1995

2. Legal And Professional Bases For Licensure Testing

William A. Mehrens
Michigan State University

2

LEGAL AND PROFESSIONAL BASES FOR LICENSURE TESTING

William A. Mehrens

Michigan State University

In this chapter the author presents the legal setting for licensure testing,[1] discusses the role of various professional standards and codes (i.e., the EEOC *Uniform Guidelines*, 1978, and the AERA/APA/NCME *Standards*, 1985), presents some of the pertinent rulings from several court decisions, and makes inferences about future changes in professional standards and their potential impact on licensure test development.

There necessarily is some minor overlap with the material in this chapter and some other chapters in this book. There is a brief discussion of the differences between licensure, certification. and employment testing and how those differences relate to the professional standards and court cases. It is necessary to mention some concepts such as task analysis, validity, and cut scores when discussing the professional standards and the court cases. However, these concepts are not dealt with in the depth that occurs in later chapters.

THE LEGAL SETTING

Licensure and certification tests are high-stakes tests and those considering using or constructing such tests should be aware of previous case law regarding

---

Portions of this chapter have been adapted from an article by Mehrens, W.A. and Popham, W.J. (1992). How to evaluate the legal defensibility of high-stakes tests. *Applied Measurement in Education, 5*(3), 265-283. Permission of the publisher and Dr. Popham to use those portions has been obtained. Special appreciation is given to Dr. Kara Schmitt and Susan Boston for their assistance in tracking down many of the legal documents used in writing this chapter.

[1]The words "test" and "testing" are to be interpreted broadly as including a variety of assessment procedures.

such testing. Some generic legal issues are discussed first. In subsequent sections, the various professional standards and some court decisions are presented.

Generic Legal Issues

Existing case law is based on constitutional requirements—primarily the 14th Amendment—and statutory requirements—primarily Title VII of the 1964 Civil Rights Act.

Constitutional Requirements: The 14th Amendment

Two basic requirements of the U.S. Constitution's 14th Amendment are discussed: equal protection and due process. For a plaintiff to win under the equal protection analysis, it must be shown that there was intent to discriminate. In *Village of Arlington Heights v. Metropolitan Housing Development Corp.* (1977), the court stated that the following factors could be considered in establishing discriminatory intent: (a) historical background, (b) the specific sequence of events leading up to the challenged decision, (c) departures from normal procedural sequences, and (d) the legislative or administrative history. Nevertheless, to prove discriminatory intent, one court has ruled that it must be shown that the user of the test "selected or reaffirmed a particular course of action at least in part `because of' not merely `in spite of,' its adverse effects upon an identifiable group" (*Personnel Administrator v. Feeney*, 1979, at 4656). Another court has stated that:

> An action does not violate the equal protection clause simply because the decision maker knows that it will have a disparate impact on racial or ethnic groups. (*United States v. LULAC*, 1986, p. 646)

It is difficult to *prove* intent. As a consequence, most plaintiffs would prefer basing their cases on the Civil Rights Acts, which do not require proof of discriminatory motive.

The due process provisions of the Constitution relate to substantive and procedural due process. Substantive due process requires a legitimate relationship between a requirement and the purpose. This legitimate relationship is easier to establish than the business necessity requirement of the Civil Rights Acts. In fact, for licensure and certification challenges Herbsleb, Sales, and Overcast (1985) concluded that:

> the rationality standard is so lenient that we were unable to find a single case where an examination was successfully challenged on this basis. (p. 1169)

Procedural due process requires fairness in the way things are done. In testing cases, this means that there must be advance notice of the requirement, an opportunity for hearings/appeals, and that the hearings must be conducted fairly. A licensure or certification testing program should not be implemented without paying careful attention to these procedures. It should be pointed out that if a plaintiff wins on procedural grounds, he/she does not necessarily get a license. However, some additional procedure—such as a hearing—must be applied.

Statutory Requirements: The Civil Rights Acts

The 1964 Civil Rights Act was a general federal statute prohibiting discrimination in employment. When first enacted it pertained to employment in the private

sector, but it was extended in 1972 to employment practices in educational institutions. The Civil Rights Act of 1991 was passed to reverse parts of several U.S. Supreme Court decisions that were unfavorable to employment complaints. There is some debate about whether licensure and certification procedures are to be considered employment practices and whether the Civil Rights Acts apply to such processes. This is discussed in more detail later.

The Acts prohibit two kinds of discrimination: disparate treatment and disparate impact. Disparate treatment involves overt discrimination—where employers treat some people less favorably than others because of their race, color, religion, or national origin. The plaintiff has the initial burden of establishing that disparate treatment occurred. Most case law related to the Civil Rights Acts regarding testing is based on disparate impact rather than disparate treatment.

Disparate impact does not require evidence of subjective discriminatory *intent*, but refers to employment practices that are ostensibly neutral in their treatment, yet result in protected groups being hired at a lower rate than unprotected groups. It is the plaintiff's responsibility to show disparate impact, but it is the responsibility of the user (e.g., employer or licensure board) to maintain documentation regarding disparate impact (see *Chance v. Board of Examiners*, 1971, 1972). The Civil Rights Act of 1991 states that the plaintiff must demonstrate that each *particular* challenged process (e.g., written test, subtest, oral exam, performance appraisal) causes a disparate impact unless the plaintiff can demonstrate that the decision-making elements cannot be analyzed separately. (This emphasis on each component may have implications for scoring procedures—should one use part scores or total scores—and conjunctive versus compensatory decision making.)

There exists some debate about what statistics to use and what groups should be considered in the statistical analysis to show disparate impact. Regarding the relevant groups, the general conclusion is that the proper comparison is between the proportions of the groups in the qualified population in the relevant job market (*Wards Cove Packing Co.*, 1989; Civil Rights Act of 1991). For the statistical analysis, the *Uniform Guidelines*[2] (EEOC, 1978) suggest a four-fifths rule. This means that the percent of protected group applicants hired should be at least 80% of the percent of unprotected group applicants hired. Others prefer a statistical inference test to discern if an observed disparity between protected and unprotected groups is statistically significant (e.g., Hazelwood, 1977). Because the issue of impact is not one of test construction and use, per se, we will not discuss it further. However, interested readers may wish to consult the literature concerning this issue (see, e.g., Meier, Sacks, & Zabell, 1984).

In cases where there has been a showing of disparate impact on members of a protected group for a particular employment practice, the burden of proof shifts to the defendants and requires them to demonstrate that the use of the test (or other assessment procedure) constitutes a *business necessity*. (Employers do not need to defend those parts of the process that do not show disparate impact.) This means that the particular challenged tests (or subtests) must be shown to be job-related and

[2]The *Guidelines* is a single work. However, for smoothness in reading it will be treated as a plural noun.

to have been professionally developed. If a test is job-related and professionally developed, it can be used even if there is disparate impact unless the plaintiffs can show that there exists an equally effective alternative selection procedure that results in less adverse impact. Although there were some Supreme Court decisions in 1988 and 1989 that lessened the burden of proof of the defendants to show business necessity, the 1991 Civil Rights Act reestablished this requirement.

Title VII and Employment, Licensure, and Certification Testing

As discussed in the previous chapter, the purposes of licensure and certification tests are different from the purpose of employment tests. The function of licensure is to protect the health, safety, and welfare of the public. There is some debate about whether Title VII of the Civil Rights Act applies to licensure tests. Some attorneys (e.g., Phillips, 1991; Pyburn, 1990; Rebell, 1986) have suggested that Title VII does not apply to state licensing agencies and their tests. Rulings in bar examination cases such as *Tyler v. Vickery* (1975, 1976) and *Woodward v. Virginia Board of Bar Examiners* (1976, 1979) support this position. For example, one court stated that:

> Title VII does not apply by its terms...because the Georgia Board of Bar Examiners is neither an "employer," an "employment agency," nor a "labor organization" within the meaning of the statute. (*Tyler v. Vickery*, 1976, p. 1096)

Smith and Hambleton (1990) concluded that:

> Most courts have been unwilling to extend Title VII...to licensure examinations. (p. 8)

Shimberg (1990) reached the same conclusion.Others believe that at least for teacher licensure, the State can be viewed as an employer (see Kuehn, Stallings, & Holland, 1990). Freeman, Hess, and Kasik (1985) discuss why teacher licensure may be unique. They suggest that:

> the history of certification in most states indicates that certification has been intimately interwoven in the employment process. (p. 14)

They argue further that:

> Teaching as a profession is somewhat peculiar because teachers are certified or licensed to work exclusively in institutions that are created, maintained, and more or less financed by the state. (p. 23)

The above quote is not precisely true because many private school, parochial school, and home school teachers are licensed. Nevertheless, some courts may view it as a relevant argument.

Based, in part, upon the number of teacher certification test cases *filed* under Title VII, and the number of employment testing cases *cited as relevant* precedent in teacher certification test litigation, Kuehn, Stallings, and Holland (1990) believe Title VII does apply to *teacher* licensure. They suggest that:

> If the Courts treat teacher certification tests as employee selection procedures, we are compelled to construct them and defend them as employee selection procedures. (p. 21)

The problem with the above quote is that it is widely recognized that licensure tests serve different purposes from employment tests and this *should* result in different test construction and validation procedures. One difference is that a person is employed to do a specific job whereas a license allows the person to engage in *diverse* jobs. Freeman et al. recognized this problem and concluded that:

> examining certification requirements to determine their job-relatedness becomes an almost hopeless task. (1985, p. 25)

The EEOC *Uniform Guidelines* address this whole issue, but the statements are not decisive. The *Guidelines* state that "licensing and certification are covered `to the extent' that licensing and certification may be covered by Federal equal employment opportunity law" (Equal Employment Opportunity Commission [EEOC], 1978, p. 38294). They further state that:

> Voluntary certification boards, where certification is not required by law, are not users ... with respect to their certifying functions and therefore are not subject to these guidelines. If an employer relies upon such certification in making employment decisions, the employer is the user and must be prepared to justify, under Federal law, that reliance as it would any other selection procedure. (1978, p. 38294)

Thus, if an employer used the results of a certification test for promotion, or a differential salary, it would be used as an employment exam and be subject to Title VII. For example, consider the proposed certification tests of the National Board of Professional Teaching Standards. These are intended to be voluntary in the sense that licensed teachers will not have to take them to maintain their licenses. However, if a state or local district chose to reward certified teachers with additional salary, that may be considered an employment decision and the Civil Rights Acts (Title VII) might apply. But it would apply to the state or local unit that uses the test for decision making.

The issue of the relevance of Title VII to licensure and certification tests is important because Title VII calls for a business necessity requirement, which is considered harder to demonstrate than the legitimate relationship requirement that would otherwise apply to licensure tests. Because there is some disagreement about whether (or under what circumstances) licensure and certification testing programs are subject to the Civil Rights Acts requirements, this chapter discusses guidelines for both types of settings. This author's view is that most licensure and certification testing programs should *not* be ruled as employment programs, but others, used in different fashions, might be.

PROFESSIONAL STANDARDS AND CODES

There are several sets of professional standards and codes that should be considered when constructing a licensure or certification examination. The two major ones are the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA/APA/NCME], 1985), hereafter referred to as the *Standards*; and the *Uniform Guidelines on Employee Selection Procedures* (EEOC, 1978), hereafter referred to as the *Guidelines*.

Prior to discussing these standards and codes, it should be emphasized that both the *Standards* and the *Guidelines* are somewhat dated. Both documents explicitly recognize that they need to be interpreted keeping this datedness factor in mind. The *Standards*[3] note that they are concerned "with a field that is evolving" (AERA/APA/NCME, 1985, p. 2) and the *Guidelines* point out that "they will have to be interpreted in light of changing factual, legal, and professional circumstances" (EEOC, 1978, p. 38292). In a later section, current psychometric views and potential future directions in the field and how they may impact legal issues and future revisions of the *Standards* and *Guidelines* are discussed.

AERA/APA/NCME Standards

The 1985 *Standards* constitute the fifth in a series of documents from the three sponsoring organizations regarding the development and use of tests and they supersede the previous documents.

> In general, the *Standards* advocates that, within feasible limits, the necessary technical information be made available so that those involved in policy debate may be fully informed. The *Standards* does not attempt to provide psychometric answers to policy questions. (AERA/APA/NCME, 1985, p. 1)

The *Standards* are divided into four parts. Part I covers technical standards for test construction and evaluation. Included in this part are chapters on such topics as validity, reliability, and norming, score comparability, and equating. Part II covers standards for test use. The chapter on licensure and certification testing is of major importance to readers of this volume although the chapter on employment testing is mentioned. Part III covers standards for particular applications and the chapter on testing the disabled is particularly important. Finally, Part IV presents standards for administrative procedures.

The *Standards* point out that their use in litigation is inevitable, but that "professional judgment ... always plays an essential role in determining the relevance of particular standards in particular situations" (AERA/APA/NCME, 1985, p. 2). Further, it is stressed that:

> evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every primary standard in this document, and acceptability cannot be determined by using a checklist. (AERA/APA/NCME, 1985, p. 2)

Although the *Standards* represent an "official" guideline to be judgmentally followed, it should be recognized that there is less than consensus in the psychometric community about various components of the *Standards*. For example, regarding the concept of test validity, Linn, comments on the Joint Committee's attempt

> to carry this unified view of validity a bit further, but not, I might add, without significant objection from a number of people. ... A number of reviewers considered such a requirement to be overly demanding. (Linn, 1984, p. 4)

Shimberg has stated that the writers of the *Standards* did not obtain consensus "among all those who prepare and use licensing and certification tests

[3]The *Standards*, like the *Guidelines*, is a single work. However, for smoothness in reading it also will be treated as a plural noun.

regarding what constitutes acceptable professional practice in these areas" (1990, p. 13).

In spite of the above comments, the *Standards* are (correctly in my opinion) used as a guide in the development of a licensure or certification test, and one should try to follow the *relevant* standards. The subsections that follow discuss some of the most pertinent standards from various chapters of the *Standards*.

Validity Standards

The validity chapter of the *Standards* states that "validity is the most important consideration in test evaluation" (AERA/APA/NCME, 1985, p. 9) and presents 25 different standards regarding validity.

Certainly many of the standards in this chapter are relevant. However, it is clear that not even all of these are relevant for any given test development/use project. For example, in the validity chapter, Standard 1.1 states that "evidence of validity should be presented for the major types of inferences for which the use of a test is recommended" (AERA/APA/NCME, 1985, p. 13). By implication, and by the comment following the standard, it is obvious that one would not have to gather all the types of validity evidences that are addressed in the *Standards* for any particular use. The separate chapters in Part II on various uses of tests make that clear also.

Validity is a technical area where the field has changed its nomenclature, if indeed not its approach. The *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. (AERA/APA/NCME, 1985, p. 9)

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

> 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)

Because content-related validity evidence is likely to be one type of validity evidence that will be gathered, it seems important to consider the validity standards that relate particularly to content-related evidence. Standard 1.3 relates indirectly and Standard 1.6 directly to content-related evidence.

> Standard 1.3: Whenever interpretation of subscores, score differences, or profiles is suggested, the evidence justifying such interpretation should be made explicit. Where composite scores are developed, the basis and rationale for weighting the subscores should be given. (Primary) (AERA/APA/NCME, 1985, p. 14).

> Standard 1.6: When content-related evidence serves as a significant demonstration of validity for a particular test use, a clear definition of the universe represented, its relevance to the proposed test use, and the procedures followed in generating

> test content to represent that universe should be described. When the content sampling is intended to reflect criticality rather than representativeness, the rationale for the relative emphasis given to critical factors in the universe should also be described carefully. (Primary) (AERA/APA/NCME, 1985, p. 14)

The last sentence in the above quoted standard is particularly important because, as will become more clear when discussing Chapter 11 of the *Standards*, one often wishes for a critical rather than representative domain in licensure testing.

Reliability Standards

The reliability chapter of the *Standards* presents 12 different standards. Some of the more important reliability standards that should be attended to are as follows:

> 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... (Primary) (p. 20)

> Standard 2.10: Standard errors of measurement should be reported at critical score levels. Where cut scores are specified for selection or classification, the standard errors of measurement should be reported for score levels at or near the cut score. (Secondary) (p. 22)

> Standard 2.12: 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) (AERA/APA/NCME, 1985, p. 23)

Test Development and Revision Standards

The chapter on test development and revision presents 25 different standards. The standards primarily relate to building a test in a correct fashion. The major overriding standard in this chapter is Standard 3.1, which states that "Tests and testing programs should be developed on a sound scientific basis" (p. 25). Standard 3.2 states that the definition of the universe or domain must be described. Many of the other standards in this chapter would also be appropriate for licensure and certification examinations.

Scaling, Norming, Score Comparability, and Equating Standards

It is certainly important that there be score comparability and equating of tests given at different times for licensure and certification exams, and the nine standards presented in this chapter relevant to those issues should be considered in test development. The standard most relevant for licensure tests is Standard 4.8 which speaks to the content and statistical requirements for anchor test items if an anchor test design is used for equating.

Setting the Cut Score

For licensure tests, the precision of the equating at the cut store is of primary importance. There is no chapter in the *Standards* directly related to this issue and the *Standards* do not make any recommendation regarding specific standard setting procedures. However, they do suggest that the method and rationale of setting the cut score, as well as the qualifications of the judges, should be documented (see *Standards* 6.9 and 10.9).

Standards Specific to Employment Testing

Chapter 10 of the *standards* is on employment testing. If a developer/user of a licensure or certification test believes that it will be regarded by the courts as an employment examination, then attention should be given to the standards in this chapter. As mentioned above, this author does *not* consider licensure tests to be employment tests, but some *uses* of *certification* tests (promotion, differential tasks or differential salaries based on the tests) may place them in that category. The major difference between the standards for employment testing and licensure and certification testing is that employment testing standards place more emphasis on criterion-related validity evidence.

Professional and Occupational Licensure and Certification Standards

Chapter 11 of the *Standards* focuses directly on professional and occupational licensure and certification examinations. As the *Standards* point out, "several hundred occupations are now regulated by state governments. Many other occupations are certified by nongovernmental agencies" (p. 63). The *Standards* discuss the different purposes of employment and licensure examinations already discussed in this book, and point out the implications of those differences for various issues of validity. For licensure and certification, the focus is on necessary skills and knowledge, whereas the employer may wish to maximize productivity. The *Standards* make clear 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. ... For licensure and certification, ...primary reliance must usually be placed on content evidence..." (AERA/APA/NCME, 1985, p. 63)

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

Although the *Standards* appropriately emphasize the importance of content-related validity evidence over criterion-related or construct validity evidence for licensure tests, builders or users of licensure tests should not think they "have it easy" in constructing licensure tests that meet the *Standards*. The requirements of content validity are quite explicit and demanding.

> Standard 11.1: The content domain to be covered by a licensure or certification test should be defined clearly and explained in terms of the importance of the content for competent performance in an occupation. A rationale should be

> provided to support a claim that the knowledge or skills being assessed are required for competent performance in an occupation and are consistent with the purpose for which the licensing or certification program was instituted. (Primary) (AERA/APA/NCME, 1985, p. 64)

The comment for Standard 11.1 points out that "job analyses provide the primary basis for defining the content domain," that "the emphasis for licensure and certification is limited appropriately to knowledge and skills necessary to protect the public," and that "skills that may be important to success but are not directly related to the purpose of licensure (i.e., protecting the public) should not be included in a licensing exam" (AERA/APA/NCME, 1985, p. 64).

Two final standards from this chapter seem particularly relevant.

> 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. (Primary) (p. 65)

> Standard 11.5: Rules and procedures used to combine scores or other assessments to determine the overall outcome should be reported to test takers preferably before the test is administered. (Secondary) (AERA/APA/NCME, 1985, p. 65)

The comment for Standard 11.5 points out that:

> In some cases candidates may be required to score above a specified minimum on each of several tests. In other cases the pass-fail decision may be based solely on a total composite score. (AERA/APA/NCME, 1985, p. 65)

These last two standards and the comment for Standard 11.5 need to be considered along with Standard 2.1 quoted above. If the test is not unidimensional, the subscores provide potentially useful information for failing candidates who wish to direct their subsequent review and study to their areas of weakness. If these subscores are reported for remediation purposes and are not used in a conjunctive model but are simply used in a total composite score in a compensatory model, it is debatable whether the scores have been used "in the decision process." They have not been used in the licensure decision, but may be used by the failed candidate for remediation purposes. In writing specifically about teacher licensure examinations, Mehrens has suggested that:

> 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* [emphasis added]. (1990, p. 85)

It seems reasonable to generalize from this point to any licensure examination use where the decision is based on a total composite score. One final point deserves emphasis. The quoted comment accompanying Standard 11.5 suggests that it is appropriate to base pass-fail decision "solely on a total composite score." Although this author agrees with that position, a common statement heard from expert witnesses for plaintiffs is that one should not make a decision on only a single piece

of data. Obviously, that stated opinion ignores the fact that there was probably a sequential decision-making model employed requiring other acceptable data on additional variables prior to being allowed to sit for the licensure examination, and it ignores this specific standard that specifically accepts making a decision solely on a composite score.

Standards on Testing Individuals with Disabilities

Chapter 14 of the *Standards* presents eight standards for testing individuals with disabilities. With the passing of the Americans with Disabilities Act (1990), which became effective in 1992, there has been much discussion regarding what accommodations need to be made for individuals with claimed disabilities. This issue has been considered in depth in other publications. For example, Millman, Mehrens, and Sackett address this issue for the New York Bar Examination in detail (1993). Clearly, there is some obligation to allow individuals with physical disabilities to be accommodated when the knowledge and skills needed for licensure are not the specific physical skills which are being accommodated. Probably the biggest areas of concern are with those who claim learning disabilities. These are hard to classify and most classification schemes result in a large number of false positives. Whether correctly or incorrectly classified, there is the issue of what is a fair accommodation for individuals with a cognitive disability when the job in question demands cognitive functioning. The largest specific issue probably relates to the amount of time extension that should be given to individuals with disabilities. If the job in question demands primarily physical skills, then it would be reasonable to grant accommodations to those with learning disabilities, but it may not be reasonable to grant them to those with physical disabilities.

Some of the major points made in the eight standards are as follows:

> Standard 14.1: People who modify tests for handicapped people should have available to them psychometric expertise for so doing. (p. 79)
>
> Standard 14.2: Until tests have been validated for people who have specific handicapping conditions, test publishers should issue cautionary statements in manuals and elsewhere regarding confidence in interpretations based on such test scores. (p. 79)
>
> Standard 14.5: Empirical procedures should be used whenever possible to establish time limits for modified forms of timed tests rather than simply allowing handicapped test takers a multiple of the standard time. (p.79)
>
> Standard 14.6: When feasible, the validity and reliability of tests administered to people with various handicapping conditions should be investigated and reported by the agency or publisher that makes the modification. (AERA/APA/NCME, 1985, p. 80)

EEOC Uniform Guidelines

The *Uniform Guidelines* (EEOC, 1978) are a set of guidelines on employee selection procedures that have been adopted by the Equal Employment Opportunity Commission, the Civil Service Commission, the Department of Justice, and the Department of Labor. In addition to being quite dated, there is, as has been

mentioned, some debate about whether (or when) they might apply to licensure and certification exams. As is stated:

> These guidelines apply to tests and other selection procedures which are used as a basis for any employment decision. Employment decisions include but are not limited to hiring, promotion, demotion, membership (for example in a labor organization), referral, retention, and licensing and certification, to the extent that licensing and certification may be covered by Federal equal employment opportunity law. (EEOC, 1978, p. 38296)

They also state that:

> Voluntary certification boards, where certification is not required by law, are not users as defined...with respect to their certifying functions and therefore are not subject to these guidelines. If an employer relies upon such certification in making employment decisions, the employer is the user and must be prepared to justify, under Federal law, that reliance as it would any other selection procedure. (EEOC, 1978, p. 38294)

Whether or not the *Guidelines* apply in licensure, it is important to realize that they "have been given great weight by the courts in Equal Protection as well as Title VII cases" (Eisdorfer & Tractenberg, 1977, p. 121; see also, Rebell, 1990a, p. 347).

Under the *Guidelines*, to use a measure that produces adverse impact, the employer

> must justify the use of the procedure on grounds of 'business necessity.' This normally means that it must show a clear relation between performance on the selection procedure and performance on the job. (EEOC, 1978, p. 38291)

Although users need not validate procedures which do not have an adverse impact,

> if one way of using a procedure (e.g. ranking) results in greater adverse impact than another way (e.g. pass/fail), the procedure must be validated for that use. (EEOC, 1978, p. 38294)

There are no major contradictions between the *Guidelines* and the *Standards*, however, the *Guidelines* are more explicit than the *Standards* on some dimensions (e.g., they require that any cutoff score be justified by reference to the "need for a trustworthy and efficient work force" [EEOC, 1978, p. 38291], and that when "cutoff scores are used, they should normally be set so as to be reasonable and consistent with normal expectations of acceptable proficiency within the work force" [EEOC, 1978, p. 38298]). The *Guidelines* terminology of "normal expectations" clearly suggests a judgmental approach for setting a cutoff score. However, the *Guidelines* suggest that rank ordering requires substantial evidence of validity and a reasonable expectation that small differences in scores would reflect real differences in job performance.

The *Guidelines* address the three types of validity evidence and state that "users may rely upon criterion-related validity studies, content validity studies or construct validity studies" (EEOC, 1978, p. 38298). They recognize the lack of a clear distinction between types of validity evidence and try to address the borderline

between content validity and construct validity. As an example, the *Guidelines* state that for typing, a typing test:

> is justifiable on the basis of content validity because it is a sample of an important or critical part of the job...but [the *Guidelines*] do not allow the validation of a test measuring a construct such as "judgment" by a content validity strategy. (EEOC, 1978, p. 38292)

Other quotes from the *Guidelines* relevant to validity are as follows:

> Any validity study should be based upon a review of information about the job for which the selection procedure is to be used. (p. 38300)

> A selection procedure can be supported by a content validity strategy to the extent that it is a representative sample of the content of the job. (p. 38302)

> A selection procedure based upon inferences about mental processes cannot be supported solely or primarily on the basis of content validity. (EEOC, 1978, p. 38302)

Finally, it should be mentioned that the *Guidelines* stress the importance of record keeping and documentation.

> Users of selection procedures...should maintain and have available for each job information on adverse impact of the selection process for that job and, where it is determined a selection process has an adverse impact, evidence of validity...Where a total selection process for a job has an adverse impact, the user should maintain and have available records or other information showing which components have an adverse impact. (EEOC, 1978, 38303).

STATE AND FEDERAL COURT DECISIONS

There are differences in case law and test construction processes between employment and licensure testing and the case law precedents will be discussed separately. For each type of test, some of the pivotal cases are identified and what made those cases important is described. In generalizing from the rulings in these cases, it should be pointed out that a legal case is binding only on lower courts in the same jurisdiction. For example, Federal Supreme Court rulings are binding on all other Federal Courts, but an Appeals Court ruling in, for instance, the 5th Circuit would be binding only on lower courts in that circuit. Also, the decisions are binding only on cases that are factually similar. Nevertheless, even cases not binding may be broadly instructive.

Employment Cases

The *Griggs v. Duke Power Company* case (1971) was the first landmark case dealing with job-related testing. The court ruled that in employment testing in private industry the defendants must show the job relatedness of the test. "Broad and general testing devices ... as fixed measures of capacity" were barred in employment testing (Griggs, 1971, p. 433). In *Albermarle Paper Company v. Moody* (1975), it was held that the EEOC *Guidelines* (revised in 1978) were the fundamental benchmark for assessing Title VII of the 1964 Civil Rights Act job relatedness requirements. These *Guidelines* constituted the administrative interpre-

tation of the act by the enforcing agency and "consequently are entitled to great deference" (*Griggs v. Duke Power Company*, 1971, 401 U.S., at 433-434). *Chance v. Board of Examiners* (1972) established a job relatedness precedent for tests used with public employees as well as private employees.

Thus, it is clear that employers can be challenged regarding the job-relatedness of their employment practices. When challenged, employers must show that their test development procedures followed acceptable professional practices, with the EEOC *Guidelines* being considered an important guide. However, in *Guardians Association of N.Y. City v. Civil Service Commission* (1980), the court ruled that the *Guidelines* adopted too rigid an approach in the selection of validation techniques and that it was inconsistent with Title VII's endorsement of professionally developed tests. The Court basically considered content validation strategies to be acceptable for a test that assessed *observable* abilities. The court stated that content validation should *not* be rejected just because the abilities measured could be classified as constructs.

In an earlier decision (*Washington v. Davis*, 1976) the Supreme Court accepted the use of a verbal skills test for entry into police training even though its use had adverse impact because the scores correlated with performance in the training program and that training program completion is a prerequisite to employment. It should be mentioned that:

> Title VII standards were not applied in *Washington v. Davis* because the statute was not applicable to federal employees when the case was initially filed. (Cohen, 1989, p. 240)

However, the Court commented that had the job-relatedness requirements of *Griggs* or *Albermarle Paper* been applied, the correlation with the training program would have been sufficient validation.

In a fairly recent court decision (*Richardson v. Lamar County Board of Education*, 1989, 1991) a school district was challenged for using the Alabama Initial Teacher Certification Test. This test was originally intended as a licensure examination. Thus, although the case was technically an employment case, it may have implications for licensure examinations. The judge ruled against the district's use of the test. Judge Thompson's decision contained a fairly extensive analysis of perceived problems in test development and standard setting processes in the Alabama Initial Teacher Certification Test. Judge Thompson ruled that:

- first try failure statistics can be used for determining the extent of adverse impact because initial failure is a discrete injury (even though another court had previously ruled otherwise—see *United States v. LULAC*, 1986);
- outside experts should have been retained to monitor the test developer's work;
- all items should have been reviewed by committee members and suggested changes in items should not have been ignored by the test developer;
- the developer should have conducted empirical bias studies (even though for many of the tests the sample sizes were small);
- the cut scores were too high;

- failure to use a backup cut score method was *not* unprofessional;
- a developer *may* change methodology across time without this constituting an admission of error; and
- a court should not eschew an idealistic view of test validity evidence, but neither should it apply an "anything goes" approach.

Although this author does not agree with all of Judge Thompson's interpretations of the data in the case, the ruling does suggest that test developers should carry out their test construction tasks very carefully.

Two recent U.S. Supreme Court rulings relate to the requirements for subjective assessments. Basically, the rulings in both cases were that nonobjective assessments are subject to legal scrutiny under the disparate impact analysis of Title VII. In *Watson v. Fort Worth Bank and Trust* (1988), it was ruled that the Griggs standards would apply to subjective testing processes such as interviews. The court wished to prevent employers from circumventing the *Griggs* standard by replacing tests with subjective assessments. However, there was sharp dispute among the Justices on how to apply the standards. A plurality of the court said the standards should be applied in a less rigorous manner in subjective testing. In the *Wards Cove Packing Co. v. Atonio* (1989), a majority of the court agreed to less rigorous standards. Rebell (1990b) has suggested that:

> The net effect of Watson/Wards Cove might be said to constitute a broadening of Title VII's reach but also a modification of its bite. (p. 5)

Nevertheless, courts will not accept an "anything goes" approach in subjective assessments. (See the discussion in the next section of a licensure case [*Musgrove et al. v. Board of Education for the State of Georgia et al.*], which was a case involving a subjective assessment process.)

Licensure Cases

Licensure testing may involve a conflict between two rights: social and individual. The tension between societal and individual rights is both a legal and a moral issue (McDonough & Wolf, 1988). No one denies that the public has a legitimate right to have competent individuals practicing in various occupations and professions. No one denies that individuals have the right to be protected from unfair employment practices. The trade-off between the two is where the controversy lies.

As mentioned, there is debate about the applicability of the Civil Rights Acts to licensure tests. However, there is a strong constitutional basis for licensing. Reeves (1984) states that:

> The constitutionality of requirements to take and pass qualifying examinations is firmly entrenched. (p. 65)

This basis is stated in *Goldfarb v. Virginia State Bar* (1975) as follows:

> The States have a compelling interest in the practice of professions within their boundaries, and that as part of their power to protect the public health, safety, and other valid interests they have broad power to establish standards for licensing practitioners and regulating the practice of professions. (p. 792)

Although a constitutional basis is well established, licensure tests must have a rational relationship to the occupation. However, as mentioned, this is relatively easy to establish.

There are several court precedents for licensure. Most of these are for licensure to the Bar although in recent years there have been several teacher licensure cases. We begin our review of licensure cases with a very early decision on the licensure of doctors. In *Dent v. State of West Virginia* (1881), ruling in favor of the licensure requirement, the court declared, in part:

> The power of the state to provide for the general welfare of its people authorizes it to prescribe all such regulations as in its judgment will secure or tend to secure them against the consequences of ignorance and incapacity, as well as of deception and fraud....The nature and extent of the qualifications required must depend primarily upon the judgment of the state as to their necessity. If they are appropriate to the calling or profession, and attainable by reasonable study or application, no objection to their validity can be raised because of their stringency or difficulty. (1881, p. 114)

In a massive review of the literature, Eisdorfer and Tractenberg (1977) suggested that: "In the post-1937 period, the standard of review has become even more relaxed than that stated in the *Dent* case" (p. 117).

Given the thorough review by Eisdorfer and Tractenberg in their 1977 chapter, this review jumps to a more recent case: *United States v. State of North Carolina* (1975, 1977). The United States brought a Title VII complaint against North Carolina for requiring a minimum score on the National Teacher Examination (NTE). The court record revealed that at least one teacher training institute had

> graduated functional illiterates and the court acknowledged that the state should have "the right to adopt academic requirements and written achievement tests designed and validated to disclose the minimum amount of knowledge necessary to effective teaching." However, the NTE was not designed for use in assessing inservice teachers, the cut-off score chosen was not validated for job performance, and the result was a disparate impact on blacks. (Cohen, 1989, p. 239)

The court ruling was vacated in 1977 following the Supreme Court's ruling in *Washington v. Davis* regarding correlation with training programs and because a validation study was conducted for the NTE in North Carolina.

The *Tyler v. Vickery* (1975, 1976) case was a challenge against the constitutionality of the Georgia Bar Examination. The decision is important for several reasons. First, as mentioned, it rejected the view that the EEOC *Guidelines* were appropriate for a bar examination. Related to cutoff scores it was ruled that:

> While the minimum passing score of 70 has no significance standing alone, it represents the examiners' considered judgment as to "minimum competence required to practice law." (p. 1102)

The court also rejected the plaintiffs' complaint that the examinations did not cover the full range of skills needed to practice law and it held that no review procedure was necessary because there was an opportunity to retake the examination within a reasonable time.

An important early teacher licensure case was the *United States v. the State of South Carolina* (1977, 1978). In this case, it was ruled that the National Teacher Examination (NTE) could be used for *both* teacher certification (licensure) and determination of salary levels. This case followed *Washington v. Davis*, and the NTE was validated against teacher training programs and not actual job performance. It was held that the content validity study was adequate under Title VII (and constitutional) guidelines. One way this case differed from the original (prior to vacating) North Carolina case was that the state did both an extensive cutoff score study and content validation study. Cohen (1989) has concluded that:

> When teacher certification tests are professionally developed in good faith to insure teacher competency and are then validated as to content, they will be upheld by courts. The public interest in having at least minimally competent teachers seems to outweigh the disparate impact that has often resulted. (p. 242)

An Alabama teacher licensure case was an example of a prolonged, complex litigation. A Basic Professional Studies Test and 45 tests for different teaching specializations were constructed and administered by the National Evaluation Systems (NES). A class action suit was brought against the state on behalf of all African-Americans who had been (or would be) denied certification because of failure to pass the tests. After considerable discussion, a settlement was approved by the court. Subsequently, the Alabama State Board of Education wished to back out of the settlement. After much legal manipulation, the United States Court of Appeals ruled that the original agreement was enforceable. The settlement incorporated the idea of the *Golden Rule* (1980) settlement that required items with minimum racial differences to be used first in any test. (The *Golden Rule* approach to choosing items has been almost unanimously viewed by measurement professionals as one that will result in psychometrically inferior examinations.) At any rate, the Alabama case was decided on procedural grounds rather than on the merits of the proposed certification programs. Nevertheless, while the settlement issue was being debated in the courts, the case was tried on its merits, but the judge never issued a ruling. Although the *Richardson* employment case discussed earlier may provide some clues regarding how the judge might have ruled, it is possible that previous legal precedent for licensure cases may have caused the judge to rule differently in a licensure case than he would have in the employment case.

Two licensure cases with important implications for testing are the *State of Texas v. Project Principle* (1987), and *United States v. LULAC* (1986). In the *Project Principle* case, use of the Texas Examination of Current Administrators and Teachers (TECAT) was ruled constitutional. It was held that there was no impairment of a contract right because teaching certificates are licenses, not contracts; state legislatures may change licensing requirements retroactively; and that teacher testing was a rational means of achieving legitimate State objectives, hence was not fundamentally unfair. Also, it was ruled that due process was not violated because applicants had a right to retake the test prior to being decertified. The court ruled that:

> teacher testing is a rational means of achieving the legitimate state objective of ensuring that public school educators meet specified standards of competency. (1987, p. 391)

In the *LULAC* case, the use of the Pre-Professional Skills Test (PPST) was upheld. The court noted that the state had considered other alternative tests before selecting the PPST, and that a validation study had been conducted which surveyed Texas educators regarding their beliefs about whether the skills measured by the PPST were necessary for success in teacher education programs and in teaching. The court agreed with the *Washington v. Davis* (1976) decision that a test only need show a relationship to the effects of a required training program, not the eventual competence of individuals on the job. Further, as noted earlier, the court held that because applicants are permitted to retake the test, and that the passing rate for minority-group students was increasing, "the ultimate impact of the PPST on the number of minority teachers in the State has not been assessed" (*United States v. Lulac*, 1986, p. 643). With respect to the issue of due process, the court held that the legislative process gave adequate notice:

> When the legislature enacts a law, or a state agency adopts a regulation, that affects a general class of persons, all of those persons have received procedural due process by the legislative process itself and they have no right to individual attention. (*United States v. Lulac*, 1986, p. 647)

Finally, the court ruled that institutions of higher education were not required to lower standards to accommodate students who had been inadequately educated due to the state's historical dual school system.

> In administering its higher education systems...a state...has no constitutional or statutory obligation to suspend or lower valid academic standards to accommodate high school students who may be ill-prepared because of prior constitutional violations by its local and elementary school systems (*United States v. Lulac*, 1986, p. 7015).

Musgrove et al. v. Board of Education for the State of Georgia, et al. (1991) was a case involving use of the Teacher Performance Assessment Instrument (TPAI) for teacher licensure. Several points were made in that ruling that have important implications for licensure testing. One issue pertained to the rule that candidates were only allowed six attempts to pass the test. The court ruled that:

> a [sic] irrebuttable lifetime presumption of unfitness after failure to pass six "TPAI"s was arbitrary and capricious because no further education, training, experience, maturity or higher degree would enable such persons to become certified in Georgia. (*Musgrove*, 1991, p. 3).

Further, the court found that two competencies ("Interpersonal Skills" and "Helps Learners Develop Positive Self-Concepts") had indicators that were "so vague, ambiguous, indefinite, arbitrary and subjective as to fail to place a reasonable person on notice of the standards of conduct expected" (Musgrove, 1991, p. 6). This court ruling focused on a performance instrument that had been carefully constructed and heavily researched. Those who are developing performance assessment instruments for high-stakes decisions should consider this court decision very carefully.

Although the ruling not limiting the number of attempts to six is different from those to be discussed in the next paragraph, consideration should be given regarding

whether additional education should result in additional attempts being permitted. Performance standards should be defined with great care to minimize the possibility of their being considered vague, arbitrary, and subjective.

Four courts have ruled in favor of limiting the number of chances an individual may have to take an exam. In *Younger v. Colorado State Board of Law Examiners* (1980) the court ruled in favor of limiting the number of examinations to three, and in *Poats v. Givan* (1981) a rule limiting the number of times an applicant could sit for the bar exam to four was declared legal. In *Jones v. Board of Commissioners* (1984) an Alabama rule limiting the number of times an applicant could take the bar exam did not create an irrebuttable presumption of incompetence. In *Yu v. Clayton* (1986) it was ruled that an RN applicant who had failed a licensure exam six times was ineligible for another chance until after recompleting an entire course of nursing studies. These four rulings are at odds with the *Musgrove* decision cited earlier.

Several other cases are worthy of brief mention. One relates to the review of exams. In *Balaklaw v. American Board of Anesthesiology, Inc.* (1990) a plaintiff who failed brought suit requesting he be allowed to review his exam and answer sheet. The request was denied. This ruling was similar, in this respect, to the *Tyler v. Vickery* decision mentioned earlier.

Finally, in *Millet v. Hoisting Engineers' Licensing Div.* it was ruled, for an *oral* exam, that:

> Failure to keep a record of the questions and answers has been held to be a constitutional violation because this deprives the failed applicant of any chance of showing that the examination was irrational and arbitrary or that the grading was in error. (1977, 1171)

Conclusions Regarding Court Decisions

A general conclusion seems to be that if tests are constructed according to procedures advocated in the *Guidelines* and *Standards*, they should withstand legal scrutiny. For employment cases, the key issue is validity. Rossein summarizes case law as follows:

> Courts readily uphold an employment practice if the employer can show that the practice actually enables the employer to screen out unqualified or less qualified candidates. (1992, p. 11)

The issue, of course, is what kinds of, and how much, evidence is required. Content validity evidence has generally been considered sufficient. For example, in *Jones et al. v. New York City Human Resources Administration* (1975) it was stated that *no* case in that Circuit had held that criterion-related evidence was required to prove job-relatedness.

Although, the Court argued in the *Richardson* decision that it should not eschew an idealistic view of test validity nor apply an "anything goes" approach, it is clear that the decision employed standards on the idealistic side of a middle position. That can perhaps be seen most clearly by looking specifically at the cut score issue. In general, the courts have accepted judgments regarding the cut score. In *Tyler v. Vickery* (1975) the court ruled that the cut score had been validated even

though there was no empirically demonstrated evidence because the score represented the examiners' "considered judgments" as to minimum competence required. In *Guardians Association* the exam was ruled as invalid, but regarding the cut score the court stated that:

> As with rank-ordering, a criterion-related study is *not* necessarily required: the employer might establish a valid cutoff score by using a professional estimate of the requisite ability levels, or, at the very least, by analyzing the test results to locate a logical "breakpoint" in the distribution of scores. (from Byham, 1983, p. 107)

Pyburn (1984) concluded that a state may set the passing grade where it chooses because it is empowered to require high standards. He references *Schware v. Board of Bar Examiners of State of New Mexico* (1957) and *Chance v. State Bar of California* (1967). The *Dent* decision was quoted above. Although all these cases suggest that professional judgment is acceptable as a means of setting cut scores, if a judge is convinced the cut scores are too high, the ruling may be unfavorable. In the *Richardson* case discussed earlier, the court ruled that:

> the developer's procedure yielded cut scores that were so astoundingly high that they signaled, on their face, an absence of correlation to minimum competence. (1989, p. 28)

> an inference as to competence will be meaningless if the cut score, or decision point, of the test does not also reflect *what practitioners in the field deem to be a minimally competent level of performance on that test.* Again, the test developer's role in setting a cut score is *to apply professionally accepted techniques that accurately marshal the judgment of practitioners.* (1989, p. 32)

One interesting point about the above quotes is that the judge seemed to support judgmental methods. Yet, when the test developers did apply what some supported as a professionally accepted technique, the judge contended that the cut scores were "astoundingly high." Certainly the attempt by the test constructors was to marshal the judgment of practitioners. Experts for the defendants did not believe the cut scores were too high. However, experts for the plaintiffs argued that the standards were too high. The judge obviously agreed.

Rebell, in discussing three recent challenges that were settled or withdrawn, pointed out the very high pass *rate* for these tests. As he suggested:

> To the extent that fear of judicial intervention caused a lowering of otherwise valid and appropriate cut scores, increased court involvement in evaluation matters is a worrisome prospect. (1990a, p. 351)

Thus, although some judges will set very high (unrealistic?) standards for test quality, the bulk of the case law suggests most judges are reasonable in their expectations and rulings. In concluding this section, it seems appropriate to quote Pyburn:

> To date, there have been very few successful challenges to licensing examinations on the grounds that the tests were "discriminatory" or were not "rationally related" to the purpose for which they were being used. (1990, p. 14)

FUTURE DIRECTIONS

The *Guidelines* are quite out of date, but no revision is being planned; the *Standards* are somewhat dated and a revision is being planned; the 1991 Civil

Rights Act, at the time of this writing, has had little chance to impact court rulings; and the *Watson* and *Wards Cove* rulings regarding subjective assessments are too recent to have had much impact on subsequent rulings. Thus, a variety of factors may impact how one should construct licensure tests and how courts may rule on their legality. Although the future is always difficult to predict, some discussion of possible future directions seems worthwhile.

New Standards

The revision of the *Standards* is being planned and, by the time this book is published, the individuals on the committee will be appointed and specific changes for the *Standards* will likely have been proposed. No revised standards are anticipated before 1996. As was mentioned, there was not total agreement among psychometricians regarding the 1985 *Standards*. Some thought they were not "tough" enough whereas others thought they set unrealistically high standards. Whether the revised standards will be more or less rigorous regarding tests used for licensure or certification will depend, in part, upon the views of the particular individuals appointed to the committee.

Although the political/social interests and psychometric views of the individuals on the new *Standards* committee will likely have an impact on the *Standards*, just what that impact will be is unknown. What is known is that some views of the psychometric profession have changed and there is likely some general agreement on the wisdom of the changes. The 1985 *Standards* predicted some specific areas where

> new developments are particularly likely, such as gender-specific or combined-gender norms, cultural bias, computer based test interpretations, validity generalization, differential prediction, and flagging test scores for people with handicapping conditions. (AERA/APA/NCME, 1985, p. 2)

Some of these new developments have been influenced by legislation. For example, the Civil Rights Act of 1991 prohibits ethnic or gender norming for employment tests. Some of the other areas have not developed as much as was surmised when the 1985 *Standards* went to print.

In my view, the major writings likely to influence the revised *Standards* are in the area of validity. As reported earlier, there was a movement in the 1985 *Standards* to unify the notion of validity under the heading of construct validity. There has been continued writing in that area and the new *Standards* may well go further in that unifying direction than the current ones do. Whether there will be any major changes in the methodologies used to establish validity is more questionable. In my view, the methodologies available for gathering validity evidence have not, in fact, expanded much. One is still likely to use the methodologies that heretofore have been referred to as content, criterion-related, and construct validity evidences. There may, in fact, be a change in that all these methodologies are referred to as providing evidence regarding the construct validity of the measures.

In addition to wishing to call all validity construct validity, there has been some suggestion that the notion of validity should extend beyond the accuracy of inferences made from the scores to encompass the social consequences of testing

(Messick, 1989; Shepard, 1993). It is unclear at the time of this writing whether that expansion of the meaning of the word "validity" will be widely accepted by the measurement community. For example, Wiley (1991) prefers to focus on the psychological processes intended to be measured rather than the use of the tests. In general, there is some concern that broadening the concept of validity into a consideration of social concerns will cause it to lose some of its scientific meaning. Nevertheless, whether consequences of test use become a part of the connotation of "validity," the measurement community has long noted the importance of considering the costs of false positives and false negatives and the new *Standards* are almost sure to emphasize the consideration of these costs more explicitly. It is hard to imagine that the costs of false positives would be taken lightly for licensure decisions.

New Legislation

Some aspects of the *Civil Rights Act of 1991* and the *Americans with Disabilities Act* have been discussed. Because both are reasonably recent, there is little legal precedent regarding what the impact of these will be. In this author's view, there will be little impact on licensure from the *Civil Rights Act of 1991* because it relates primarily to employment testing and it basically reaffirms the business necessity requirement that was the basis for many of the previous decisions. The only two decisions that would have allowed for a lessening of the business necessity requirement were the *Watson* and the *Wards Cove* cases. There will likely be some consideration of the *Americans with Disabilities Act* in the new *Standards*. Whether or not that occurs, test constructors and test users do need to attend to the necessity of providing *appropriate* accommodations for individuals with documented disabilities.

Subjective Assessments

Although portions of the *Watson* and *Wards Cove* cases have been made impotent as precedents due to the 1991 Civil Rights Act, the act did not address the issue of subjective assessments. It is reasonable to assume that many more cases will arise where subjective assessments are being challenged. Both Rebell (1990b) and Phillips (1993) have pointed out that the testing issues in *Watson* were less complex than those posed by some of the currently proposed performance tasks.

> The question remaining is whether it is reasonable and technically feasible to apply the EEOC Guidelines to such performance (subjective) tasks. (Phillips, 1993, p. 735)

It is too soon to know how demanding the courts will be regarding the psychometric properties of subjective assessments. However, it would seem that the psychometric community would desire high quality assessments whether they be considered objective or subjective. Thus, one should not anticipate support from the psychometric community for subjective assessments that have low reliability, low validity, inadequate equating procedures, etc. (It is true that the specific operational definitions of validity and reliability may be somewhat different for subjective assessments.)

SUMMARY

The general legal setting within which employment and licensure tests are judged has been described in this chapter. Generic legal issues include the constitutional requirements (primarily of the 14th Amendment) and the statutory requirements of the Civil Rights Acts. Basically the Constitution requires equal protection and due process. The Civil Rights Acts prohibit disparate treatment and disparate impact.

A distinction was made between employment and licensure/certification testing. The purposes of these types of testing are quite different and logically should lead to different test development procedures. There is some uncertainty about whether the Civil Rights Acts and the EEOC *Guidelines* are applicable to licensure tests. This is an important issue because the Civil Rights Acts call for a business necessity requirement, which is considered harder to demonstrate than the legitimate relationship requirement that the 14th Amendment calls for.

The more relevant portions of a variety of professional standards and codes for licensure tests were summarized. Although both the AERA/APA/NCME *Standards* and the EEOC *Guidelines* are somewhat dated, they have been used extensively in previous court cases (the *Guidelines* for employment tests) and, thus, there is some legal precedent based on these standards.

Several of the more important employment and licensure court decisions were discussed. In general, it would appear that higher test development/validation standards have been set for employment decisions than for licensure decisions. The courts have accepted a variety of kinds of validity evidence and are (generally) reluctant to second-guess cut scores that have been established by obtaining the judgments of individuals in the profession/occupation in question.

Future directions with respect to legal precedents will be somewhat dependent upon the upcoming revision of the *Standards*. It is unclear what recent legislation such as the *Civil Rights Act of 1991* and the *Americans with Disabilities Act* will have on court decisions. Basically, the new Civil Rights Act reaffirms the business necessity requirement that was the basis for many previous decisions. The *Americans with Disabilities Act* may result in increased accommodations for those with claimed disabilities. The movement to more subjective based assessments coupled with the *Watson* and *Wards Cove* rulings that subjective assessments are subject to test development standards should result in some interesting court cases.

Although an agency can always be sued, and one can never predict how a judge will rule, there has been enough precedent to suggest that if one develops an exam with professional care, there should be a good chance that the test will be declared legally acceptable.

REFERENCES

Albermarle Paper Co. v. Moody, 422 U.S. 405, 431 (1975).

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

Americans with Disabilities Act, 42 U.S.C. Section 12101 *et seq.* (1990).

Balaklaw v. American Board of Anesthesiology, Inc., 562 N.Y.S. 2d 360 (Sup. 1990).

Byham, W. C. (1983). *Review of legal cases and opinions dealing with assessment centers and content validity.* Monograph IV, Pittsburgh: Development Dimensions International.

Chance v. Board of Examiners, 303 F. Supp. 203, 209 (SDNY, 1971) aff'd, 458 F. 2d 1167 (2nd Cir., 1972).

Chance v. State Bar of California, 386 F. 2d 962, 964 (9th Cir., 1967).

Civil Rights Act of 1991. (1991). Washington, DC: The Bureau of National Affairs, Inc.

Cohen, J. H. (1989). Legal challenges to testing for teacher certification: History, impact and future trends. *Journal of Law and Education, 18*(2), 229-265.

Dent v. State of West Virginia, 129 U.S. 114, 122 (1881).

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.

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.

Freeman, L. D., Hess, R., III, & Kasik, M. M. (1985, March). *Testing teachers and the law.* Presentation made at the American Educational Research Association annual meeting, Chicago, IL.

Golden Rule Insurance Co. et al. v. Mathias et al. 86 Ill.App 3d 323, 326, 41Ill. Dec. 888, 891, 408 N.E. 2d 310, 313 (1980).

Goldfarb v. Virginia State Bar, 421 U.S. 773 (1975).

Griggs v. Duke Power Company, 292 F. Supp. 243 (MD NC, 1968), 420 F.2nd 1225 (4th Cir., 1970) and 401 (U.S. 424, 1971).

Guardians Association of New York City v. Civil Service Commission, 431F. Supp. 526 (Southern District of New York, 1977); U.S. District Court of Appeals, Second Circuit (No. 849), July 31, 1980.

Hazelwood School District v. United States, 97 S.Ct. 2736 (1977).

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

Jones v. Board of Commissioners, 737 F. 2d (11th Cir. 1984).

Jones et al. v. New York City Human Resources Administration, U.S. District Court, Southern District of New York, January 10, 1975; 73 (1) 3815; U.S. Court of Appeals, Second Circuit (New York), January 26, 1976.

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

Kuehn, P. A., Stallings, W. M., & Holland, C. L. (1990). Court-defined job analysis requirements for validation of teacher certification tests. *Educational Measurement: Issues and Practice*, 9(4), 21-24.

Linn, R. L. (1984, April). *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.

McDonough, M. W., Jr., & Wolf, W. C., Jr. (1988). Court actions which helped define the direction of the competency-based testing movement. *Journal of Research and Development in Education, 21*(3), 37-43.

Mehrens, W. A. (1990). Assessing the quality of teacher assessment tests. In J. V. Mitchell, Jr., S. L. Wise, & B. S. Plake (Eds.), *Assessment of teaching: Purposes, practices, and implications for the profession* (pp. 77-136). Hillsdale, NJ: Lawrence Elbaum Associates.

Meier, P., Sacks, J., & Zabell, S. L. (1984). What happened in Hazelwood: Statistics, employment discrimination, and the 80% rule. *American Bar Foundation Research Journal, 1*, 39-164.

Messick, S. (1989). Validity, In R.L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). New York: American Council on Education and Macmillan.

Millet v. Hoisting Engineers' Licensing Div., 377 A.2d 229 (R.I. 1977).

Millman, J., Mehrens, W. A., & Sackett, P. R. (1993, May). *An Evaluation of the New York State Bar Examination.* A study commissioned by the New York State Court of Appeals. Albany, NY.

Musgrove et al. v. Board of Education for the State of Georgia et al. (Feb., 1991). Civil Action File No. D-62016.

Personnel Administrator v. Feeney, 442 U.S. 256, 279, 99 S.Ct. 2282, 2296, 60 L.Ed.2d 870 (1979).

Phillips, S. E. (1991). Extending teacher licensure testing: Have the courts applied the wrong validity standard? *Thomas M. Cooley Law Review, 8*(3), 513-550.

Phillips, S. E. (1993, March 11). Legal issues in performance assessment. *Education Law Reporter, 79*, 709-738.

Poats v. Givan, 651 F. 2d 495 (7th Cir. 1981).

Pyburn, K. M., Jr. (1984, April). *Legal challenges to licensing examinations.* Paper presented at the AERA-NCME Annual Meeting, New Orleans, LA.

Pyburn, K. M., Jr. (1990). Legal challenges to licensing examinations. *Educational Measurement: Issues and Practice, 9*(4), 5-6, 14.

Rebell, M. A. (1986). *Pre-Trial Memorandum of Law on behalf of Amicus Curiae National Evaluation Systems, Inc.* 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.

Rebell, M. A. (1990a). Legal issues concerning teacher evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers* (pp. 337-355). Newbury Park, CA: Sage.

Rebell, M. A. (1990b). Legal aspects of subjective assessments of teacher competency. In NES (Eds.), *The assessment of teaching: Selected topics* (pp. 1-10). Amherst, MA: National Evaluation Systems.

Reeves, R. (1984). *The law of professional licensing and certification.* Charlotte, NC: Publications for Professionals.

Richardson v. Lamar County Board of Education, et al., Civil Action No.87-T-568-N (1989); U.S. Court of Appeals, Eleventh Circuit, Nos.90-7002, 90-7336 (July 17, 1991).

Rossein, M. (Feb., 1992). *Disparate Impact Theory After the Civil Rights Act of 1991: Restoring the Job Performance Standard.* 429 PLI/Lit 155 PLI Order No. H4-5127.

Schware v. Board of Bar Examiners of State of New Mexico, 353 U.S. 232, 238-239 (1957).

Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of research in education: 19* (pp. 405-450). Washington, DC: American Educational Research Association.

Shimberg, B. (1990). Social considerations in the validation of licensing and certification exams. *Educational Measurement: Issues and Practice*, *9*(4), 11-14.

Smith, I. L., & Hambleton, R. K. (1990). Content validity studies of licensing examinations. *Educational Measurement: Issues and Practice*, *9*(4), 7-10.

State of Texas v. Project Principle, Inc., 724 S.W. 2d 387, 391 (Tex. 1987).

Tyler v. Vickery, 517 F.2d 1089 (5th Cir. 1975), cert. denied, 426 U.S. 940 (1976).

United States v. LULAC, 793 F.2d 636, 640 (5th Cir. 1986).

United States v. North Carolina, 400 F. Supp. 343 (E.D.N.C. 1975), *vacated*, 425 F. Supp. 789 (E.D.N.C. 1977).

United States v. State of South Carolina, 445 F. Supp. 1094 (DSC), (1977), *aff'd.* 434 U.S. 1026 (1978).

Village of Arlington Heights v. Metropolitan Housing Development Corp., 429 U.S. 252, (1977).

Wards Cove Packing Co. v. Atonio. 490 U.S., 109 S. Ct. 2115 (1989).

Washington v. Davis, 348 F. Supp. 15 (D.C., 1972), 512 F. 2d 956(D.C. Cir., 1975) and 426 U.S. 229, 250. (1976).

Watson v. Fort Worth Bank and Trust. 487 U.S. 977 (1988).

Wiley, D. E. (1991). Test validity and invalidity reconsidered. In R. E. Snow & D. E. Wiley (Eds.), *Improving inquiry in social science: A volume in honor of Lee J. Cronbach* (pp. 75-107). Hillsdale, NJ: Erlbaum.

Woodward v. Virginia Board of Bar Examiners, 420 F. Supp. 211, 18 FEP 836 838 (E.D. Va 1976), *aff'd per curiam*, 598 F.2d 1345 (4th Cir. 1979).

Younger v. Colorado State Board of Law Examiners, 625 F.2d 372 (10th Cir. 1980).

Yu v. Clayton, 497 N.E. 2d 1278 (Ill. App. 1 Dist. 1986).