta, GA.

Gallup, G.H. (1984). The 16th Annual Gallup Poll of the public's attitudes toward the public schools. Phi Delta Kappan, 66, 1, 23-38.

Gallup, G.H. (1986). The 18th Annual Gallup Poll of the public's attitudes toward the public schools. Phi Delta Kappan, 68, 1, 43-59.

Garvue, R. et al. (1983). The 1983 National Teacher Examinations Core Battery Louisiana validation study: Final Report. Baton Rouge, LA: Office of Research and Development, Department of Education, State of Louisiana.

Georgia Department of Education (1985). A Standard of Quality: The Georgia Teachers Certification Testing Program. Atlanta: Author.

Gifford, B.R. (1985). Teacher competency testing and its effects on minorities: Reflection and recommendations. In Freeman, E.E. (Ed.), Educational Standards, Testing, and Access. Proceedings of the 1984 ETS Invitational Conference. Princton: N.J.: Educational Testing Service.

Gifford, B.R. (1986). Excellence and equity in teacher competency testing: policy perspective. The Journal of Negro Education, 55, 3, 251-271.

Gifford, B.R. & Stoddard, T. (1985). Teacher education: Rhetoric or real reform. In W. J. Johnston (Ed.), Education on Trial: Strategies for the future (pp. 177-197). San Francisco: Institute for Contemporary Studies Press.

Glass, G.V. (1978). Standards and criteria. Journal of Educational Measurement, 15, 237-261.

Goertz, M. (1986). State educational standards: A 50-state survey. Princeton, N.J.: Educational Testing Service.

Goertz, M. E., Ekstrom, R. B., & Coley, R. J. (1984). The impact of state policy on entrance into the teaching profession. Final report, NIE grant no. G.83-0073. Princeton, N. J.: Division of education policy research and services, Educational Testing Service.

Gronlund, N. E. (1985). Measurement and evaluation in teaching (5th ed.). New York: Macmillan.

Gubser, L. (1979). A triangular model for assessing and assuring professional competence. Paper presented at the annual Texas Conference on Teacher Education, Austin, Tx.

Guion, R. M. (1977). Content validity, the source of my discontent. Applied Psychological Measurement, 1, 1-10.

Hackley, L.V. (1985). The decline in the number of black teachers can be reversed. Educational Measurement: Issues and Practice, 4, 3, 17-19.

Hall, C. L. (1984). Validating the NTE for the initial certification of teachers and administrators in New Mexico--and beyond. Presented as part of a symposium on Validation of the National Teacher Examinations: A Multi-state perspective, at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hambleton, R. K. (1984). Validating the test scores. In R. A. Berk, (Ed.). A guide to criterion-referenced test construction (pp. 199-230). Baltimore: The Johns Hopkins University Press.

Hankins, B. J. & Hancock, J. J. (1984). Validation study of the NTE for certification of entry level teachers in the State of Mississippi. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA.

Hathaway, W. E. (1980). Testing teachers. Educational Leadership, 38, 3, 210-215.

Hecht, K. A. (1979). Current status and methodological problems of validating professional licensing and certification. In M. A. Bunda & J. R. Sanders (Eds.), Practices & Problems in Competency-Based Measurement (pp. 16 - 27). Washington, D. C.: National Council on Measurement in Education.

Herbsleb, J.D., Sales, B.D., & Overcast, T.D. (1985). Challenging licensure and certification. American Psychologist, 40, 11, 1165-1178.

Hills, J. R. (1981). Measurement and evaluation in the classroom (2nd ed.). Columbus, OH: Charles E. Merrill Publishing Company.

Hilldrup, R. P. (1978, April). What are you doing about your illiterate teachers? The American School Board Journal, pp. 27-28.

Holmes, B.J. (1986). Do not buy the conventional wisdom: Minority teachers can pass the tests. The Journal of Negro Education, 55, 3, 335-346.

Holmes Group, The (1986). Tomorrow's Teachers: A Report on the Holmes Group. The Holmes Group. East Lansing, Michigan.

Hopkins, K. D. & Stanley, J. C. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, N. J.: Prentice Hall.

Hunter, J. E. (1983). A causal analysis of cognitive ability, job knowledge, job performance, and supervisor ratings. In F. Landy, S. Zedeck, & J. Cleveland (Eds.). Performance Measurement and Theory (pp. 257-266). Hillsdale, New Jersey: Lawrence Erlbaum Associates.

IOX Assessment Associates. (1983). Appraising the National Teacher Examinations for the State of Kentucky. Culver City, CA: IOX Assessment Associates.

Jaeger, R.M. (1986). Policy issues in standard setting for professional licensing tests. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for Teacher Certification (pp. 185-199). Hillsdale, N.J.: Lawrence Erlbaum.

Johnson, S.T. & Prom-Jackson, S. (1986). The memorable teacher: Implications for teacher selection. The Journal of Negro Education, 55, 3, 272-283.

Kane, M.T. (1982). A sampling model for validity. Applied Psychological Measurement, 6, 2, 125-160.

Kane, M. T. (1984). Strategies in validating licensure examinations. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Lahey, M.A. (1985). Personal communication.

Lehmann, I. J. & Phillips, S. E. (1985). Teacher competency examination programs: A national survey. Paper presented at the annual meeting of the National Council on Measurement in Education Chicago, IL.

Levine, E. L., et al. (1981). Evaluation of seven job analysis methods by experienced job analysts. Unpublished paper, University of South Florida.

Linn, R.L. (1984). Standards for validity in licensure testing. Paper presented at the "Validity in Licensure Testing" symposium at the annual meeting of the American Educational Research Association, New Orleans, LA.

Linn, R.L. & Miller, M.D. (1986). Review of test validation procedures and results. In R.M. Jaeger and J.C. Busch (Principal Investigators), An evaluation of the Georgia Teacher Certification Testing Program. (Chapter 3). Greensboro, N.Carolina: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.

Livington, S. A. & Zieky, M. J. (1982). Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests. Princeton, N. J.: Educational Testing Service.

Madaus, G. (1985). Public policy and the testing profession - You've never had it s good? Educational Measurement: Issues and Practice, 4, 4, 5-11.

McMorris, R. F. (1972). Effects of violating item construction principles. Journal of Educational Measurement, 9, 4, 287-295.

McPhee, S. A. & Kerr, M. E. (1985). Scholastic aptitude and achievement as predictors of performance on competency tests. Journal of Educational Research, 78, 3, 186-190.

Medley, D. M. (1982). Teacher competency testing and the teacher educator. Charlottesville: Association of Teacher Educators and the Bureau of Educationl Research, University of Virginia.

Mehrens, W.A. (1986). Measurement specialists: Motive to achieve or motive to avoid failure? Educational Measurement: Issues and Practice, 5, 4, 5-10.

Mehrens, W. A. & Lehmann, I. J. (1984). Measurement and Evaluation in Education and Psychlogy (3rd ed.). New York: Holt, Rinehart & Winston, Inc.

Mehrens, W. A. & Lehmann, I. J. (1987). Using Standardized Tests in Education (4th ed.). New York: Longman.

Messick, S. A. (1975). The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 30, 955-966.

Millman, J. (1986). Review of test development and score reporting procedures. In R. M. Jaeger and J.C. Busch (Principal Investigators). An evaluation of the Georgia Teacher Certification Testing Program (Chapter 3). Greensboro, N. Carolina: Center for Educational Research and Evaluation, University of North Carolina at Greensboro.

Nassif, P. M. (1978). Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, Canada.

National Center for Educational Statistics. (1982). Projections of education statistics to 1990-1991. Washington, D.C.: National Center for Education Statistics.

National Commission for Health Certifying Agencies. (1980, April). Perspectives on health occupational credentialing. Washington, D. C.: U.S. Department of Health and Human Services, DHHS Publication No. (HRA) 80-39.

Newsnotes. (1984). School programs are changing in response to reform reports: ERS. Phi Delta Kappan, 66, 4, 301.

Nitko, A. J. (1983). Educational tests and measurement. New York: Harcourt, Brace & Jovanovich.

Pecheone, R.L. Tomala, G., & Forgione, P.D., Jr. (1986). Building a competency test for prospective teachers. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for Teacher Certification (pp. 99-113). Hillsdale, N.J.: Lawrence Erlbaum.

Piper, M.K. & O'Sullivan, P.S. (1981). The National Teacher Examination: Can it predict classroom performance? Phi Delta Kappan, 62, 5, 401.

Popham, W. J. (1984). Specifying the domain of content or behaviors. In R. A. Berk (Ed.), A guide to criterion-referenced test test construction (pp. 29-48). Baltimore: The Johns Hopkins University Press.

Pottinger, P. S. (1979). Competence testing as a basis for licensing: Problems and prospects. In M. A. Bunda & J. R. Sanders (Eds.). Practices and problems in competency-based education (pp. 28-46). Washington, D. C.: National Council on Measurement in Education.

Pugach, M. C. & Raths, J. D. (1983). Testing teachers: Analysis and recommendations. Journal of Teacher Education. 34, 1, 37-43.

Pyburn, K. M. Jr. (1984). Legal challenges to licensing examinations. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA.

Quirk, T. J., Witten, B. J. & Weinberg, S. F. (1973). Review of studies of the concurrent and predictive validity of the national teacher examinations. Review of Educational Research, 43, 1, 89-114.

Raspberry, W. (1983, April 29). Teachers should pass tests, too. Washington Post.

Rebell, M.A. (1986a). Disparate impact of the teacher competency testing on minorities: Don't blame the test-takers--or the tests. Yale Law & Policy Review, 4, 2, 375-403.

Rebell, M.A. (1986b). Recent legal issues in competency testing for teachers. In W.F. Gorth & M.L. Cheroff (Eds.), Testing for teacher certification (pp. 59-73). Hillsdale, N.J.: Lawrence Erlbaum.

Rebell, M.A. & Koenigsberg, R.G. (1986). Post-trial memorandum of law on behalf of amicus curiae National Evaluation System, Inc. Margaret T. Allen, et al. and Board of Trustees for Alabama State University and Erica P. Smith v. Alambama State Board of Education et al.

Roid, G. H. (1984). Generating the test items. In R.A. Berk (Ed.), A guide to criterion-references test construction (pp. 49-77). Baltimore: The Johns Hopkins University Press.

Rosen, G.A. (1986, August). A perspective on predictive validity and licensure examination. Paper presented at the 94th annual convention of the American Psychological Association, Washington, D.C.

Rosenfeld, M., Shimberg, B., & Thornton, R.F. (1983). Job Analysis of Licensed Psychologists in the United States and Canada. Princeton, N.J.: Center for Occupational and Professional Assessment, Educational Testing Service.

Rosenfeld, M., Thornton, R.F. & Skurnik, L.S. (March, 1986). Analysis of the professional function of teachers: Relationships between job functions and the NTE Core Battery. Research Report 86-8. Princton, N.J.: Educational Testing Service.

Roth, R. (1984). Validation study of the National Teacher Examinations for certification in the State of Arkansas. Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, New Orleans, LA.

Rovinelli, R. J. & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. Dutch Journal of Educational Research. 2, 49-60.

Rubinstein, S. A., McDonough, M. W. & Allan, R. G. (1982). The changing nature of teacher certification programs. A paper presented at the annual meeting of the National Council on Measurement in Education, New York.

Ruch, W. (1984). Personal communication.

Rulon, P. J. (1946). On the validity of educational tests. Harvard Educational Review. 16, 4, 290-296.

Sax, G. (1980). Principles of educational and psychological mesurement and evaluation (2nd ed.). Belmont, CA: Wadsworth Publishing Company.

Schmidt, F. L. & Hunter, J. E. (1981). Employment testing: Old theories and new research findings. American Psychologist, 36, 10, 1128-1137.

Scriven, M. (1979). Recommendations for modification in external assessment process. State of California: Commission on Teacher Preparation and Licensing.

Shanker, A. (1985). A national teacher examination. Educational Measurement: Issues and practice, 4, 3, 28-31.

Shanker, A. & Ward, J. G. (1982). Teacher competency and testing: A natural affinity. Educational Measurement: Issues and practice, 1, 2, 6-9 & 26.

Shannon, G.A. (1986, April). Usefulness of score interpretive information for examinees who fail criterion-referenced tests. Paper presented at the Annual Meeting of the American Education Research Association, San Francisco.

Shimberg, B. (1981). Testing for licensure and certification. American Psychologist, 36, 10. 1138-1146.

Shimberg, B. (1982) Occupational licensing: A public perspective. Princeton, N. J.: Educational Testing Service.

Shimberg, B. (1984) The relationship among accreditation, certification and licensure. Federation Bulletin, 71, 4, 99-116.

Shimberg, B. (1985). Overview of professional and occupational licensing. In J. C. Fortune & Associates (Eds.), Understanding Testing in Occupational Licensing (pp. 1-14). San Francisco. Jossey Bass.

Shulman, L.S. (1986). Those who understand: Knowledge growth in teaching. Educational Researcher, 15, 2, 4-14.

Shulman, L.S. (No date). Knowledge and teaching: Foundations of the New Reform. Paper prepared for the Task Force on Teaching as a Profession. Carnegie Forum on Education and the Economy, New York.

Solomon, L.M. (1986). Analysis of candidates who retake the teacher tests. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for Teacher Certification (pp. 89-97). Hillsdale, N.J.: Larwence Erlbaum.

Southall, C. (1982) Trends in student teacher evaluation in selected mid-western states. Unpublished manuscript, College of Education, University of Missouri-Columbia.

Spencer, T.L. (1986). Teacher education at Grambling State University: A move toward excellence. The Journal of Negro Education, 55, 3, 293-303.

Stark, J. S. & Lowther, M.A. (1984) Predictors of teachers' preferences concerning their evaluation. Educational Administration Quarterly, 20, 4, 76-106.

State of Florida, (1982). Florida Teacher Certification Examination: Registration Bulletin. Tallahassee: Teacher Certification Section, Department of Education, Author.

Sykes, G. (1983). Contradiccions, ironies, and promises unfulfilled: A contemporary account of the status of teachings. Phi Delta Kappan, 62(2), 87-93.

Sykes, G. (1986). The social consequences of standard-setting in the professions. Paper prepared for the Task Force on Teaching as a Profession, Carnegie Forum on Education and the Economy, N.Y.

Tenopyr, M. L. (1977). Content-construct confusion. Personnel Psychology, 30, 47-54.

Thomas, M.C. & Tyler, K.P. (1984). Early identification and preparation for successful entry into teacher education programs. In M. Davis (Ed.), Prospective Black Teachers and the Closing Door: Strategies for Entry (pp. 8-14). Montgomery, AL: Alabama Center for Higher Education.

Time Magazine. (1980, June 16). Help! Teachers can't teach. Time Magazine.

Title, C. K. (1982). Use of judgmental methods in item bias studies. In R. A. Berk (Ed.), Handbook of methods for detecting test bias (pp. 31-63). Baltimore: The Johns Hopkins University Press.

The Washington Post. (1979, September 28). Civille Right. The Washington Post.

U.S. Civil Service Commission. (1973). Job Analysis: Key to Better Management. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office.

U S. Department of Health, Education and Welfare. (1971) Report on licensure and related health personnel credentialing. DHEW publication 72-11. Washington D. C.: Author.

U.S. News and World Report. (1984). Why many teachers don't measure up. U.S. News and World Report. September 10, p. 14.

USES. (1983). Overview of validity generalization for the U.S. Employment Service. USES Test Research Report No. 43. Division of counseling and test development employment and training administration. U.S. Department of Labor, Washington, D.C.

Vertiz, V. C. (1985). Legal issues in licensing. In J. C. Fortune & Associates (Eds.), Understanding Testing in Occupational Licensing (pp. 87-106). San Francisco: Jossey Bass.

Vold, D.J. (1985). The roots of teacher testing in America. Educational Measurement: Issues and Practice, 4, 3, 5-6.

Vorwerk, K.E. & Gorth, W.P. (1986). Common themes in teacher certification testing program development and implementation. In W.P. Gorth & M.L Chernoff (Eds.), Testing for Teacher Certification (pp. 35-43). Hillsdale, N.J.: Lawrence Erlbaum.

Watts, G.D. (1986). And let the air out of the volleyballs. Phi Delta Kappan, 67, 10, 723-724.

Weaver, J.R. (1986). Study guides and their effect on programs. In W.P. Gorth & M.L. Chernoff (Eds.), Testing for Teacher Certification (pp. 235-251). Hillsdale, N.J.: Lawrence Erlbaum.

Weaver, W. T. (1980). In search of quality: The need for new talent in teaching. In C.C. Mackey (Ed.). Assuring qualified educational personnel in the eighties (pp. 1-20). National Association of State Directors of Teacher Education and Certification, Albany, New York.

Webb, L. D. (1983) Teacher evaluation. In S. B. Thomas et al. (Eds). Educators and the Law (pp. 69-80). Elmont, NY: Institute for School Law and Finance.

Webster, W. J. (1984). Five years of teacher testing: A retrospective analysis. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Werner, E. (1982) What a licensing board member needs to know about testing. Paper given at the annual conference on The Clearinghouse on Licensure, Enforcement and Regulation, The Council of State Governments, Chicago, IL.

Williamston, J. W. (1979). Improving content validity of certification procedures by defining competence in specialty practice: Directions, resources, and getting started. In Definitions of competence in specialties of medicine, conference proceedings, (pp. 61-68). Chicago: American Board of Medical Specialties.

Wood, B. D. (1940). Making use of the objective examination as a phase of teacher selection. Harvard Educational Review, 10, 277-282.

Yalow, E. S. & Collins, J. L. (1985). Meeting the challenge of content validity. Presented as part of a symposium session "The assessment boomerang returns: Competency tests for Educatory" at the annual meeting of the American Educational Research Association, Chicago, IL.

Yalow, E. S. & Popham, W. J. (1983). Content validity at the crossroads. Educational Researcher, 12, 8, 10-14,21.