1990

4. Assessing the Quality of Teacher Assessment Tests

William A. Mehrens
Michigan State University

4

Assessing the Quality of Teacher Assessment Tests[1]

William A. Mehrens
Michigan State University

This chapter discusses some of the types of evidences that are appropriate for assessing the quality of teacher-licensure tests. Licensure tests are used to make dichotomous decisions, so reliability estimates of the consistency of decisions are needed. 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 for criterion-related validity, construct validity and "curricular validity" are also discussed. The issue of whether the cut score on a licensure test should in any way be related to the supply and demand and the requirements for reporting test scores and documenting the quality of the test are also discussed.

It is concluded that teacher-licensure tests allow valid in-

[1]Portions of this chapter have been adapted from Validity Issues in Teacher Competency Tests, *Journal of Personnel Evaluation in Education*, 1987, *1*, 2, 195–229 and from *Issues in Teacher Competency Tests* which was prepared for the Commission on Testing and Public Policy, Graduate School of Education, University of California, Berkeley.

ferences 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 help ensure that those individuals who are licensed have a minimal level of competence on some important subdomains of knowledge and skills relevant to their profession

ASSESSING THE QUALITY OF TEACHER-ASSESSMENT TESTS

> Scott wont pass in his assignment at all, he had a poem to learn and he fell tu do it. (*Time*, 1980)
>
> If selection of the most suitable people to be teachers is a matter of importance to the five percent of the population who become teachers, it is no less important a matter to the 100 percent who become students. (Pratt, 1977, p. 16)
>
> If education is the cornerstone upon which a great nation builds, then teaching is our most important human activity. (Sweeney & Manatt, 1986, p. 446)

It seems so obvious. Quality education is important to the nation. Quality teachers are important for quality education. But historically not all who received licenses to teach were of high quality—or necessarily even competent. We do not want incompetent teachers. Licensure tests are used in over 900 other occupations in an effort to protect the public, and in those occupations the public typically has a choice of whom to go to for services. Teachers have conscripted clients. Licensure tests should be able to weed out prospective teachers with skills at a level such as that demonstrated in the first of the preceding quotations. Isn't it obvious licensure tests should be useful in a profession as important as teaching?

But things are not always as obvious 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 chapter is intended to take a close look at several issues regarding the quality of teacher-competency tests. A general conclusion of the paper is that if such tests are constructed properly they will be of sufficiently high quality to be valid for a delimited set of inferences.

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.

Gabrys (1987) reported that "as of June 1, 1986, 46 states had regulations requiring some form of competency assessment of teachers. Three additional states were actively planning programs to test teachers . . ." (p. 27).

The very rapid spread of teacher-testing programs is politically based and supported by the public. Gallup polls (1984)indicate that 89% of the public (and 63% 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" (p. 107). Many educational leaders also support teacher testing. The recent Holmes Group (1986) and Carnegie (1986) reports on teaching both support examinations for prospective teachers. Both the American Federation of Teachers AFL–CIO and the National Education Association (NEA) currently support examinations for licensing new teachers (Cameron, 1985; Shanker, 1985).

WHY TEACHER-COMPETENCY TESTS?

The motivating factor behind teacher-competency tests is that the public and many educators believe that both our colleges and our state-licensing boards have failed as gatekeepers. Although debateable, there is considerable evidence for those beliefs (see Mehrens, 1986a; 1987a; 1987b).

A few educators may discount, or perhaps even support, the deplorable standards (arguing that love, patience, compassion, and so forth, are the important criteria to be a teacher (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 earlier in this chapter. 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 (see Carnegie Task Force, 1986; Holmes Group, 1986; Shulman, n.d.; 1986).

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 (see Mehrens, 1986a; 1987a; 1987b). Graduation from one of the 1200 institutions with teacher-education programs simply does not ensure sufficient competence. This is partly due to political considerations (see Scriven, 1979), but 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 (1977) 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." (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

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) made the following useful distinctions between 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.

This chapter is limited to a discussion of issues in assessing the quality 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. Furthermore, 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 dis-

tinctions among the three terms. Thus, some definitions and explanations are in order.

The U.S. Department of Health, Education, and Welfare (1971) 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. (p. 7)

The same agency defined 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. (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.

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 certified. It is a selecting-in process. Certification typically goes beyond the minimum requirements. (The type of examinations Shulman [n.d.] and Shanker [1985] advocated would not appear to be minimal.)

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, which are discussed in this chapter, are, in fact, state-licensure programs. Their purpose is to protect the public from incompetents.

Employment tests have a quite different purpose from licensure tests. Employment tests 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 *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, NCME], 1985) clarified 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" (p. 63).

Because employment and licensure examinations serve different purposes, they may well be constructed somewhat differently. Whether or not the examinations differ, we make different inferences from the scores of examinations used for employment and licensure and, therefore, the kinds of evidence gathered to support their uses should differ. Because many people confuse the validity requirements of the two types of examinations, they will be discussed in more detail at various points in the chapter, which is primarily devoted to technical measurement issues of relevance for licensure tests.

RELIABILITY

Licensure tests are used to make dichotomous decisions. As the *Standards*[2] pointed out: "Estimates of the consistency of decisions are needed whenever decision rules assign people to categories according to specified test score intervals. An estimate of the standard error of measurement at the cut score is helpful" (AERA, APA, NCME, 1985, p. 20).

Standard 2.12 is quite specific about what the authors believe is conditionally desired. "For dichotomous decisions, estimates should be provided of the percentage of test takers who are classified in the same way on two occasions or on alternate forms of the test. (conditional)" (p. 23).

Other literature (e.g., Millman, 1979) suggests that the Kappa index, which corrects the proportion of agreement for chance effects, should also be computed. Because two different scores per person are not typically obtained in licensure tests it is necessary to estimate the proportion of consistent decisions from the distribution of scores for a single administration. The *Standards* do

[2]*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 chapter. The defense, in addition to the sound, is that there are a set of standards within the single book.

not specify any particular formula to estimate this. The literature suggests that "there is no procedure for estimating this quantity that is clearly preferred over all others" (Traub, 1986, pp. 5–6). (See also Subkoviak, 1984.) Certainly, the Subkoviak, Huynh, and Marshall procedures would all be considered acceptable (Subkoviak, 1984).

Standard 2.12 just quoted, by its calling for estimates of the consistency of *decisions*, and Standard 11.3 both show a clear preference for reliability indices that are based on a threshold-loss function. Novick, chair of the *Standards* committee also previously had argued for this approach (Hambleton & Novick, 1973). The threshold-loss function assumes that all misclassifications are equally serious regardless of their size. If misclassifying an individual close to the cut score is a less serious error than misclassifying one far above or below the cut score, then one should use a procedure that involves a squared error-loss function. Examples would be Livingston's (1972) approach or the Brennan and Kane index (1977).

Traub (1986) has recently argued for the threshold loss function for licensure tests because "an error of classification has consequences that are as serious *for the candidate* [italics added] whose true score lies relatively near the cutting score as for the candidate whose true test score lies relatively far from the cutting score" (pp. 5–14). However, not all specialists would wish to use the threshold loss function. Recall that the purpose of licensure tests is to protect the public. Most measurement specialists believe that knowledge measured in a licensure test is a continuous variable and that the cut score artificially divides the variable into two categories. One could argue that a false positive teacher candidate with a true score far below the cut score would be more costly in terms of harm to the *public* than a false positive whose true score was just one point below the cut score. One could make a comparable argument for false negatives. For a more thorough discussion of this issue see Berk (1984a), Subkoviak (1984), and Brennan (1984).

A third approach would be to estimate the reliability of the domain score estimates—consistency across parallel or randomly parallel test forms. The traditional K-R 20 is commonly used if one assumes parallel tests. As Traub (1986) pointed out, although such an estimate is not *required* by the *Standards* for licensure tests, it does provide useful information. It does not replace one of the other estimates discussed earlier.

As indicated in the first quotation from the *Standards* in this section, an estimate of the standard error of measurement at the cut score is helpful (see also Standard 2.10). Again, a variety of

formulas could be used. They make slightly different assumptions, and there is no consensus as to what method is best (Feldt, Steffen, & Gupta, 1985).

Of course if one knew the cut score in advance of test construction and had item statistics on a large number of items one could construct a test with a small standard error at the cut score. However this typically would not work in the initial construction of licensure tests because the cut score is based on item judgments—not determined in advance. If one assumed a constant cut score and had dependable item statistics, one might build subsequent test revisions on such a basis. However such tests may not be truly equivalent to the first.

Standards 2.1 and 11.4 (AERA, APA, NCME, 1985) speak of estimating reliability of the subscores that are reported and used. Because subscores are not typically used in teacher licensure decisions they would not need to be reported. If they are reported they might be used as study guides by candidates who failed and thus it would be useful to report their reliabilities and standard errors. The reliabilities are frequently low and candidates should recognize their limitations as study guides. However, it should be stressed that low subscore reliabilities are irrelevant in litigation regarding the legality of using the total score for licensure decisions.

Some individuals like guidelines as to how reliable a test should be. Traub (1986) chose 0.80 as an arbitrary value for an acceptably high decision consistency index. However, he did not suggest discontinuing tests with lower estimates. Rather he suggested they "give cause for concern" (p. 5–23).

VALIDITY: SOME GENERAL NOTIONS

> 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)

The AERA, APA, NCME (1985) *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. (p. 9)

Although validity is a unitary concept, evidence may be accumulated in many ways. Traditionally, psychometricians have cate-

gorized 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" (AERA, APA, NCME, 1985, 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. Criterion-related evidence demonstrates that test scores are systematically related to one or more outcome criteria" (p. 10–11). 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* [italics added] constructions about the nature of human behavior" (p. 9).

The lack of a rigorous distinction among the categories of validity evidence is especially true between the categories of content and construct validity evidence. A distinction Tenopyr (1977) 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* [italics added] only in test construction, by specifying a domain of tasks and sampling rigorously. The *inference back to the domain* [italics added] can then be purely deductive" (Cronbach, 1980, p. 105). This wording holds more appeal to me. We do make deductive *inferences* from the score on the test 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.

It is unfortunate, but not incapacitating, that measurement specialists do not all use the validity terms the same way. In this chapter 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, that reader is surely capable of handling the internal translation he or she must engage in to comprehend the discussion. In fact, some who argue that licensure tests need construct validity evidence might well feel appeased if some of the evidence here placed under content were recategorized to construct.

The terms *curricular validity* and *instructional validity* are being used increasingly in the educational measurement literature. Al-

though 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 (Mehrens & Lehmann, 1987; Yalow & Popham, 1983). However, both are generally considered irrelevant in judging the quality of licensure examinations. Reasons for this are discussed later.

INFERENCES FROM TEACHER-COMPETENCY TESTS

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 is 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 responsibility 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 cannot 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" (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 (Rulon, 1946). 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 (1940) made this point over 45 years ago: "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 . . ." (278–279).

Of course, if test builders/users do not wish others to make incorrect inferences 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. Almost all those who write in the professional literature regarding licensure examinations, would agree with Kane (1984) that such exams should "be interpreted as providing evidence of an examinee's present competence on specific abilities that are needed for practice" (p. 2).

CONTENT-VALIDITY EVIDENCE FOR TEACHER-COMPETENCY TESTS

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. (e.g., Bond, 1987, p. 19; Linn & Miller, 1986, pp. 4–3; Shimberg, 1982, p. 62; Vertiz, 1985, p. 97.)

However, the content-validity evidence should differ for licensure and employment purposes. 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, employment tests may measure *aptitude* to learn a *specific* job, whereas licensure is usually intended 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 although 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 quotation. Linn (1984) and Kane (1984) have made the same point. 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).

Note that the quotation 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. The problem with the "necessary" requirement is that very, very few specific compe-

tencies 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).

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. 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 questions was *absolutely* necessary.

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 al-

most 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 subdomain of pedagogy would be considered relevant. We could all probably agree at the abstract level that a teacher could know so little about that subdomain that he or she should not be licensed to teach. That indeed, giving a license to teach to someone who knew almost nothing 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 relevant objectives within that subdomain might help contribute to a correct inference about whether prospective teachers know the minimum amount necessary about the subdomain 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 absolute necessity of each specific piece of knowledge tapped by each question, should not be something about which the measurement community would disagree.

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.

The major points of concern in establishing the content validity of a licensure test appear to be (a) developing an original list of competencies, (b) doing some type of job analysis survey, (c) spec-

ifying the domain for the test, (d) writing and validating the items, and (e) obtaining an overall judgment of the content validity of the test. These five steps will be discussed in some detail in the following sections plus the additional sixth step of communicating the domain to the test takers and the general public.

Developing an Original List of Competencies

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 constituencies. It is probably useful to have a variety of perspectives represented on the committee (Yalow & Collins, 1985).

The starting point for the committee should be an understanding of the purpose(s) of licensure tests. The next task should be a thorough review of the relevant literature (Burns, 1985; Kane, 1984). This 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 (see Rosenfeld, Thornton, & Skurnik, 1986).

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 cannot 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 *not* being suggested that teachers only need subject-matter knowledge at the level they are teaching (see Shulman, 1986).

Doing the Job Analysis Survey

Professional standards, logic, and legal precedent all stress the importance of job relevance or job relatedness in both employment and licensure exams. There is no specific Standard on how to do the job analysis. The *Uniform Guidelines* state that "Any method of job analysis may be used if it provides information for the specific validation strategy used" (Equal Employment Opportunity Commission, 1978, p. 38300).

The *Guidelines* do impose two basic requirements for a job analysis to be used in content validation: (a) The analysis must yield an operational definition of the domain and (b) the content of the domain should be necessary for critical or important work behaviors. Two commonly accepted methods of determining job-relatedness are through document review and group discussion. These two methods (sometimes called logical analyses) 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* (American Psychological Association [APA], 1980) this process of using the pooled judgment of experts is a recognized approach to determining job relatedness (or job analysis).

A major advantage of a logical analysis is that "it makes use of the extensive body of existing knowledge and can focus on the main goals of the job or category of jobs" (Kane, Kingsbury, Colton, & Estes, 1986, p. 1.6). The main disadvantage is that it may overlook important aspects of work.

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. Although a few educational measurement experts would wish the job analyses to include observations, they appear to be in the minority (see Kane et al., 1986, p. 1.7).

Typically a job-analysis survey (or task inventory) of people in the profession is conducted to confirm, disconfirm, or add to the judgments of the committee experts (Pecheone, Tomala, & Forgione, 1986; Yalow & Collins, 1985). (This survey is often referred to as an *empirical* analysis and some have confused this with crite-

rion related validity.) 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 or objectives 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. 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).

Some researchers (Colton, Kane, Kingsbury, & Estes, 1987; Elliot, 1987; Kane et al., 1986) discuss how to examine the construct validity of the job analysis. This is accomplished through setting hypotheses about the dimensional structure of the data; anticipated differential responses (or lack thereof) across different groups of respondents; agreement of responses with the professional literature; and so forth. This testing of the hypotheses about the job analysis data can also be used as indirect construct validity evidence for the test data. It adds credence to the supposition that the test measures teacher competence.

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. Elliot (1987) found that there were no differences in amount of *time spent* on various job content areas between individuals with and without Masters' Degrees and among individuals of varying years of teaching experience. However, he did find differences across grade levels taught.

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.

Obviously, most surveys done to help 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 survey 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 are discussed later. These procedures are not reasonably considered a part of the job analysis.

Determining the Domain Specifications

As Elliott and Nelson (1984) pointed out: "There is little to guide the developer of teacher licensing tests in making the huge leap from job analysis to domain specifications" (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 algorithms one might set up to weight the subcategories of the domain or to sample within the subcategories. Determining the domain specifications is obviously a judgmental task, and as Cronbach (1980) suggested, "the defense must be prepared to show that the domain is relevant and that weight is properly distributed over it" (p. 105).

Three general points need to be made about the domain of licensure tests: (a) the domain should be fairly broad because a certificate is not for a specific job but for a general kind of job; (b) the domain does not have to cover the total set of tasks determined by the job analysis, and related to that; (c) a licensure test does not need to have, and probably should not have, subcategory 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* knowledge and skills. The weighting of an area should be based on the degree of its 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 spent in teaching or using an objective, as well as the criticality of the objective. Often, these data are combined in some

fashion to determine a single value of "importance" for each objective. There is no single established algorithm for combining the two pieces of information or for arriving at weights for the table of specifications. Some evidence suggests the algorithm used to combine the two variables does not matter a great deal because the correlations between the responses to the two questions is quite high. 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 0.93, between the first and third it was 0.91, and between the second and third it was 0.996 (M. A. Lahey, personal communication, 1985). Schmeiser (1987) also found that three different methods of obtaining composite scores produced a very high degree of consistency. However, Kane et al. (1986) in a survey of nursing practices did find some differences across the algorithms they used. More research should probably be conducted in this area.

Once one has information about the objectives (or tasks), it is both appropriate, and common practice, to use it along with any subdomain information to select a proportional number of important objectives within each subdomain. 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 algorithm based on the ratings (see Millman, 1986).

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, & Hast-

ings, 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. However, some critics have suggested this format is inappropriate. Pottinger (1979) argued against such a format because he believed it does not do an adequate job of protecting the public. That is, multiple choice tests 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 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 cannot 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. Some professionals prefer item writers to work from what are commonly called "item specifications" (Popham, 1984). Others prefer to write items directly from objectives. There is no particular reason to prefer either approach although Millman (1986, pp. 3–8) suggested, 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; Roid, 1984, for positions advocating item specifications.)

Whether or not items are written from item specifications, it is necessary to have the items reviewed by a panel of experts. 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) gave an excellent overview of some of the methods of judging item validity. He 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. However, as Hambleton (1984) pointed out, this procedure is very time consuming to implement. 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. Although 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 essence, 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 nonminorities (W. Ruch, 1984, personal communication). Berk (1984b, 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.

Whatever particular methods are used the *Standards* state conditionally that "the relevant training, experience, and qualifications of the experts should be described" (AERA, APA, NCME, 1985, p. 15).

Overall Judgment of Content Validity

Because content validity is established only in test construction the judgment of the adequacy of the content validity should be based on a judgment of the adequacy of the construction process. 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 those tests 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.

Communicating the Domain to the Public

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 is sampled for the test.

It is appropriate to tell those who will be taking the test that all objectives will not be tested. This is what is done in many of the licensure tests. For example in the Examination for the Certification of Educators in Texas (ExCET) the *Study Guide* (National Evaluation Systems, n.d.) specifically states that the test measures only a portion of the objectives.

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. Apparently the Texas Examination of Current Administrators and Teachers (TECAT) covers *all* the basic reading and writing skills that educators need to perform adequately. Thus the list of objectives made public was limited to the objectives actually tested. Shepard and Kreitzer (1987) found that monumental effort went into preparing for the TECAT. "As soon as test specifications were available, the Continuing Education Division of the University of Texas at Austin, in cooperation with the Texas Classroom Teachers Association, developed a review course and a 300 page self-study book" (p. 6). Furthermore, 12 videotapes were prepared and used extensively in preparation for the test. Now this is fine *if* the objectives were all inclusive of the basic skills in reading and writing that teachers should know and *if* the information and preparation helped the teachers develop the skills as opposed to just passing the test. But, as Shepard and Kreitzer (1987) reported, "at some point legitimate teaching to the test crossed over an ill-defined line and became inappropriate. . . . Over and over again, . . . teaching to the test involved exploitation of the test specifications . . ." (p. 9).

I agree with Shepard and Kreitzer that the line between legitimate teaching to the test and illegitimate teaching of the test is not well defined. But there is a legitimate worry that if too much information is released in advance about a test—such as which specific objectives are tested and detailed item specifications regarding how the questions and multiple-choice options are developed—one will no longer be able to infer competence in the domain from a passing score on a test. In the extreme one could give out advance copies of the test; most of us would feel that to be

inadvisable for licensure tests where the goal is to protect the public from incompetents.

Not all would agree with my position. An expert witness at one trial testified that he found it misleading to communicate a larger set of objectives to candidates than are actually 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 we wish to make 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 at both the objective and item levels.

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 (1985) suggested, 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" (p. 5). Johnson and Prom-Jackson (1986) pointed out that the cognitive abilities of teachers "constitute a necessary but not nearly sufficient condition" [for teacher success] (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 (1979), 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" (p. 21). However her opinion is certainly not the common view held by most measurement experts. Shimberg (1982) 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.

. . .

> 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. (p. 60)

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" (Kane, 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. Kane (1982; 1984), Linn (1984), Rosen (1986), Shimberg (1982; 1984), and others all have argued that it is both unfeasible and inappropriate to expect criterion-related validity of a licensure examination.

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 gave as examples 10 different criteria for teacher evaluation.

Ebel (1961) and Kane (1984) both discussed the many criterion problems for licensure exams. As Cronbach (1970) 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" (p. 127). Petersen (1987) reported that administrative reports were bunched up at the high end of the

scale, showed little variance, and did not correlate with other measures used to evaluate teachers. Berliner (1986) discussed the lack of training of judges of teacher competence: "Only a few states provided any training for their judges, and when training was provided it was usually for one-half day" (p. 11). He compares that to the training of other judges "We learned that to become a livestock judge in Arizona you ordinarily have to take a year of livestock evaluation courses at a college. . . . The American Kennel Club's application for a judge requires 10 years documented experience in the field. . . . Written testing and an oral interview are also required. . . . In figure skating it can take 10 to 15 years" (pp. 12, 13).

It is probably safe to suggest that if teachers had high supervisors' ratings on teaching effectiveness but could not pass a test on the content they were supposed to be teaching, most reasonable people would (and should) 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 MD 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 MDs is not needed. Rather, they typically interpret the situation as an instance of an incompetent not getting caught sooner.)

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 (1977) 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. (p. 60)

In another article, Ebel (1975) 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* [italics added]. (pp. 26–27)

In summary, 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; and 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 preceding statements that it is inappropriate to attempt to find out what, if any, correlates of teacher-licensure tests exist. Although correlational data are somewhat sparse, they are consonant with the logical inference that knowledge about teaching and the subject matter being taught (competence) should be related to both performance and effectiveness in teaching.

Webster (1984) conducted 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. (p. 4)

Using a class average residualized composite score (CARCS), Webster found a correlation of 0.47 between CARCS and the NTE Common, 0.47 between CARCS and WPCT-Verbal, 0.37 between CARCS and WPCT-Math, and 0.48 between CARCS and WPCT-Total. All correlations were significant at $p \leq .01$.

Piper and 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 (0.43)

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 (Ayres, 1983; McPhee & Kerr, 1985).

CONSTRUCT-VALIDITY EVIDENCE

Although some measurement experts believe that all validity is construct validity, other measurement experts worry some about this labeling because "theoretical constructions about the nature of human behavior" (AERA, APA, NCME, 1985, p. 9) often implies hypothetical constructions (Ebel, 1974). Haertel (1987) suggested that in achievement testing either of two positions may be taken.

> A *domain* [italics added] of test items may be considered to operationalize achievement outcomes, so that achievement is defined in terms of test performance, or items may be treated as partial and imperfect indicators of student proficiency, so that achievement cannot in principle be defined in terms of any single operational procedure for its measurement. The former position is consonant with a behaviorist orientation, which treats mental entities as no more than interviewing variables, and the latter with a cognitivist orientation, which treats mental entities as *hypothetical constructs* [italics added]. (pp. 5–6)

Given Haertel's conception of cognitive learning outcomes, he believes they cannot be defined in behavioral terms. However, Ebel (1977), in speaking specifically about educational and employment testing, 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. . . .
>
> Why do we continue to talk about construct validity 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 (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 chapter would not necessarily lead to such a conclusion. According to Medley's (1982) 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. This does not seem like a theoretical construct. It would seem a set of items could *sample* this repertoire rather than being a *sign* of a theoretical construct (see Mehrens & Lehmann, 1987).

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 (and do) 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" which may allow some individuals to infer a construct. The writers of the *Standards*, apparently very alert to this issue, wisely used only 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 labeled "construct validity" evidence (in spite of the illogical but real legal dangers of so doing) how should it be done? 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 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 chapter is usually available for well-developed licensure tests. There are some criterion related validity 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).

If all the aforementioned 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. If one is going to suggest that a test measures a theoretical construct, it would appear necessary to define the construct. Kane (1984) 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" (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 in 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!

As has been pointed out by a variety of writers (e.g., Darling-Hammond, Wise, & Pease, 1983), the evaluation of teaching in any generic sense depends on one's conceptions. The Medley distinctions made early in the chapter between teacher competence, teacher performance, and teacher effectiveness must be kept in mind. Whereas 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 and/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/instructional 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: (a) confusing the situation in the Debra P. case with licensure decisions (see Rebell, 1986), and (b) 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.* . . ." (Educational Testing Service [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* state 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 academic 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. It would make little sense to build a licensure examination based on the curriculum of an inadequate college! Roth (1984) 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. (p. 4)

Roth went on to argue that the validity question for licensure examinations is job relevance, not training program relevance. This is the commonly—almost universally—accepted position.[3]

---

[3]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 chapter.

Most states using the NTE tests have "validated" them both for their match to the colleges' curricula and to the requirements of the job. 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.

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. 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. 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. The relationship of course holds only for tests over knowledge of the profession of education. It cannot 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).

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 ad-

ditional 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). Ekstrom and Goertz (1985) argued 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. (p. 9)

What the state must *not* 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 grievous 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. It would seem this confusion should have ended 25 years ago (see, Cronbach 1963, p. 680).

HOW VALID MUST A TEST BE? IDEALISTIC VERSUS 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. But 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 in a conjunctive model 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 (potentially) 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).

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. (p. 9)

Werner (1982) 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. (pp. 7–8)

Finally, consider some excerpts from an article on the *Examination for Professional Practice in Psychology* (Carlson, 1978), 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* [italics added]. (pp. 491–492)

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, pp. 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 chapter, 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.

FURTHER POSSIBLE RESEARCH 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 language, 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 emanates from tests that when it emanates 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 chapter in the development of content valid teacher-competence 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 or competencies? How should we "validate" the competencies the committees produce? Must we accept them on faith? If not, what external criteria would we use?

It is also reasonable to conclude that we would profit from more research in job-analyses procedures. 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 criticality? Would there be any better way to determine how critical 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-analysis strategies. Of course none of this research would empirically answer the question of which strategy produces the "best" domain of competencies.

Licensure 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 *not* having the basic competencies will be more likely to harm the public (students) than those who do have the minimum competen-

cies. 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 effectiveness*—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 intermittent reinforcement over continuous reinforcement. The problem is to define and measure the *harmful* effects, and to show that the lack of teacher competence 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 be difficult indeed. 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 precede 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 dust-

bowl 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 reveal an effect of a lack of "necessary" knowledge on 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 dichotomous decisions made from the scores. I am inclined to believe these studies would be useful. However, these studies should probably *not* 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 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 indicate 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 (which includes test data) to determine who should be licensed.

THE CUT SCORE

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 (e.g., Berk, 1986; Jaeger, 1986; Livingston & Zieky, 1982). There is considerable debate about what method is "best." In discussing the various drafts of the *Standards*, Linn (1984) 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 consensus on this issue within the area of certification and licensure testing to justify a specific standard on cut scores within this chapter" (p. 12).

Avoiding debates over specific cut score methodologies, there are still some cut scores 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 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 excerpts:

> 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 selec-

> tion. . . . There is not an explicit limit on 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 rates. (Shimberg, 1984, p. 3)

All of the preceding excerpts 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 (1979) addressed 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. (p. 41)

At a theoretical level, there is much in all the preceding excerpts 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. Furthermore, tests are never designed perfectly. Many licensure tests, in fact, have many of the same characteristics as employment tests. Schmidt and Hunter (1981) suggested 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 *continuum* 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. (1130)

In theory, employment tests should be positively correlated with the criterion true scores 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 important 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 the items that an individual gets correct, the less danger to the public. Surely, all else being equal, 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 continuum 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 critical 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 *should* 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 MD were available, that person would probably prefer going to a nonlicensed graduate of a medical college than going to someone with no medical training whatsoever. Particularly if the graduate was a false negative who failed the examination! If there were generally a shortage of doctors, it might make some sense to lower the qualification a bit. If there were generally a surplus, it might make sense to raise the standards.

Whatever the merits of the views just expressed, 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 and Koenigsberg (1986, p. 65) also supported 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.

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 purposing, 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 (1986) discussed 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. (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 (1986, p. 55) 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% |

Although there is disagreement about the supply/demand issue, it is obvious that the placement of the cut scores *has* 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 *minimally* qualified candidate should know in order to be licensed. They have then *reduced* 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 there is a 50% chance of licensing an examinee whose true score is three

standard errors *below* the judged standard, whereas there is less than a probability of 0.0014 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 three 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.

The public should rightly be concerned about the profession's apparent concern for false rejects and its lack of concern for false acceptances. 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 that we have, in several states, intentionally set the cut score three 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 been 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 (1986) may have been unfortunately correct when they suggested 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 (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

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, 1986b, 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 (1983a) *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" (p. 18). As Shannon (1986) suggested, 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) submitted 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 subtest 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 subscores are frequently reported.

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

> Although CRT subject 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) pointed out, the AERA, APA, NCME (1985) *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 should be given information on their performance in parts of the test for which separate scores or reports are produced *and used in the decision process* [italics added]. (AERA, APA, NCME, 1985, p. 65)

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). [italics added]. (AERA, APA NCME, 1985, p. 20)

Note the emphasis added to the preceding excerpt. 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 a low subscore reliabilitie 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 (1987), however, suggested that the second goal of teacher competency testing programs "is to provide diagnostic information about candidates' strengths and weaknesses to the candidates and to the teacher training institutions" (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 chapter. 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 support systems including study guides for applicants. See Downs and Silvestro (1987) and Weaver (1986) for brief discussions of such support systems and their effects.

EVIDENCE OF TEST QUALITY

In this chapter I have addressed primarily the issue of validity for licensure tests and have addressed to a lesser degree the issues of reliability and the setting of cut scores. Other issues related to test

quality include test administration and scoring procedures, bias considerations, and equating of different forms. These could all be discussed, but the most general conclusion I would draw is that the considerations of such issues for teacher-licensure tests would be the same as for any high-stakes criterion-referenced achievement test.

A final issue to be discussed is the evidence that should exist regarding the quality of the test scores. There should be evidence documenting all stages of test construction, administration, scoring, and reporting. Evidence must be gathered and maintained regarding how the issues of reliability, validity, setting cut scores, equating and determining lack of bias were addressed.

The *Standards* (AERA, APA, NCME, 1985) have a chapter "Test Publication: Technical Manual's and User's Guides" which contains 11 standards. The index to the *Standards* references other standards that pertain to publishers' materials. The background section of the chapter makes the following relevant points: "Publishers should provide enough information for a qualified user or a reviewer of a test to evaluate the appropriateness and technical adequacy of the test. Even when a test (or test battery) is developed for use within a single organization, a brief manual will be useful" (p. 35).

There has not been total agreement either about the degree of detail that should be in a manual or about the kinds of statements that a publisher should be able to document. In the Alabama lawsuit the documentation issue received consideration attention. Plaintiffs' experts argued for the necessity for very complete documentation, whereas Defendant's experts, though obviously not opposed to documentation, felt that a rule of reason should apply. As was stated in the *Defendant's Post-Trial Memorandum:* "Regardless of how much documentation one created or maintained, a reviewer could always find something that was not documented; again a "rule of reason" applies" (Boyd & Coody, 1986, p. 70).

One possible rule of reason is to use what commercial publishers do. Hall (1985) reviewed the technical data of 37 published achievement tests. He reported that only 54% provided information on the manner in which they selected their test items, only 46% reported item discrimination information, 49% item difficulty, 32% logical techniques for race bias, and 11% empirical techniques for race bias. For the criterion-referenced tests within the sample the percentages were even lower. For example while 89% of the norm-referenced tests provided reliability data, only 11% of the criterion-referenced tests reported such information.

However, what has been done is not the same as what should be done. It is obvious that many publishers are less diligent than they should be. Because critics will want to audit the test construction process for high stakes tests such as licensure tests, the publishers would be wise to be particularly diligent in the accuracy and thoroughness of their documentation. Manuals should be quite complete and records should exist to verify the information in the manuals. However, again we should apply a rule of reason. For example, in Texas approximately 200 educators in each of 63 different fields were involved in the job analysis survey. Would it be reasonable to keep the original approximately 12,600 response sheets for years after they had been scored and recorded on computer tapes? A publisher building a variety of tests for a number of different agencies would soon need an exorbitant amount of storage space. Of course some records need to be maintained. For example, it would seem necessary to retain the examination answer sheets for at least some period of time (perhaps 1 year) to allow for verification by those who may wish to challenge the accuracy of the scoring process. To many experts it would seem sufficient to have the item ratings from individuals on validity panels recorded on computer tape. However, others believe the original rating sheets should be maintained.

Most argue that a *detailed* resume on each person on a validity panel need not be placed in a formal report, or even necessarily kept on file. Of course their *names* should be available so that someone who wishes can check their qualifications. Indeed there should be some sort of summary statement about the training, the experience and the qualification of the panel members. But there is room for disagreement among experts about the extent of the documentation required and, as mentioned, publishers would be wise to be diligent. However, it is *not* logical to infer that a test produces scores with low validity because of a lack of documentation. Documentation of the test construction process, per se, does not influence the scores or their validity. The correct inference from inadequate documentation is simply that there is inadequate documentation not that the test scores are invalid.

CONCLUSION

In the past our colleges and state licensing boards have failed as adequate gatekeepers to the teaching profession. Consequently, almost all states have implemented some sort of competency assess-

ment of teachers. In this chapter I have discussed the issues of reliability, validity, setting the cut score, reporting scores, and the necessity to provide documented evidence of test quality.

Even if we have top notch tests of competence and set the "correct" cut score we need to recognize the limitations of the inferences we can make from licensure tests. Such tests do not address the issues of overall teacher performance or teacher effectiveness. A high quality 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 (because of false positives) even all incompetent teachers. It should help ensure that those individuals who are licensed have a minimal level of competence on some important subdomains of knowledge and skills relevant to their profession. That is a step in the right direction.

REFERENCES

American Educational Research Association, American Psychological Association, & National Council on Measurement. (1985). *Standards for educational and psychological testing.* Washington, DC: 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, DC: Author.

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

Berk, R. A. (1984a). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 199–230). Baltimore: Johns Hopkins University Press.

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

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

Berliner, D. C. (1986, April). *In pursuit of the expert pedagogue.* Presidental Address of the 1986 Annual Meeting of the American Educational Research Association. San Francisco, CA.

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

Bond, L. (1987). The Golden Rule Settlement: A minority perspective. *Educational Measurement: Issues and Practice. 6*(2), 18–20.

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.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk

(Ed.). *A guide to criterion-referenced test construction.* (pp. 292–334). Baltimore: Johns Hopkins University Press.

Brennan, R. L., & Kane, M. T. (1977). An index of dependability for mastery tests. *Journal of Educational Measurement, 14,* 277–289.

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.

Cameron, D. (1985). The NEA position on testing in-service teachers. *Educational measurement: Issues and practice, 4*(3), 26–27.

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, DC: U.S. Department of Health, Education and Welfare, Office of Education.

Colton, D., Kane, M., Kingsbury, C., & Estes, C. (1987, April). *Examining the construct validity of job analysis data: A strategy and an example.* Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

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 & 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, Vol. 5* (pp. 99–108). 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.

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), 285–328.

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.

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

Downs, S. L., & Silvestro, J. R. (1987, April). *Support systems for teacher certification testing programs.* Paper presented at the annual meeting of the American Educational Research Association. Washington, DC.

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, NJ: Prentice Hall.

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

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

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

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

Ekstrom, R. B., & Goertz, M. E. (1985, April). *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. (1987, April). *Validating job analysis survey instruments used in developing teacher certification tests: A construct validity study.* Paper presented at the annual meeting of the National Council of Measurement in Education. Washington, DC.

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, August 25). Uniform guidelines on employee selection procedures. *Federal Register, 43*(166), 38290–38315.

Feldt, L. S., Steffen, M., & Gupta, N. C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement, 9,* 351–361.

Freeman, L. D. (1977). State interest, certification, and teacher education program approval. In W. A. Hazard, L. D. Freeman, S. Eisdorfer, & P. Tractenberg (Eds.), *Legal issues in teacher preparation and certification* (pp. 67–108). Washington, DC: ERIC Clearinghouse on Teacher Education.

Gabrys, R. E. (1987). State reaction to national teacher testing and certification issues. In National Evaluation Systems, Inc. (Ed.). *Trends in teacher certification testing* (pp. 25–29). Amherst, MA: National Evaluation Systems.

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.

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

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

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

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

Haertel, E. (1987, April). *Structuring item domains to map the school curriculum.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Hall, B. W. (1985). Survey of the technical characteristics of published educational achievement tests. *Educational Measurement: Issues and Practice, 4*(1), 6–14.

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

Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests. *Journal of Educational Measurement, 10,* 159–170.

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, DC: National Council on Measurement in Education.

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

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

Holmes Group, The (1986). *Tomorrow's teachers: A report on the Holmes Group.* The Holmes Group. East Lansing, MI.

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

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, NJ: Lawrence Erlbaum Associates.

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.

Kane, M. T., Kingsbury, C., Colton, D., & Estes, C. (1986). *A study of nursing practice and role delineation and job analysis of entry-level performance of registered nurses.* Chicago, IL: National Council of State Boards of Nursing.

Levine, E. L., Ash, R. A., Hall, H. L., & Sistrunck, F. (1981). *Evaluation of seven job analysis methods by experienced job analysts.* Unpublished manuscript. 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 M. Jaeger & J. C. Busch, (Eds.), *An evaluation of the Georgia Teacher Certification testing program* (Chap. 4). Greensboro, NC: Center for Educational Research and Evaluation. University of North Carolina.

Livingston, S. A. (1972). Criterion-referenced applications of classical test theory. *Journal of Educational Measurement, 9,* 13–26.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting stan-*

dards of performance on educational and occupational tests. Princeton, NJ: Educational Testing Service.

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, VA: Association of Teacher Educators and the Bureau of Educational Research, University of Virginia.

Mehrens, W. A. (1986a). *Validity issues in teacher competency tests.* Gainesville, FL: University of Florida, Institute for Student Assessment and Evaluation.

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

Mehrens, W. A. (1987a). *Issues in teacher competency tests.* Prepared for the Commission on Testing and Public Policy Graduate School of Education, University of California, Berkeley.

Mehrens, W. A. (1987b). Validity issues in teacher competency tests. *Journal of Personnel Evaluation in Education, 1,* 195–229.

Mehrens, W. A., & Lehmann, I. J. (1984). *Measurement and evaluation in education and psychology* (3rd ed.). New York: Holt, Rinehart, & Winston.

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. (1979). Reliability and validity of criterion-referenced test scores. In R. E. Traub (Ed.), *New directions for testing and measurement:* (No. 4): *Methodological developments* (pp. 75–92). San Francisco, CA: Jossey–Bass.

Millman, J. (1986). Review of test development and score reporting procedures. In R. M. Jaeger & J. C. Busch (Principal Investigators). *An evaluation of the Georgia Teacher Certification Testing Program* (Chap. 3). Greensboro, NC: 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 Evaluation Systems. (n.d.). *Study Guide: Examination for the certification of educators in Texas.* Amherst, MA. Author.

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, NJ: Lawrence Erlbaum Associates.

Peterson, K. D. (1987). Teacher evaluation with multiple and variable lines of evidence. *American Educational Research Journal, 24*(2), 311–318.

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 construction* (pp. 29–48). Baltimore: 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, DC: National Council on Measurement in Education.

Pratt, D. (1977). Predicting teacher survival. *The Journal of Educational Research, 71*(1), 12–18.

Rebell, M. A. (1986). Disparate impact of the teacher competency testing on minorities: Don't blame the test-takers—or the tests. *Yala Law & Policy Review, 4,* 2, 372–403.

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. Alabama State Board of Education et al. Civil Action No. 81-697-N.

Roid, G. H. (1984). Generating the test items. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 49–77). Baltimore: 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, DC.

Rosenfeld, M., Shimberg, B., & Thornton, R. F. (1983). *Job analysis of licensed psychologists in the United States and Canada.* Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service.

Rosenfeld, M., Thornton, R. F., & Skurnik, L. S. (1986, March). *Analysis of the professional function of teachers: Relationships between job functions and the NTE Core Battery* (Research Report 86–8). Princeton, NJ: 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.

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 measurement and evaluation* (2nd ed.). Belmont, CA: Wadsworth.

Schmeiser, C. B. (1987, April). *Effects of translating task analysis data into test specifications.* Paper presented at the annual meeting of the National Council on Measurement in Education. Washington, DC.

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.

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 Educational Research Association, San Francisco.

Shepard, L. A., & Kreitzer, A. E. (1987, April). *The Texas teacher test.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Shimberg, B. (1982). *Occupational licensing: A public perspective.* Princeton, NJ: 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. (n.d.). *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.

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

Subkoviak, M. J. (1984). Estimating the reliability of mastery-nonmasterly classifications. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 267–291). Baltimore: Johns Hopkins University Press.

Sweeney, J., & Manatt, R. P. (1986). Teacher evaluation. In R. A. Berk (Ed.), *Performance assessment: Methods and applications* (pp. 446–468). Baltimore: Johns Hopkins University Press.

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 Economy, New York.

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

Time. (1980, June 16). Help! Teachers can't teach. *Time Magazine,* pp. 54–63.

Tittle, 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: Johns Hopkins University Press.

Traub, R. E. (1986). Review of test reliability assessment procedures and results. In R. M. Jaeger & J. C. Busch (Eds.), *An evaluation of the Georgia Teacher Certification testing program* (Chap. 5). Greensboro, NC: Center for Educational Research and Evaluation, University of North Carolina.

U.S. Civil Service Commission. (1973). *Job Analysis: Key to better management.* Washington, DC: 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.* PHEW publication 72-11. Washington, DC: Author.

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, DC.

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, NJ: Lawrence Erlbaum Associates.

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, NJ: Lawrence Erlbaum Associates.

Webb, L. D. (1983). Teacher evaluation. In S. B. Thomas, N. H. Cambron-McGabe, & M. M. McCarthy (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.

Willimason, 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. 61–86). 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.