

1-1-1985

# Testing teachers : legal and psychometric considerations 1965 to 1985.

Matthew W. McDonough  
*University of Massachusetts Amherst*

Follow this and additional works at: [https://scholarworks.umass.edu/dissertations\\_1](https://scholarworks.umass.edu/dissertations_1)

---

## Recommended Citation

McDonough, Matthew W., "Testing teachers : legal and psychometric considerations 1965 to 1985." (1985). *Doctoral Dissertations 1896 - February 2014*. 4022.  
[https://scholarworks.umass.edu/dissertations\\_1/4022](https://scholarworks.umass.edu/dissertations_1/4022)

This Open Access Dissertation is brought to you for free and open access by ScholarWorks@UMass Amherst. It has been accepted for inclusion in Doctoral Dissertations 1896 - February 2014 by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact [scholarworks@library.umass.edu](mailto:scholarworks@library.umass.edu).



TESTING TEACHERS: LEGAL AND PSYCHOMETRIC  
CONSIDERATIONS, 1965 TO 1985

A Dissertation Presented

By

MATTHEW W. McDONOUGH, JR.

Submitted to the Graduate School of the  
University of Massachusetts in partial fulfillment  
of the requirements for the degree of

DOCTOR OF EDUCATION

MAY 1985

Education

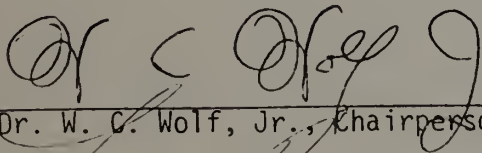
TESTING TEACHERS: LEGAL AND PSYCHOMETRIC  
CONSIDERATIONS, 1965 TO 1985

A Dissertation Presented

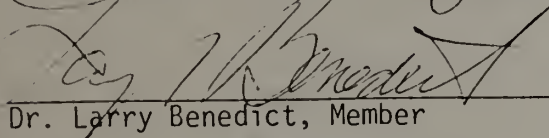
By

MATTHEW W. MCDONOUGH, JR.

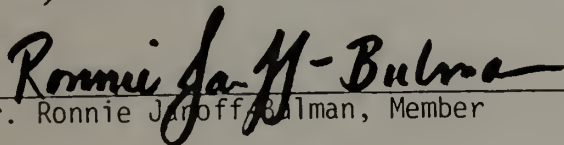
Approved as to style and content by:



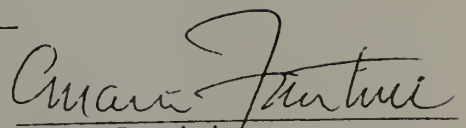
Dr. W. C. Wolf, Jr., Chairperson of Committee



Dr. Larry Benedict, Member



Dr. Ronnie Jaroff-Bulman, Member



Mario Fantini  
Dean, School of Education

Matthew W. McDonough, Jr.



All Rights Reserved

## ACKNOWLEDGEMENT

I wish to acknowledge a sincere debt of thanks, first to the staff of National Evaluation Systems, Inc., and the staffs of the Alabama, Georgia, and Oklahoma Departments of Education for providing me with the opportunities that are the foundation of this dissertation. I also extend my appreciation to Drs. Benedict, Janoff-Bulman, and Wolf, the members of my committee, whose encouragement and expectations kept me on task. And finally, to Janice Frazier, who not only provided all of the typing and much-needed editorial advice, but who also gave me the moral support that spelled the difference between a burden and an enjoyable and enlightening experience.

ABSTRACT

Testing Teachers: Legal and Psychometric  
Considerations, 1965 to 1985

May 1985

Matthew W. McDonough, Jr., B.A., University of Massachusetts  
M.A., Springfield College, Ed.D., University of Massachusetts

Directed by: Professor W. C. Wolf, Jr.

This study examines the legal and regulatory decisions made during the last two decades that have influenced the teacher certification testing movement. Chapter I presents an overview of the study and an outline of the major legal, psychometric, and programmatic pressures that have influenced the direction of teacher certification testing. A synopsis of significant litigation relevant to employment testing in general and teacher testing in particular is presented in Chapter II. Chapter III concerns the technical challenges that psychometrists have faced in their effort to address problems identified by the courts. Chapter IV provides an example of one state's (Oklahoma) and one test developer's (National Evaluation Systems, Inc.) attempt to develop a psychometrically-sound, legally-defensible teacher certification test. Finally, Chapter V presents summaries of interviews with leading experts in the legal, psychometric, and programmatic areas of teacher testing. The experts offered their insights into the future of the teacher certification movement.

## TABLE OF CONTENTS

ACKNOWLEDGEMENT . . . . .	iii
LIST OF FIGURES . . . . .	vi
Chapter	
I. LEGAL CHALLENGES TO THE TEACHER CERTIFICATION PROCESS . . . . .	1
Statement of the Problem . . . . .	1
The need to test teachers . . . . .	1
The legal/regulatory environment of teacher testing . . . . .	4
Professional standards for teacher testing . . . . .	8
Purposes of the Study . . . . .	11
Significance of the Study . . . . .	15
Definitions . . . . .	15
Limitations and Delimitations . . . . .	17
Synopsis of Research Design . . . . .	19
II. REGULATORY AND JUDICIAL PRECEDENTS . . . . .	21
The Major Issues . . . . .	21
Chronology . . . . .	23
The Cases . . . . .	24
Moody v. Albermarle Paper Company . . . . .	28
Baker v. Columbus Municipal Separate School District . . . . .	29
Chance v. Board of Examiners . . . . .	31
Washington v. Davis . . . . .	32
U.S. v. State of North Carolina and U.S. v. State of South Carolina . . . . .	34
Connecticut v. Teal . . . . .	35
III. TECHNICAL CHALLENGES TO THE TEST DEVELOPER . . . . .	38
Introduction . . . . .	38
The Technical Challenges . . . . .	39
Validation procedures . . . . .	39
Reliability . . . . .	51
Error compensation . . . . .	53
Job analysis . . . . .	56
IV. A TEST DEVELOPMENT EXAMPLE . . . . .	60
Introduction . . . . .	60
Overview of the Oklahoma program . . . . .	61
The Development Process . . . . .	65
Formation of committees . . . . .	65



Content outlines . . . . .	66
Objectives . . . . .	68
Job analysis . . . . .	70
Selection of final objectives . . . . .	71
Item development . . . . .	74
Field test . . . . .	76
Content validation . . . . .	77
Standard setting . . . . .	78
Test form development . . . . .	79
First administration . . . . .	79
Standard error of measurement . . . . .	80
Monitoring activities . . . . .	82
V. THE FUTURE OF TEACHER CERTIFICATION TESTING . . . . .	84
Introduction . . . . .	84
Legal Survey . . . . .	88
Psychometric Survey . . . . .	92
State Department Survey . . . . .	97
Summary . . . . .	102
SELECTED BIBLIOGRAPHY . . . . .	109

LIST OF FIGURES

2.1. Timeline for major legislative, regulatory, professional, and judicial events between 1965 and 1985 . . . . .	25
3.1. Type I and type II errors in the identification of masters and non-masters . . . . .	54
4.1. Test development process . . . . .	67
4.2. Job analysis results for a hypothetical field, showing the distribution and classification of objectives . . . . .	72
4.3. Passing score determination based on hypothetical 20-item data bank . . . . .	81

C H A P T E R    I  
LEGAL CHALLENGES TO THE TEACHER CERTIFICATION PROCESS

Statement of the Problem

The need to test teachers

The last decade has witnessed a growing demand for accountability in the American educational system. Some criticism has been leveled at educational philosophies that became popular in the 1960s. Some critics point to the changing priorities in the federal government's attitude toward education. Today, a large portion of the responsibility for falling SAT scores, functionally illiterate graduating seniors, and the perceived deficits of young people entering the work force has fallen squarely on the shoulders of the front line soldier -- the classroom teacher. One need not look far to see the extent and intensity of this concern about the poor quality of American teachers. The recently published report A Nation at Risk: The Imperative for Educational Reform (National Commission on Excellence in Education, 1983) not only castigates the present cadre of educators but also predicts that the young professionals now graduating from teacher education institutions are less prepared and less motivated than the senior faculty they will replace.

In recent years, the New York Times has focused two major news series on the crisis in teacher competency (Montgomery, 1979; Maeroff, 1983). Popular periodicals such as Esquire (1983),

Harper's (1983), and Newsweek (1984) report that the brighter and more competent college students are shunning careers in education and that the best veteran teachers, when asked, report that if they had it to do over again, they would not choose careers in education. The implication here is that the individuals currently being certified to teach are not adequately prepared to do their jobs and that all evidence points to a worsening of the situation in the future. The current interest in teacher certification testing is a direct response to the growing lack of confidence in teacher competence.

Early in the nineteenth century, the sole credential required of teachers of public school children was basic proficiency in reading, writing, and arithmetic. With the advent of mass compulsory education later in the century came the state's interests in extending these criteria to include proficiency in professional techniques and specific subject-matter knowledge. These three aspects -- basic skills, competence in teaching techniques, and knowledge of subject matter to be taught -- have continued through to the present as the mainstays of teacher assessment systems.

This characterization seems to suggest and underscore a considerable consensus about some important aspects of teacher evaluation -- although some would rather interpret this status as a reflection of the slow growth in our understanding of the elements of effective teaching and the way to test for their presence. Naysayers notwithstanding, the last decade has been marked by dramatic change in approaches to credentialing public school teachers. The nature of the change in credentialing practice is evidenced by the significant nationwide

increase in efforts to re-examine and modify those state-level programs charged with the responsibility of licensing teachers.

Prior to the late 1960s most states credentialed prospective teachers on the basis of successful completion of a teacher education program of study. Only a few states went so far as to require accreditation or "approval" of such programs; and even fewer states took the additional measure of requiring entrants into the teaching field to pass a nationally standardized, norm-referenced test. Such state policies had been stable for a considerable length of time, which suggested a prevailing opinion that certification programs were fulfilling their purpose. From the lack of controversy, one might conclude that most groups and individuals concerned with public education were satisfied that these programs were adequate to ensure that unqualified individuals were excluded from teaching and that all qualified applicants had fair and unbiased access to the profession.

The decade of the 1970s stands in marked contrast to the earlier complacency. During this time, teacher certification programs were taken to task by a variety of interest groups concerned with the quality of teaching in the nation's schools, and state departments of education faced strong and often contradictory demands for change. As a result, teacher certification programs were subjected to considerable scrutiny and underwent extensive changes.

### The legal/regulatory environment of teacher testing

As public pressure was being brought to bear on teacher certification programs, a number of legal and regulatory precedents were being set that influenced the direction of the movement. These were an outgrowth of Title VII of the Civil Rights Act and the Equal Employment Opportunity Commission (EEOC) Guidelines on Employee Selection Procedures. Additionally, there was the influence exerted by development of the 1974 version of the Standards for Educational and Psychological Tests (APA, AERA, NCME, 1974). The promulgation of these regulations and standards reflected increasing legislative, judicial, and professional concern with fair employment practices both in and out of education.

Stated simply, Title VII of the Civil Rights Act of 1964 outlawed employment discrimination on the basis of sex, race, color, religion, or national origin, and established the EEOC with the power to enforce the stipulations of the law. The 1970 EEOC Guidelines, a revision of the first version published in 1966, included a set of stipulations founded on the premise that standardization and proper validation in employee selection procedures would build a foundation for the nondiscriminatory personnel practices required by Title VII. These stipulations (EEOC, 1970) included the following:

- (a) Empirical data should be made available to establish the predictive validity of a test, that is the significant correlation of test performance with job-relevant work behaviors; such data must be collected

according to generally accepted procedures for establishing criterion-related validity.

- (b) Where predictive validity is not feasible, evidence of content validity may suffice as long as appropriate information relating test content to job requirements is supplied.
- (c) Where validity cannot otherwise be established, evidence of a test's validity can be claimed on the basis of validation in other organizations as long as the jobs are shown to be comparable.
- (d) Differential failure rates for members of groups protected by Title VII constitute discrimination unless the test has proven valid (as defined above) and alternative procedures for selection are not available.
- (e) Differential failure rates must have a job-relevant basis and, where possible, data on such rates must be reported separately for minority and non-minority groups.

As a result of Title VII and the EEOC Guidelines, many concepts that had previously been the purview of psychometricians took on important legal ramifications. In the first major challenge to employment tests (Griggs v. Duke Power Company<sup>1</sup>), the Supreme Court unanimously interpreted Title VII as prohibiting "not only overt discrimination but also practices that are fair in form, but discriminatory in operation" (p. 431).

<sup>1</sup>Griggs v. Duke Power Co., 401 U.S. 424 (1971).

This decision decreed that absence of intent to discriminate was insufficient to justify the use of a test that had a disproportionate impact on protected minorities; even the employer with the best of intentions bore the responsibility of demonstrating "that any given requirement . . . [bears] a manifest relationship to the employment in question" (p. 431). The Court further commented that the tenets of the Guidelines were "entitled to great deference" (p. 434) because they were drafted by the enforcing agency for Title VII. It was in this way that the concepts of "job relatedness" came to be incorporated into the law of employment testing (Bersoff, 1981) and virtually came to have the effect of law (Rebell, 1976).

Two other early cases are worthy of note. In Chance v. Board of Examiners,<sup>2</sup> the New York licensing exams for principals and other administrators were declared invalid for lack of job relevance. Later, in Albemarle Paper Company v. Moody,<sup>3</sup> the Court invoked EEOC and, in effect, established criteria to be used in proving whether employers' tests were job related.

Most significantly for teacher certification programs was passage of a 1972 amendment (Public Law 92-261) to the Civil Rights Act that struck out the exemption for educational personnel in public institutions, extending the provisions of EEOC beyond private industry to state and local government agencies. Prior to the amendment, court challenges against public employers (e.g., Chance v. Board of Examiners) were initially brought on equal protection grounds under the Fourteenth Amendment, which

<sup>2</sup>Chance v. Board of Examiners, 330 F.Supp. 203 (S.D.N.Y. 1971).

<sup>3</sup>Albemarle Paper Company v. Moody, 422 U.S. 405 (1975).



unanimously interpreted Title VII as prohibiting "not only overt discrimination but also practices that are fair in form, but discriminatory in operation" (p. 431). This decision decreed that absence of intent to discriminate was insufficient to justify the use of a test that had a disproportionate impact on protected minorities; even the employer with the best of intentions bore the responsibility of demonstrating "that any given requirement . . . [bears] a manifest relationship to the employment in question" (p. 431). The Court further commented that the tenets of the Guidelines were "entitled to great deference" (p. 434) because they were drafted by the enforcing agency for Title VII. It was in this way that the concepts of "job relatedness" came to be incorporated into the law of employment testing (Bersoff, 1981) and virtually came to have the effect of law (Rebell, 1976).

Two other early cases are worthy of note. In Chance v. Board of Examiners (1972), the New York licensing exams for principals and other administrators were declared invalid for lack of job relevance. Later, in Albemarle Paper Company v. Moody (1975), the Court invoked EEOC and, in effect, established criteria to be used in proving whether employers' tests were job related. Specifically, the Court made reference to the importance of analyzing "the attributes of, or the particular skills needed in" (p. 432) a given job as a basis for creating a job-relevant test.

Most significantly for teacher certification programs was passage of a 1972 amendment (Public Law 92-261) to the Civil Rights Act that

required only that employers demonstrate a rational basis for use of a test. Arguments only indirectly cited, but amassed consensual support for, EEOC Guidelines that were not technically binding at the time (Rebell, 1976). The 1972 Amendment paved the way for later litigation (e.g., United States v. State of North Carolina<sup>4</sup>), which successfully challenged the NTE as a teacher selection test. For an excellent review of these cases and an overview of the law and teacher certification, see Licensing and Accreditation in Education: The Law and the State Interest (Levitov, 1976).

Throughout the decade, the concepts contained in the 1970 EEOC Guidelines were refined through the process of litigation and resulting Court opinion. Concurrently, various federal agencies were debating related issues, a debate that culminated in publication of the 1978 Uniform Guidelines (EEOC, CSC, Department of Labor, and Department of Justice, 1978), a document that contained "specific statements in most sections, in contrast to the more general statements of the 1970 Guidelines" (Novick, 1981, p. 1040). The intent was made clear: that a test must be a representative measure of the actual domain of skills used on the job and must be validated for its intended purpose.

<sup>4</sup>U.S. v. State of North Carolina, Civil No. 4476 (E.D.N.C. 1975).

### Professional standards for teacher testing

A discussion of the legal and regulatory environment affecting teacher certification testing cannot exclude the process whereby professionals and practitioners regulate themselves. An example of this self-regulation is reflected in the publication of the Standards for Educational and Psychological Tests (APA, AERA, NCME, 1974). Unlike earlier documents of its kind that stressed the obligations of test producers, the 1974 Standards addressed competency in testing practice and test use (Novick, 1981). Novick (1981) presents an excellent review of the evolution in professional standards over the last three-quarters of a century, but most revealing is his comment that this first document on test use "might not have happened, had it not been for the emergence of the social questions to which the EEOC Guidelines clearly responded, and the concomitant civil rights pressure of numerous advocacy groups" (p. 1043).

The Standards display many similarities to the EEOC Guidelines and, in fact, both the 1974 document and its 1966 precursor were cited in numerous court cases (e.g., Albemarle) to bolster the credibility and importance of the Guidelines themselves (Bersoff, 1981). Beyond the emphasis on validation strategies, however, the Standards stressed the requirement to investigate potential bias in the measures and to report results for separate sub-samples (i.e., minority groups). Further, the Standards specified that any pass-fail scores used should be accompanied by "a rationale, justification, or explanation" (p. 66) for their adoption. It was provisions such as these that were taken seri-

ously by the designers and implementers of the newer teacher certification program.

Taken together, Title VII, the EEOC Guidelines, resulting court challenges, and the Standards can be seen as catalysts and guides to the restructuring of teacher certification programs. Their impact is evidenced in several aspects of these programs:

- (a) Because it has not been feasible to conduct predictive validity studies (based primarily on difficulties in obtaining reliable and valid measures of the criterion), the response has been to more fully incorporate other validation efforts. Courts have paid increased attention to the validity of certification tests, and their focus has been almost exclusively on content validity. Also, content validity has been shown to be more appropriate to tests of content knowledge than to tests of traditional skills and abilities.
- (b) The focus on content validity has much expanded involvement of incumbent teachers and subject-matter specialists in the test development process, through both committee review work and participation in full-scale job analyses. This emphasis on job analysis is a direct response to the regulatory agencies' guidelines and the court decisions surrounding job relatedness.
- (c) There is increased awareness of the potential for differential impact, with expanded efforts to include

diverse interest groups in the test development process and to report test results separately for relevant minority groups. Where the courts have found adverse impact to exist in a job-related testing program, it has become the responsibility of the employer to prove that 1) the test is valid, 2) alternative procedures for selection are not available, and 3) the differential failure rates have a job-relevant basis.

These trends reflect the significant impact of the legal/regulatory environment on the design of teacher certification programs.

Those individuals who have the responsibility for improving teacher preparation and certification now face a number of problems:

- (1) The pressure to improve public education, exemplified by the report of the National Commission on Excellence in Education (NCEE, 1983), is leading program designers to increase the demands on those who want to teach.
- (2) The regulatory agencies and the courts are making it clear that the performance demands of competency tests have legal limits, particularly in the area of fair employment.
- (3) The area is interdisciplinary. Most program designers do not have the legal background necessary to perceive many of the consequences of a particular decision, and as a result some altogether avoid legitimate avenues of pursuit while others suddenly find themselves in a

legally indefensible situation.

- (4) No comprehensive interdisciplinary analysis or review of the various fields has been conducted. The most current thinking on recent legal decisions, for example, is not available to practitioners who are trying to identify the most appropriate standard-setting procedures or job analysis methodology.

#### Purposes of the Study

The purpose of this study is to examine changes in testing for certification -- and in particular the certification of teachers -- in light of regulatory (e.g., agency guidelines and decisions) and court (e.g., threat of litigation, direct court decisions) actions. Three aspects of this proposal are developed. First, the study examines the major certification issues that lead to litigation, including test validity, bias, and adverse impact. Second, the study also examines in greater detail a set of court cases that directly affect the certification of public school teachers. And third, the study presents the views and interpretations of various experts from each of the related fields that impact upon testing for certification.

At the present time, no two states wrestling with the myriad of legal and regulatory issues surrounding the development of a teacher certification program have chosen exactly the same approach to program development. There has been substantial activity in all areas of educa-

tion. This spate of activity has taken a variety of forms, including the appointment of study panels, the commissioning of position papers, the hosting of conferences, and the review of concrete proposals. These activities at a minimum suggest an interest in re-analyzing teacher certification requirements and, in a significant number of cases, this interest has been followed by action. A number of states have made significant modifications to their existing certification programs; others have chosen to design totally new programs to replace existing ones. Changes have been variously brought to bear on the policies and practices of four phases of teacher certification programs: those effective 1) upon admission to teacher training programs, 2) upon completion of such programs (initial certification), 3) during the first year of incumbency in a teaching position, and 4) during later incumbency (certification renewal).

One major form of revision has been the elimination of the common policy of automatically granting certification to a graduate of any teacher education program. During the period from 1970 to 1975, 26 states revised such policies and implemented systems of "approving" teacher education programs (Pittman, 1975). By far the most dramatic action (or at least the most publicly visible one), however, was to require that graduates of teacher education programs pass a state-sponsored test to obtain a license to teach. Between 1977 and 1984, 18 states enacted legislation or state board of education policies that either initiated or modified tests whose purpose was state licensing of teachers. And last fall, Arkansas became the first state to require

incumbent teachers to pass a competency test or be fired.

Recent changes in testing practices appear in two areas: 1) the testing of prospective teacher education program entrants, and 2) the testing of teacher education program graduates as eligible and prospective license holders. An example of the former is Alabama's English Language Proficiency Test, which assesses basic skills in reading, writing, language skills, and listening. It is the installation of tests such as this one that reveals a heightened emphasis on "the basics" in the screening of prospective teachers. This trend is mirrored in end-of-program testing. An increasing number of states are including a basic skills test as one component of initial certification requirements; Florida's new program is a prime example.

The courts have played a significant role as states attempt to institute many of the changes described above. In 1971, in Chance v. Board of Examiners, the court held that the City of New York's tests for supervisory positions were not sufficiently job-related to justify the evidenced adverse impact. The court issued an injunction restraining the Board of Examiners from giving tests in the future and from promulgating eligibility lists based on the examinations.

In 1975, in United States v. State of North Carolina, the court determined that the National Teacher Examinations (NTE) did not measure teaching skills, although they did measure the content of the academic preparation of teachers. The court further determined that the established cutoff score of 950 was arbitrary. The court held that the state could reinstate a written test cutoff score if that score was first validated.



In 1976, in Georgia Association of Educators, Inc. v. Jack P. Nix,<sup>5</sup> the court held that the Georgia State Board of Education's use of an NTE cutoff score of 1225 was arbitrary and enjoined the state from requiring the attainment of any minimum score on the NTE as a condition for obtaining a six-year certificate.

In 1977, in United States v. State of South Carolina,<sup>6</sup> the court found that although the NTE did have an adverse impact on minorities, the evidence was sufficient to establish the validity of the examinations as an appropriate measure of minimum teacher competence. In a subsequent research study, eight NTE examinations (Trades and Industries, Distributive Education, German, Latin, Earth Science, Psychology, Speech and Drama, and Health) were deemed to be invalid measures of the South Carolina teaching positions.

In 1982, suit was filed in federal court in the State of Alabama alleging that the teacher testing program was having an adverse impact on certain minorities and that the tests were biased and lacking in validity. As of September 1984, plaintiff and the state were attempting to agree on an out-of-court settlement.

The teacher testing movement and the litigation can be expected to continue, at least for the foreseeable future. In light of this, a clear understanding of the present issues and future directions would seem critical for those responsible for the development of teacher certification programs. The proposed study is designed to meet that need.

<sup>5</sup>Georgia Association v. Nix, 407 F.Supp. 1102 (N.D.GA. 1976).

<sup>6</sup>U.S. v. South Carolina, 445 F.Supp. 1094, 1110 (D.S.C. 1977).

### Significance of the Study

The significance of the study lies in the fact that the public pressure for educational accountability and the demands for equal opportunity and fair employment practices are both continuing to mount. As state departments of education, officials, test development practitioners, and educational researchers continue to try to respond to these often opposing pressures, litigation and regulatory revision will continue to be the avenues for definition.

A clear understanding of the existing precedents and the suggested direction of change will be needed if teacher certification testing is to continue on its present course.

### Definitions

A number of the terms used in this study are critical to an understanding of the questions being raised. Many of the terms have popular definitions that are sometimes at variance with the technical or literal definitions. Some are still being defined through the continuing process of litigation. In order to aid the reader, the most important terms and concepts are defined below. A fuller explanation of the application of these terms is provided in the various sections of this study.

For the purposes of the study, the following definitions will be used:

Adverse impact refers to the situation that arises when selection procedures act to disqualify a disproportionately high number of minority or female applicants.

Certification grants the use of a title (e.g., "teacher") to an individual who has met a predetermined set of standards or qualifications set by a credentialing agency (Shimburg, 1981).

Job-relatedness, as the term is used here, refers to the degree to which an instrument or procedure used for job selection or for certification actually measures the knowledge, skills, and abilities necessary to do the job.

Licensure is the "process by which an agency of the government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency required to ensure that the public health, safety, and welfare will be reasonably well protected" (U.S. Department of Health, Education, and Welfare, 1977, p. 4). An individual without a teaching license from a particular state is legally barred from the practice of public school teaching in that state.

Professional standards refers to standards and guidelines established by professional organizations concerned with the use of tests, such as the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. The prime document in this field is the Standards for Educational and Psychological Tests.

Regulatory agency will be used to refer to any and all agencies of

government empowered to set guidelines and standards relevant to fair hiring practices. This may include the Equal Employment Opportunity Commission, the U.S. Department of Labor, and the U.S. Department of Education.

Validity refers to how an instrument measures what it purports to measure. The APA Principles for the Validation and Use of Personnel Selections Procedures (APA Principles, 1980; p. 2) defines validity operationally as "the degree to which inferences from scores on tests or assessments are justified or supported by evidence." There are three approaches to the question of validity that are traditionally employed in educational or psychological measurements: construct validity, criterion-related validity, and content validity.

### Limitations and Delimitations

Testing, job selection practices including certification and licensure, and the legal issues surrounding equal opportunity and fair employment practices are each broad areas with substantial fields of literature. This study is limited to a relatively narrow intersection of these three areas.

The study concerns all federal litigation around the use of tests for the certification of public school teachers between January 1, 1965, and December 31, 1983. It also examines state litigation that has a direct bearing on policies for other states.

In order to understand the background and implications of teacher certification testing litigation, it is also necessary to examine some litigation in the general area of employment selection testing. This investigation, however, is limited to cases with direct relevance to issues of concern in the more specific teacher certification testing cases.

The review and analysis of regulatory guidelines and professional standards is also limited to those materials and those issues that directly influence the proceedings and decisions in teacher certification testing cases or that influence policy direction in state department of education offices of teacher certification.

The focus of interest in each of the case studies is the legal precedent and its implications for educational policy-making. The focus is not on the motivations or the culpability or innocence of the parties to the litigation. The study does not concern itself with "who did what to whom, and why," but rather with the implications of the findings for future policies and future litigation.

Finally, the study limits itself to those areas of testing and psychometrics that are relevant to the litigation and policy decisions. Those areas include discussions of validity, reliability, standard-setting practices, and test construction and administration procedures. Each of these areas is reviewed in sufficient depth only to permit a thorough analysis of the issues considered in the litigation or policy analysis components of the study.

### Synopsis of the Research Design

This study examines the legal and regulatory decisions made during the last two decades that have influenced the teacher certification testing movement. Conversely, the study explores elements and trends in the teacher certification testing movement that have served or are likely to serve as the bases for legal challenges to a state's procedures for deciding who is qualified to teach in the public schools.

The research for this study is based on these sources:

- (1) A three-year period (1980-83) of employment with National Evaluation Systems, Inc., during which time the author assisted in the development and administration of criterion-referenced teacher certification tests for Alabama, Georgia, Oklahoma, and South Carolina.
- (2) A thorough review of the literature (1965-1984) in the fields of psychometrics, law, and public administration.
- (3) Interviews with selected test developers, state administrators, and legal experts concerned with the development and implementation of teacher certification testing programs.

The information gathered from these sources is presented in the remaining chapters of this study. Chapter Two presents a history of the teacher certification movement, including a review of the major legal and regulatory decisions that have shaped the present course. Chapter Three offers a detailed explanation of certain technical issues

that today's practitioners are attempting to address. Many of these, such as validity and job-relatedness, stem from legal and regulatory decisions documented in Chapter Two. Chapter Four presents one model for the development of a valid and legally-defensible teacher certification test, as implemented by National Evaluation Systems, Inc. Chapter Five offers an assessment of the present legal and psychometric issues facing the users of teacher certification tests. This chapter is based in part on interviews with experts who are currently trying to resolve these issues.

C H A P T E R    I I  
REGULATORY AND JUDICIAL PRECEDENTS

The Major Issues

As this dissertation is being completed, the Alabama State Department of Education is preparing to meet a legal challenge to its three-year-old teacher certification testing program. On December 15, 1981, approximately six months after the first administration of the new tests, three minority applicants who had failed the tests filed a class action suit against the Alabama State Board of Education.<sup>1</sup> The plaintiffs charged:

- (a) That the defendants' requirement of Teacher Certification Tests was arbitrary and capriciously devised, promulgated, and implemented;
- (b) That said tests covered materials much of which was not taught students in Alabama's colleges and universities, particularly the predominantly black state and private colleges and universities in Alabama;
- (c) That the class was not given adequate notice to prepare for the administration of said tests;
- (d) That the tests have not been properly validated;
- (e) That said tests have had an adverse and disproportionate

---

<sup>1</sup>Allen v. Alabama State Board of Education, (Trial pending).



- impact on black college and university students seeking regular certification as teachers from defendants;
- (f) That the tests penalize the class members for having to acquire all or part of their formal education in segregated public schools;
  - (g) That said tests systematically exclude black graduates from obtaining regular certification needed to enter the teaching profession in this state;
  - (h) That the Teacher Certification Tests perpetuates the vestiges of racial discrimination which had their genesis in the dual system of higher education in this state; and
  - (i) That there is no demonstrative correlation between said tests and a graduate's performance in the teaching profession.

The issues raised by this litigation are not new. Rather, they represent the latest link in a chain leading back to the early 1960s. This chapter will present an outline of the major legislative, judicial, and regulatory events that form the precedents upon which the Alabama case will be argued and adjudged. Because of the current litigation, the particulars of the Alabama suit will not be considered. Likewise, the discussion of general issues and previous findings should not be construed to imply any merit or lack of merit in the cases of the present litigants.

The causes of action, as stated above for the Alabama case, however, do suggest some broader issues. It is these broader concerns that will be considered in this chapter. The issues that relate to established legal precedents would seem to include the following:

- (1) The arbitrary and capricious development or implementation of a test or employee selection procedure;
- (2) The statistical and conceptual validity of a test or procedure;
- (3) The adverse or disproportionate impact of a testing program or selection procedure on a "protected group";
- (4) The relevancy of a test or procedure to the identified requirements of the job (i.e., job-relatedness); and
- (5) The use of tests or selection procedures to violate an individual's or group's civil rights.

### Chronology

The chronology of events leading to the present approaches to teacher certification testing is rather complex. First, the events arise from a number of different sources, which include legislation such as Title VII of the Civil Rights Act of 1964; regulatory guidelines such as the Equal Employment Opportunity Commission's Guidelines on Employee Procedures (EEOC, 1966); professional standards such as the APA, AERA, NCME Standards for Educational and Psychological Meas-

urement (1966); and court decisions such as the one arising from Griggs v. Duke Power Co.<sup>2</sup>

The second reason for the complexity stems from the fact that publications are updated and court cases reargued through the appeals process. Thus, one may encounter citations for EEOC guidelines dated 1966, 1970, or 1978, or for Griggs dated 1968, 1970, or 1971. In order to clarify the progression of events, Figure 2.1 presents a timeline of the major legislative, regulatory, professional, and judicial events that occurred between 1965 and 1985.

The information is presented by case. Within each case, the relevant regulation or professional standard is referenced and elaborated upon. An exhaustive list of applicable court cases would be beyond the scope of this dissertation. Instead, eight landmark cases have been selected. Taken together, these eight cases have provided the precedents upon which current decisions have been, and future decisions will be, made.

### The Cases

#### Griggs v. Duke Power Company

The Griggs case was brought by 13 black employees at one of the defendants' power stations. The action was initiated under Title VII of the Civil Rights Act of 1964. Section 703(a) of the Act makes it

<sup>2</sup>Griggs v. Duke Power Co., 292 F.Supp. 243 (M.D.N.C. 1968).

---

	1985	
Connecticut v. Teal (1982)	—	Allen v. Alabama State Board of Education (Trial scheduled for April 1985)
	—	Allen v. Alabama State Board of Education (Filed 1981)
<u>APA Principles for the Validation and Use of Personnel Selection Procedures</u> published (1980)	—	<u>Educational Testing Service Principles, Policies, and Procedural Guidelines</u> published (1979)
EEOC, Civil Service, Dept. of Labor, Dept. of Justice <u>Joint Guidelines</u> published (1978)	—	U.S. v. State of South Carolina (1977)
Washington v. Davis, Supreme Court (1976)	—	U.S. v. State of North Carolina (1975)
Albermarle Paper Co. v. Moody, Supreme Court (1975)	—	Davis v. Washington, Court of Appeals (1975)
APA, AERA, NCME <u>Standards</u> revised (1974)	—	Moody v. Albermarle Paper Co., Court of Appeals (1973)
Davis v. Washington, District Court (1972)	—	Chance v. Board of Examiners (1972)
Public Sector Amendment to Title VII implemented (1972)	—	Baker v. Columbus Municipal Separate School District (1971)
Griggs v. Duke Power Co., Supreme Court (1971)	—	Moody v. Albermarle Paper Co., District Court (1971)
EEOC <u>Guidelines</u> revised (1970)	—	Griggs v. Duke Power Co., Court of Appeals (1970)
Griggs v. Duke Power Co., District Court (1968)	—	APA, AERA, NCME <u>Standards for Educational and Psychological Measurement</u> published (1966)
EEOC <u>Guidelines on Employee Procedures</u> published (1966)	—	Title VII of the 1964 Civil Rights Act implemented (1965)
	1965	

---

Figure 2.1. Timeline for major legislative, regulatory, professional, and judicial events between 1965 and 1985.

an unlawful employment practice for an employer to limit, segregate, or classify employees to deprive them of employment opportunities or adversely to affect their status because of race, color, religion, sex, or national origin. Section 703(h) authorizes the use of any professionally-developed ability test, provided that it is not designed, intended, or used to discriminate.

Prior to the implementation of Title VII, the Duke Power Company had openly discriminated against black applicants and employees. After the effective date of the statute the company began to require that job applicants score satisfactorily on two professionally-prepared tests.<sup>3</sup> The company also dropped a previous requirement of a high school diploma for transfer from a lower to a higher paying department, but retained it for entry-level employment. The plaintiffs alleged that these requirements discriminated against blacks and that the company was in violation of Title VII.

The plaintiffs claimed that under section 703(h) of Title VII a test must measure the ability to perform a particular job. The District Court dismissed the complaint. The court recognized that the plaintiffs' view was in accord with the Equal Employment Opportunity Commission's Guidelines (1966) but decided that the Guidelines did not conform to the intent of the statute. The court held that even a totally unrelated test is acceptable under the federal statute, so long as there is no intent on the part of the employer to discriminate.

<sup>3</sup>The Wonderlic Personnel Test, which purports to measure general intelligence, and the Bennett Mechanical Comprehension Test. The two tests are widely used in industry.

The plaintiffs appealed the decision to the Fourth Circuit Court of Appeals.<sup>4</sup> The appellate court reversed part of the lower court's decision. They held that to add new requirements for promotion from the lowest department, when blacks had been restricted to that department prior to 1965, was illegal because of the discriminatory impact on blacks. However, the court held that if such requirements were not employed to perpetuate past discrimination -- as in hiring new employees -- their use would be acceptable. The appellate court also upheld the lower court's decision that section 703(h) does not bar the use of tests absent of intent to discriminate. Finally, the appellate court stated the EEOC command that the test be job-related was not substantiated by the legislative history of the statute. They cited the fact that Congress had rejected an amendment that would have required employment tests to be job-related. The court did not examine the legislative history of the Tower Amendment, which actually was adopted and which does require job-relatedness.

In 1971 the Supreme Court reversed the appellate decision.<sup>5</sup> The justices concluded that a discriminatory effect alone takes the test outside the protection of section 703(h) unless the test is shown to be job-related. The court's decision gave great weight to the EEOC interpretation and to the legislative history of the statute. From now on it would be necessary to show the job-relatedness of an employment test used in private industry.

<sup>4</sup>Griggs v. Duke Power Co., 420 F.2d 1225 (4th Cir. 1970).

<sup>5</sup>Griggs v. Duke Power Co., 401 U.S. 424 (1971).

Moody v. Albermarle Paper Company

The Moody<sup>6</sup> case followed a similar progression to Griggs, moving through the lower courts to a Supreme Court hearing in 1975. Albermarle Paper Company, like Griggs Power Company, had segregated black employees to the lowest job classifications prior to 1965. In December of 1964 the company offered black employees in certain lines of progression the opportunity to take its personnel tests in order to be considered for other lines of progression. At the same time, the company waived its high school education requirement for incumbent black employees. Most of the blacks scored below the passing score on the two tests.<sup>7</sup> In certain instances, testing requirements were waived for black and white incumbents seeking transfer. The plaintiffs charged the company and union with discriminatory employment practices, including the test requirement for promotion and transfer into better paying jobs, which was alleged to operate to the disadvantage of black employees on the basis of race alone.

The lower court held that a validation study carried out by a consultant to the company and based solely on supervisors' ratings constituted proof of their validity as well as of their necessity to the safe and efficient operation of the business. The high school requirement was found to be unlawful on the grounds that the tests alone were an adequate measure of the mental ability and reading skills required for the jobs in question.

<sup>6</sup>Albermarle Paper Co. v. Moody, 422 U.S. 405 (1975).

<sup>7</sup>The Revised Beta Examination and the Wonderlic Personnel Test.

The appellate court reversed, holding that the company's validation studies were deficient in several respects. First, only eight of the 14 lines of progression were studied; therefore the results were not adequate as a basis for claiming test validity for all jobs for which test requirements existed. Second, although one of the tests proved valid for nine of the 10 job groups studied, both tests were found valid for only one job group. A third fatal defect was the absence of any formal job analysis. Finally, the overall supervisory rating criterion was judged to be inadequate as an unbiased, meaningful measure of job performance.

The Supreme Court upheld the decision by the Court of Appeals that the company's validation studies were deficient. The Supreme Court held that the fundamental benchmark for assessing compliance with Title VII job-relatedness requirements was to be the EEOC Guidelines, which "draw upon and make reference to professional standards of test validation established by the American Psychological Association."<sup>8</sup>

Baker v. Columbus Municipal Separate School District

Baker<sup>9</sup> represented the first Title VII challenge to a teacher examination. The Columbus Municipal Separate School District of Lowndes County, Mississippi, ruled that in-service teachers with one year of service and applicants for teaching positions must obtain a minimum composite score of 1,000 on the National Teacher Examinations (NTE) as

<sup>8</sup>422 U.S. 431 (1975).

<sup>9</sup>Baker v. Columbus Municipal Separate School District, 329 F.Supp. 706 (N.D. Miss. 1971).



a requirement for retention or hiring. The plaintiffs were eight black teachers formerly employed by the Columbus School District. They were joined in the suit by the National Education Association and the Mississippi Teachers Association. The plaintiffs alleged that the NTE requirement had a disproportionate effect on blacks, and that the examination was not job-related.

Educational Testing Service (ETS), the developers and administrators of the NTE, testified for the plaintiffs. Dr. James Deneen, Senior Program Director for Teacher Examinations at ETS, stated that the NTE examinations are not intended to be a measure of teacher performance and that they do not provide information on a number of important factors that should be considered in evaluating teachers.

The court concluded that the use of the NTE score requirement was racially discriminatory. In part, the evidence consisted of a study conducted by the ETS at Mississippi institutions of higher learning. The study revealed that about 90 percent of the students graduating from predominantly white institutions obtained NTE scores of 1,000 or better, whereas only 11 percent of the students graduating from predominantly black institutions obtained scores of 1,000 or higher. Reflecting the testimony given by Dr. Deneen, the court further concluded that the NTE examinations were not a useful predictor of the classroom effectiveness of teachers.

Chance v. Board of Examiners

The Chance<sup>10</sup> case moved the question of job-relatedness from the actual employment process to the area of certification and licensure. This was also one of the first cases to arise under Title VII, because it was not until 1972 that Title VII was extended to cover state and local governmental agencies. The law formerly applied only to private companies. In the pre-1972 cases, although the challenge would initially be brought on general equal protection grounds under the Fourteenth Amendment, the plaintiffs would then "indirectly" refer to EEOC Guidelines. This "indirect application" of the EEOC Guidelines provided a strong motivation for Congress to formally amend Title VII to specifically include public employees under the Act.

The plaintiffs in this case were minority applicants for supervisory positions in the New York City public schools. They sought a preliminary injunction against awarding supervisory positions to applicants on the basis of previous examinations and against the further use of the examinations.

Although data on the passing rates of minority and white candidates were not complete, the data that existed, together with the numbers of white and minority supervisors actually assigned, led the court to conclude that the test did have a disproportionate impact on minority groups. The court observed that the experts for both the defense and the plaintiffs had indicated that the plan used by the Board for

<sup>10</sup>Chance v. Board of Examiners, 330 F.Supp. 203 (S.D.N.Y. 1971).

constructing tests was proper. However, the court also noted that there were many instances in which the planned procedures for constructing the tests had not been properly carried out. In addition, the court reviewed the content of the tests and noted that much of the material seemed to have little relevance to the duties of a school supervisor and appeared to measure only the applicant's memory for isolated facts.

A preliminary injunction was issued, restraining the Board of Examiners from giving tests in the future and from promulgating eligibility lists based on the examinations.

The decision was upheld by the appellate court in 1972.<sup>11</sup>

#### Washington v. Davis

Washington v. Davis<sup>12</sup> represents a major turnaround from the cases discussed above. The plaintiffs, black police officers in the District of Columbia, charged that the use of a test with adverse impact, which had not been shown to be job-related, violated Title VII.

The District Court<sup>13</sup> ruled that the test was valid. They observed that the police department had followed a systematic and vigorous affirmative action effort to recruit black police officers. The court ruled that the higher passing rates for whites placed the burden of proof on the District of Columbia to show that the test was job-

<sup>11</sup>Chance v. Board of Examiners, 458 F.2d 1167 (2nd Cir. 1972).

<sup>12</sup>Washington v. Davis, 426 U.S. 250 (1976).

<sup>13</sup>Davis v. Washington, 348 Supp. 15 (D.C. 1972).

related. The court took note of the fact that the test was widely used in selecting individuals for Civil Service positions throughout the country. The court based its decision on the fact that the test was clearly related to success in recruit training, even though there had been no job performance validation of the test.

The Court of Appeals,<sup>14</sup> in line with previous Title VII related actions, reversed, citing the fact that job-relatedness had not been established.

In a surprising move that has caused considerable confusion in the use of employment tests, the Supreme Court reinstated the holding of the district court. The decision did not address the Title VII precedents concerning test validation upon which the court of appeals had ruled. The Supreme Court was able to avoid confronting those holdings because of the dichotomy it created in Davis between constitutional liability (based on impact). Since Davis did not include a Title VII claim, and there was no basis for a claim of intentional discrimination (the Department had a "model" affirmative action program), the constitutional claim was easily disposed of. The only relevant validation standards left to consider were regulations of the United States Civil Service Commission. These regulations, according to the Court majority's reading, specifically included "success in training" as a proper criterion for assessing the validity of a selection instrument. Therefore, the test was upheld.

<sup>14</sup>Washington v. Davis, 512 F.2d 956 (D.C. Cir. 1975).

A reasonable interpretation of the Supreme Court's consideration of test validation issues in Davis would appear to be that the basic substance of Title VII job-relatedness standards, as articulated in the agency guidelines, should continue to be enforced; but that if a defendant is acting in good faith and a simple entry-level examination has obvious job relevance, courts may not insist on strict psychometric requirements.

U.S. v. State of North Carolina and U.S. v. State of South Carolina  
North Carolina<sup>15</sup> and South Carolina<sup>16</sup> bring the issues raised in previous cases to the focus of this study: the use of a professionally developed test to certify or license teachers.

In North Carolina the United States brought suit against the State, claiming that the North Carolina General Assembly, against the advice of the State Superintendent of Instruction and the Educational Testing Service, had enacted legislation requiring a prescribed minimum score to be achieved on the NTE. A score of 950 had been chosen as the minimal total score for a Class A teaching certificate.

The court concluded that the NTE did not measure teaching skills, but did measure the content of the academic preparation of teachers. Although the State has a right to protect the public from incompetence by establishing minimum standards of knowledge and skills, the establishment of 950 as the cutoff score was deemed arbitrary. The court

<sup>15</sup>U.S. v. State of North Carolina, Civil No. 4476 (E.D.N.C. 1975).

<sup>16</sup>U.S. v. State of South Carolina, 445 F.Supp. 1094 (D.S.C. 1977).

held that the State could reinstate a written test cutoff score if that score were first validated.

In South Carolina the State had been using the NTE for certification since 1945. During that period, the minimum scores for passing were raised several times. Local school boards within the state used NTE scores, along with other requirements, to select teachers and to determine salary level. The NTE had an adverse impact, resulting in fewer blacks than whites being certified and hired. The plaintiffs charged that the use of the minimum cutoff score violated the Fourteenth Amendment and Title VII.

In South Carolina the court found in favor of the defendants. Although the NTE did have an adverse impact on minorities, the court concluded that the evidence was sufficient to establish the validity of the examinations as appropriate measures of minimum teacher competency. It was held that a content validity study was adequate under Title VII and that the NTE examinations were fair and objective measures for determining both teacher certification and salary levels.

Taken together these two cases suggest that a test used for teacher certification must be properly developed and that its use, including the determination of a passing score, must be proven valid. The content must be shown to be related to the job in question.

#### Connecticut v. Teal

One final case is of interest to the present discussion. In 1982 the Supreme Court ruled in favor of black employees of a Connecticut

state agency.<sup>17</sup> The employees claimed that a written promotion examination discriminated against them on account of their race, in violation of Title VII of the 1964 Civil Rights Act. The U.S. District Court for the District of Connecticut dismissed the suit for failure to establish a prima facie case, and the plaintiffs appealed. The Court of Appeals<sup>18</sup> reversed the decision and remanded the case back to the District Court.

The case was brought before the Supreme Court in 1982. The case involved the fact that the original plaintiffs were claiming that the test was discriminatory and not job-related, even though the promotion figures did not support de facto segregation. The examination was given to 48 black and 259 white candidates. The passing rate for blacks was approximately 68% of the white passing rate. Between the time that the plaintiffs brought suit and the case came to trial, some of the plaintiffs were promoted. When the case came to trial, 22.9 percent of the black candidates and only 13.5 of the white candidates had received permanent promotions.

The Supreme Court ruled that the "bottom line" is no defense against a claim of racial discrimination. The Court pointed out that there was a clear distinction between unlawful discrimination and discriminatory intent. A majority of the justices suggested that Congress never intended to give an employer license to discriminate against some employees on the basis of race or sex merely because the employer fa-

<sup>17</sup>Connecticut v. Teal, 457 U.S. 440 (1982).

<sup>18</sup>Connecticut v. Teal, 645 F.2d 133 (2nd Cir. 1980).

vorably treats other members of the employee's group.

With Teal the Court appears to have completed the intent and de facto discrimination equation first stated in Griggs. According to this ruling, Title VII is violated when there is discrimination, even in the absence of intent and when there is an intent to discriminate, even in the absence of discriminatory results.



C H A P T E R    I I I  
T E C H N I C A L   C H A L L E N G E S   T O   T H E   T E S T   D E V E L O P E R

Introduction

The court cases presented in Chapter II raise a number of technical concerns for the professional faced with the task of developing a fair, useful, and legally-defensible test to certify new teachers. Over the last 20 years some of the concerns have been addressed very well. Practitioners have developed ingenious ways to ascertain the job-relevancy of some instruments, and psychometricians have made great strides in the development of better and more useful validity and reliability procedures. But as the success of some recent legal challenges would suggest, the job is far from complete. In Chapter III, a number of technical concerns are presented for consideration. They include the following:

- \* validation procedures
- \* reliability
- \* error compensation
- \* job analysis

These issues, of course, are not the only ones facing practitioners today. Nor does attention to the topics listed here guarantee protection from a legal challenge. The topics do, however, reflect the ways in which practitioners have viewed the job at hand in light of the court decisions presented in Chapter II. In this respect they

represent the recent, present, and anticipated challenges that face a test developer.

### The Technical Challenges

#### Validation procedures

Test validity is a ubiquitous concern in the teacher certification cases. It takes many forms. One may look at the validity of the development process, or one may question how valid a particular passing score is. A "valid" employment test is one that can document its technical rigor and that can prove its job-relevancy through a well-developed job analysis procedure or a formal post-development validity study.

In any question about validity, it is necessary to identify the type of validity that is being established. The concept of test validity refers to the degree to which an instrument measures what it is designed to measure. Validity is an indication of how precise the instrument is. A math test, for example, that could not distinguish highly competent mathematicians from individuals with little or no math background would be virtually absent of validity.

There are three types of validity relevant to employment testing: construct validity, predictive (or criterion-related) validity, and content validity. A discussion of each and its relevance to the legally-defensible teacher certification test is presented below.

Construct validity. When an investigator is attempting to establish the construct validity of a test, he or she is attempting to relate the test to some psychological trait or construct. The construct may be a personality variable, such as sociability or maturity, or it may be an intellectual variable, such as intelligence or creativity. To the degree that a test can illuminate the trait in question, the test is said to have construct validity. The most difficult aspect of a construct validity study is often the problem of defining the construct. No attempt to measure the instrument's validity can begin until the construct is clearly defined. The results and interpretation of a construct validity study are often limited by a degree of "fuzziness" around the definition of the construct. The most familiar example of this problem is found in the area of intelligence testing. After years of refinement, the well-established intelligence tests are good measures of IQ. The instruments have reasonable reliability estimates, and the test results are useful to practitioners concerned with intelligence (this utility in the clinical and research settings is one indicator of the instrument's construct validity). The problem arises when one asks, "What's IQ?" The instruments that measure IQ are "valid" only to the degree that one agrees with the stated definition of the construct.

A second problem for construct validation experts is developing a measure of something as intangible as a psychological trait. In the case of intelligence testing, it has taken many years of refinement to identify the right questions and the right interpretation of responses.

And even after rigorous refinement, there is never a perfect match between the construct and the measure.

Does this mean that construct validity is not useful in employment testing? Not at all. Some traits or constructs may be critical to certain jobs. There are, for example, certain traits that we would want to see in an individual responsible for a nuclear power plant. There are, likewise, certain constructs that we would not want to see in that individual. A properly-conducted construct validity study can tell the test users how confident they can be about the absence or presence of the traits or constructs in question. That they can never be certain (i.e., that there is never perfect validity) does not mean that the information that is available is not useful.

While construct validity is a common approach in many employment settings, it is not commonly used in teacher employment or teacher certification. There are two reasons for this. First, much of what is tested (e.g., content knowledge, rules of pedagogy, philosophy of education) is best validated by other measures (e.g., criterion-related validity, content validity). Second, those areas that might conform to a construct are extremely difficult to define. The argument as to what makes a good teacher has continued unabated since the time of Socrates. And while various task forces and institutes continue to take on the task, it does not appear that agreement is near at hand. This is a critical problem and one that deserves more serious attention. But until the attributes can be identified, construct validity will not play a major role in constructing the legally-defensible teacher certification test. -

Criterion-related validity. This second approach to validity involves comparing performance on the test to some other direct but independent measure of what the test is trying to predict. The criterion can be based in the present, in which case it is referred to as concurrent validity, or it can exist in the future, in which case it is known as predictive validity. Testing incumbent employees and comparing their test results with their most recent performance evaluation would be an example of a concurrent validity study. Testing a group of newly-hired employees and then holding the results for comparison with some future measure, such as a performance appraisal after three years on the job, would be an example of a predictive validity study.

Even though the two procedures sound similar, they are vastly different in practice. A concurrent validity study is relatively easy to carry out, but its utility is limited. Most employers want tests that will tell them something about the new person who has just applied, not about the employees who are already hired. The ability of a test to predict how well a new person will be able to perform "down the road" can be estimated only by a predictive validity study. But unlike the concurrent validity study, a predictive validity study requires a substantial investment of time and resources, as well as some difficult political and ethical choices.

National Evaluation Systems (1976) lists six requirements for implementing a predictive validity study:

- (1) admission of all applicants for employment in the field

- (2) sufficient lapse of time before observing the criterion variable
- (3) unexamined, unused results of the test -- i.e., the predictor -- stored until correlated with the criterion
- (4) the measurability of the criterion -- i.e., a mechanism for collecting accurately and reliably the reasons for teacher dismissal that clearly separate knowledge of content as one of those reasons
- (5) sufficient sample size
- (6) stability of the criterion -- i.e., it must be unaffected by maturation or learning

Taken together, the requirements mean that a group of teacher candidates would be randomly selected and that no criterion could be used to include or exclude any individuals. After taking the test, all would have to be hired and placed in teaching positions, without benefit of the analyzed test results that will have been sealed, unscored, for future use. All subjects would teach for at least three to five years without developing new skills or learning anything on the job. At the end of the three-to-five-year period, the subjects would be assessed using a clear measure of teacher competence, the development of which is a major hurdle in its own right. Finally, the tests would be scored and correlated with the independent criterion to determine which subjects should not have been hired five years previously.

The courts, faced with this dilemma, have suggested that predictive validity should not be a requirement in teacher certification

tests. In Chance<sup>1</sup> the court noted that

Predictive validity is of greater significance in evaluating aptitude tests than proficiency tests. Furthermore, it often takes a long time to establish such validity and even then the evaluation depends upon the reliability and fairness of the field appraisal of performance on the job.

While some still have the goal of developing an instrument that will be able to predict the candidate who will be able to successfully "get it across" in the classroom, most parties, including the courts, are settling for narrow indices of teacher competence such as minimum content knowledge and satisfactory completion of a course of study.

Content validity. The third approach to validation, and the one most often employed in licensing and certification, is content validity. The objective in a content validation study is to determine to what extent the content measures the domain or domains of the job or area of study. To content validate a test of mathematics knowledge, for example, the investigators would examine all elements of the test and try to ascertain how well the test covered the essential areas of mathematics knowledge. This is done by defining the domain or domains of knowledge. For an expanded explanation of domain specification procedures, the reader is directed to Ebel (1962); Hively, Patterson, and Page (1968); and Popham (1975, 1980).

After the domain has been specified, expert judges review each of the test items in an attempt to determine whether the use of the test

<sup>1</sup>Chance v. Board of Examiners, 458 F.2d 1167 (2nd Cir. 1972).

(i.e., the test score) is appropriate. Hambleton (1980) describes the logic behind this procedure as applied to criterion-referenced tests.

Generally speaking, the quality of criterion-referenced test items can be determined by the extent to which they reflect, in terms of their content, the domains from which they were derived. The problem here is one of item validation; unless one can say with a high degree of confidence that the items in a criterion-referenced test measure the intended instructional objectives, any use of the test score information will be questionable. (p. 86)

Most content validity procedures use a combination of two approaches; they combine expert judgments with empirical evidence. While the specifics of the expert judgment task vary widely between item-validation approaches, the task remains one of looking at the item and judging whether or not the test item is technically sound and measures some significant aspect of the domain. In addition, most approaches require some judgment as to the representativeness and proportionality of the items to the domain. In order for a test to meet the criterion of representativeness and proportionality, it must reflect the entire breadth of the domain and it must place the greatest emphasis (i.e., largest number of items) on the most significant aspects within the domain. A test that sampled knowledge or behavior from only one "corner" of a domain would not be representative. A test that put great weight on tangential or insignificant aspects of a domain would be disproportionate.

In recent years, these issues have been of paramount concern in determining the legal defensibility of employment, licensing, and cer-



tification tests. If a test developer or test user wants to make a claim as to the content validity of an instrument, he or she should be prepared to provide answers to four questions:

- (1) Are all test items technically sound?
- (2) Does each test item measure some meaningful aspect of the domain?
- (3) Do the test items, taken as a whole, fairly reflect the breadth of the domain?
- (4) Do the most important aspects of the domain receive the greatest emphasis within the test?

A number of approaches have been developed to help the practitioner prepare to answer these questions. All the approaches rely on some degree of expert judgment. This requirement has often raised concern about arbitrary standards. Quoting James Popham, Hambleton (1980) provides a sharp rebuttal to these concerns.

Unable to avoid reliance on human judgment as the chief ingredient in standard-setting, some individuals have thrown up their hands in dismay and cast aside all efforts to set performance standards as arbitrary, hence unacceptable.

But Webster's Dictionary offers us two definitions of arbitrary. The first of these is positive, describing arbitrary as an adjective reflecting choice or discretion, that is, "determinable by a judge or tribunal." The second definition, pejorative in nature, describes arbitrary as an adjective denoting capriciousness, that is, "selected at random and without reason." In my estimate, when people start knocking the standard-setting game as arbitrary, they are clearly employing Webster's second, negatively loaded definition.

But the first definition is more accurately reflective of serious standard-setting efforts. They represent genuine attempts to do a good job in deciding what kinds of standards we ought to employ. That they are judgmental is inescapable. But to malign all judgmental operations as capricious is absurd. (p. 102)

As Popham indicates, judgmental approaches to content validation are also referred to as standard-setting procedures. This term identifies the essential element in the validity of a minimum competency test such as those used in teacher certification. It is not only the internal aspects of the test that must be judged valid, but also the way in which the test is used to identify masters and non-masters, successes and failures. The establishment of a passing score becomes the standard by which the candidate is judged, and it is this standard that must be shown to be valid. In United States v. North Carolina,<sup>2</sup> the use of the NTE was judged to be illegal because the selection of a passing score was not based on any clear rationale. In United States v. South Carolina,<sup>3</sup> however, the NTE was found to be a valid and appropriate measure, because the passing score was based on a systematic, empirical approach.

Three content validity or standard-setting procedures stand out in teacher certification; they are Nedelsky (1954), Angoff (1971), and Jaeger (1978). These three approaches have been combined, modified, and refined in a number of states. In each procedure judges are asked to consider how a minimally competent teacher would perform on each

<sup>2</sup>United States v. North Carolina 400 F.Supp. 343 (E.D.N.C. 1975).

<sup>3</sup>United States v. South Carolina 425 F.Supp. 789 (E.D.S.C. 1977).

item. In the Nedelsky method, judges examine each of the item distractors (i.e., A, B, C, D of a multiple-choice question). They are asked to identify which distractors the minimally competent teacher should be able to eliminate as incorrect. A minimal passing level (MPL) is then established for that item by the reciprocal of the number of remaining responses. Thus, if one out of five responses was eliminated, the MPL would be  $1/4$ , or the reciprocal of the four remaining responses. This reciprocal represents the "chance score" for the minimally competent teacher. The reciprocals for all of the judges' items are summed, and the judges' sums are then averaged to produce a passing score or standard. A final statistical procedure is then applied to the data to account for the scores of candidates who fall close to the passing score. The standard deviation of the judges' standards is computed, and this standard deviation is then multiplied by a constant,  $K$ , decided upon by the test users. The result of this computation is then added to (or subtracted from) the original standard to produce the final passing score.

In the Angoff (1971) procedure, each judge is asked to imagine the minimally competent teacher and to estimate the probability that this person will make a correct response. The Angoff judges envision a group of minimally competent teachers and offer an estimate of the proportion of this hypothetical group who could answer the item correctly. These proportions or probabilities are then summed to produce the passing score or standard.

Jaeger's (1978) process involves an iterative approach, using different types of experts. Rather than ask judges to envision the minimally competent teacher, the Jaeger approach uses questions such as "Should every candidate for certification be able to answer this item correctly?" and "If a teacher candidate does not answer this item correctly, should he or she be denied certification?" Groups of judges from several areas of expertise (e.g., incumbent teachers, administrators, professors from teacher training programs) review normative data and respond to the two questions for each item. The ratings of all judges within a group are pooled, and a median is computed. The minimum median across all of the groups becomes the standard.

Variations of Nedelsky, Angoff, and Jaeger have been applied to teacher certification tests in some states. In Georgia and in Alabama a modification of the Nedelsky approach (Nassif, 1978) was implemented. The procedure actually incorporates elements of both Nedelsky and Angoff. According to Nassif (1978),

Panels of expert judges reviewed items independently on an item-by-item basis. The following was asked about each valid item: "Should a person with minimum competency in the teaching field be able to answer this item correctly?" Each judge was asked to imagine the skills of a hypothetical candidate with minimum competency in the content of a teaching field. Within this frame of reference the item was examined as to whether it required too sophisticated a knowledge of the content or whether it required content knowledge of a trivial or minor importance.

Judges responded "yes" if the item was considered appropriate for measuring minimum competency or "no" if otherwise. The "I don't know" option

was available for judges unfamiliar with the content of an item.

The significance of agreement was determined by comparing the number of "yes" responses with probability tables for the binomial distribution. The ratings of "I don't know" were not considered for any item, so that dichotomous ratings with different numbers of judges were generated. If the probability of receiving a given number of "yes" ratings (i.e., appropriate for minimum competency) was less than a chance of 1 in 10, the item was classified as an appropriate requirement for minimum competency.

In South Carolina three different approaches were applied to different aspects of the testing program. For the validation of the NTE, researchers used a variation of the Angoff procedure. Instead of asking judges for the probability (.01 to 1.00) that minimally competent candidates could answer the item correctly, the researchers had the judges make their ratings against a more restricted scale (1 to 7). While this restricted the range of the judges' estimates, it greatly simplified data reduction and analysis. In a separate study of 10 customized, criterion-referenced tests, another group of researchers applied the Angoff approach as described above. Finally, in a content validation procedure for the newly-developed Basic Skills test, researchers used the Jaeger approach.

Nassif (1982) lists several reasons that these three approaches continue to be selected for teacher certification tests.

- \* These procedures are based on and permit an item-by-item review. This is a very important consideration for tests that are regenerated in part

quite frequently, due to test security and job-analysis requirements.

- \* The procedures permit the incorporation of performance data in judgment, if desired as additional information in the decision-making process.
- \* These procedures allow the establishment of single or multiple cut scores as necessitated by the testing program. In the case of multiple cut scores, compensatory or disjunctive scoring can take place.
- \* These models are easy to understand -- a factor that should contribute to the reliability of judges' ratings and to the comprehensibility by constituent audiences.
- \* These involve and rely on expert judges.
- \* The cut score that is set does bear a relationship to necessary job performance -- a legal requirement. It allows all competent candidates to pass, without restriction from quotas.
- \* They do not require information (statistical or demographic) not generally available.
- \* These methods produce a cut score that can be adjusted easily by standard error of measurement to incorporate relevant employment factors.

- \* These methods can be employed on any number of items, although the original Nedelsky and Jaeger approaches are prohibitive due to the length of the process.

### Reliability

Reliability studies attempt to estimate the degree of consistency in a test's application. The more consistently that a test measures whatever it is designed to measure, the more reliable it is. The concern with reliability is closely related to the concerns about validity. It is a tenet of psychometrics that if an instrument is significantly lacking in reliability, it cannot be valid for any given purpose. Reliability is a requisite of validity.

There are two aspects of the teacher certification test that must be consistent: performance on individual items must be consistent over time or across forms, and the decisions (e.g., mastery, non-mastery) that result from the administration of a test must be consistent over time or across forms. As regards the performance of the test items, there are three methods that are employed with criterion-referenced tests. While each of the three is theoretically applicable to teacher certification testing, ones that require test-related analysis or parallel forms are not likely to be found. Parallel test forms are not used because teacher testing programs cover many fields (79 in Oklahoma). The additional development and administration expense would be substantial. The problem with establishing test-retest data (i.e.,

asking a candidate if he or she would mind retaking his or her certification test) is evident.

As a result of these restrictions, estimates of stability, which reflect a test's consistency over time, and estimates of equivalence, which reflect a test's consistency across forms, have not been commonly employed in certification testing. The remaining method of test item consistency is referred to as internal consistency. Internal consistency, like stability and equivalence, is based on a correlational analysis. In the case of a stability estimate, it is a correlation between individuals' performances on two administrations of the same form. The equivalence estimate uses a correlation between the equivalent test forms. When test-retest and parallel forms are not available, however, it must be internal components of the test that are correlated.

Two approaches to the estimation of internal consistency are commonly used. Split-half reliability involves dividing the test items into two groups (e.g., odd-numbered items and even-numbered items) and performing a correlational analysis on the two groups. The second approach takes the same concept one step further. The Kuder-Richardson indices of item homogeneity (KR-20, KR-21) examine the average of all possible split-half reliability coefficients (Elliot, 1982).

The second aspect of reliability for teacher certification tests involves decision consistency: if the certification test were readministered, would the same people be issued certificates? Subkoviak (1982) reviews four methods to decision-consistency reliability that



are relevant to the dichotomous decision-making found in a certification program. Each of the four methods provides an estimate of the proportion of individuals consistently classified as masters and non-masters ( $P_0$ ), and the proportion of individuals consistently classified beyond that expected by chance (Kappa). One method, developed by Swaminathan, Hambleton, and Algina (1974), uses two administrations of one test form or two parallel test forms. One method, developed by Huynh (1976), and another developed by Subkoviak (1976), use two different estimates of  $P_0$  and Kappa based on a single test administration. The fourth method, developed by Marshall and Haertel (1976), uses one estimate of  $P_0$  based on one administration. For a practical application of each method and a thorough comparison of the four methods, the reader is referred to the full Subkoviak (1982) review.

#### Error compensation

For all mastery tests, researchers are concerned that (1) the total error surrounding the passing score be minimal ( $P_0$ ), and that (2) the likelihood that the classification as master or non-master was due to chance be less than some prescribed level (Kappa). In the case of licensing and certification, however, another concern often arises. In addition to being able to estimate the amount of error, it may also be necessary to predict the direction error. The task of classifying individuals as masters or non-masters is analogous to the task of avoiding type I and type II errors during hypothesis testing. Figure 3.1 indicates the possibilities involved in the classification task.

Fail Masters Type II Error	Correct Decision
Correct Decision	Pass Non-Masters Type I Error

Figure 3.1. Type I and type II errors in the identification of masters and non-masters.

There are many instances in certification and licensing when one type of error is more acceptable than another. In the case of brain surgeons -- to use a dramatic example -- the consequences of licensing one incompetent far outweigh the consequences of denying licensure to even a few qualified candidates. As there is always some error in the measurement of human attributes, we must make sure that we can minimize the more critical type of error, even if it means tolerating some moderate level of error in the opposite direction.

Which type of error should be minimized in a teacher certification test? Licensure, as in the case of a doctor, grants the recipient the right to practice. Once a doctor is licensed, he or she can hang out a shingle and begin to practice in the profession. Certification, on the other hand, merely permits the holder of the certificate to apply for a position within the profession. A teacher candidate who has completed an approved program of study and passed the certifying examination can not simply begin teaching. He or she must first be hired and then, in many cases, must perform a supervised first-year internship. The

teacher candidate's situation, then, is very different from that of the doctor's. Here we are more concerned that we do not bar a competent candidate from consideration than we are that some border-line candidates may be able to apply for teaching positions. In other words, there are enough checks in the system, after the certification test, that we can be more concerned with not losing the potentially good teacher than with guaranteeing that we have eliminated every possible non-master.

One way to decrease the likelihood of type I or type II error is to adjust the passing score or standard by some standard error of measurement (SEM). The SEM is a statistical estimation of the average amount of error surrounding the scores on a test. Like the standard deviation, which indicates how much variation there is in a sample of people or events, the SEM is an indicator of the degree of variability around a statistic or score. If the variability is normally distributed, we can determine the percentage of scores that are likely to be in error. If, for example, the passing score on a certification test is 80 and the SEM has been calculated at 3, we know, from an examination of the normal distribution, that approximately 68% of the type II error (failing to certify people who are actually qualified) will be found within 3 points of the passing score. Likewise, we know that 95% of the type II error will be between 74 and 80, and 99.7% will be between 71 and 80. The practitioner, then, has a statistical tool that can be used to compensate for type I or type II errors. The decision to raise or lower the original standard by some SEM, however, is still

a programmatic one and not a statistical one. The test user must carefully consider the consequences of type I and type II errors for a particular program and apply an SEM based on a clearly-defined rationale.

### Job analysis

Job relevancy has been a major issue in many of the employment-test cases of the past 20 years. The valid use of an employment test is based on a clear understanding of the rational relationship between the test and the knowledge, skills, and abilities required to do the job. For the pencil-and-paper, criterion-referenced or norm-referenced teacher certification tests, the primary interest is content knowledge. The defense of teacher licensing tests is based on an ability to show that the test is a fair and accurate measure of what the candidate needs to know in order to be a successful teacher. Because the emphasis is on content knowledge rather than on performance skills, many traditional job analysis techniques, such as the critical incidence technique, are inappropriate.

The appropriate approach to job analysis for a teacher certification test involves an examination of the subject matter to be taught. A content knowledge test for a math test, for example, focuses on the material that a math teacher presents in class. This is the knowledge component of the job "math teacher." While there is some variation in the job analysis process used by different states, there is also considerable commonality. Most procedures use panels of content experts to rate, on some dimension considered important, the information that a

teacher needs in order to do his or her job. Regardless of the particular method selected, there are a number of requirements that should be considered.

Raters. A great part of the legal defensibility of a test rests with the credibility of the expert judges. Where at all possible, the job analysis should include ratings by job incumbents. These are the people who know what the current teaching position actually involves. It may also be useful to get input from other knowledge sources, such as administrators, state department curriculum experts, or college and university professors of teacher education. It is also advisable to develop a stratified random-sampling procedure for selecting raters. The composition of a panel assembled to rate the job of teaching a particular subject should reflect some important demographic characteristics. These might include race, sex, geographic distribution, school size, or years of experience.

Level of competence. In asking people to consider the requirements of a job, it is necessary to specify a particular level of competence. An examination of any position will reveal a range of talent from the barely competent (or incompetent) to high performer. The reliability of the job-analysis procedure is dependent on all raters' having a similar image of "what it takes to do the job." For the teacher certification test, this usually requires some definition of the minimally competent teacher. The clearer the definition of the level of competence, the greater the reliability of the job-analysis procedure. And since reliability is a necessary requirement for validity, the definition plays an important role in the defense of the test.

The job versus training. One aspect that will invariably arise during the job-analysis process concerns the difference between what is required on the job and what is taught in teacher education institutions. If there is a discrepancy between these two, the job analysis must focus on what is required on the job. This is not to say that what is taught in the teacher education institutions is irrelevant. If a state requires a candidate to meet certain standards (i.e., pass a certification test) on the one hand, and on the other hand approves programs of education based on a different set of standards, it may be opening up an area for a legal challenge. This will be discussed more in Chapter V. For the job analysis, however, the focus is the job and not the preparation for that job.

Local versus national perspective. It has been pointed out that an instrument must be validated for a given purpose. A nationally-developed history test may not be a valid job measure for a history teacher in a particular state. South Carolina was able to defend the use of the NTE because they conducted a validity study that indicated that the component tests were valid measures of the job of teaching certain subjects in South Carolina public schools. The job analysis should focus on the job as it is performed in a particular state. In fact, there may be little difference between the way that algebra is taught in Colorado or Ohio, but unless the study was designed to assess the elements of the local job, the situation remains open to challenge.

Emerging areas. There is one exception to the rule that the job analysis be limited to the job as it is performed here and now. There are instances when new information or new approaches will be entering the profession in the immediate future. It may be, for example, that some schools are not teaching courses involving the use of computers. A job analysis of incumbents would suggest that this is a low-incidence aspect and it should, therefore, not be included in a teacher certification test. The case might be made, however, that the state is in the process of purchasing computers for most of the schools and that the demand for computer courses is on the rise. The state may then reasonably say that it will be an important aspect of the job for those incoming teachers who are now taking the test. While there is no specified limit to the amount of material that can apply to this "emerging area" rule, it would be inadvisable to develop a purportedly job-relevant testing program that had more than 10% of the test dedicated to new or emerging material.

C H A P T E R    I V  
A TEST DEVELOPMENT EXAMPLE

Introduction

A number of the legal pressures that led state departments to reconsider their testing and certification procedures have been identified. A number of psychometric and programmatic responses to these pressures have also been identified. In this chapter, a test development example that uses many of the approved practices is presented.

Between September 1980 and February 1983, this writer served as the project manager for National Evaluation Systems, Inc., during the development of the Oklahoma Teacher Certification Testing Program. The responsibilities of the project manager included supervision of the day-to-day development activities. Working with other staff at National Evaluation Systems (NES), and with the Oklahoma State Department of Education (OSDE), the project manager carried out each of the steps in the development of a set of programmatically sound, psychometrically rigorous, and legally defensible teacher certification tests.

The example presented here was custom developed to meet specific needs at a specific time. It is not suggested that the Oklahoma example be taken as a model for all certification testing programs at all times. Equally valid techniques are currently available for some elements of the development process, and it is expected that new and better procedures will be developed in the future. The utility of this



example, then, lies not in its specifics but in the logic that guided its development. Each step in the process was guided by a concern for the issues raised in Chapters II and III of this dissertation and by a concern for the programmatic needs of the state of Oklahoma.

#### Overview of the Oklahoma program

In 1980 the Oklahoma State Legislature enacted a bill designed to establish provisions for better opportunities at the preservice level to improve the competence of those who teach in the Oklahoma schools. The bill established three criteria as the measure of competence:

- (1) the successful completion of an approved teacher education program;
- (2) the successful completion of an entry-year teaching experience; and
- (3) a passing score on a standardized, externally developed, administered, and scored content knowledge examination. In the summer of 1980 the OSDE contracted with NES to develop, administer, and score this examination.

The program called for the development of a content knowledge test for each of the certificates offered by the OSDE. For some certificates, such as home economics or speech pathology, a single 120-item test was developed. For other certificates that are part of a larger field, such as the branches of mathematics or the sciences, a 100-item "umbrella" test and 80-item specific area tests were developed. In all, 79 separate tests were developed. These include the following:

#### General Tests (120 Items)

Art

School Superintendent

Audiovisual Specialist  
 School Counselor  
 Distributive Education  
 Driver and Safety Education  
 Health and Physical Education  
 Reading Specialist  
 German  
 Spanish  
 Psychology  
 Agriculture  
 Secondary Principal  
 Home Economics

Umbrella Tests (100 Items)

Industrial Arts

Business Education

Speech Pathology  
 Early Childhood Education  
 Elementary Education  
 Journalism  
 Librarian  
 Psychologist  
 Psychometrist  
 French  
 Latin  
 Speech and Drama  
 Deaf and Hard of Hearing  
 Elementary Principal

Specific Area Tests (80 Items)

Drafting

Metalwork

Woodwork

Accounting

Business Economics

Business English

Business Law

Business Machines

Business Mathematics

Office Practice

Shorthand

Language Arts	Typewriting Grammar and Composition American Literature English Literature World Literature Library Science
Mathematics	Algebra Geometry Trigonometry Mathematical Analysis Calculus
Music	Vocal Music Instrumental Music
Science	Biology Botany Chemistry Earth Science General Science Physics Physical Science Anatomy and Physiology Zoology
Social Studies	American History Economics Geography

	Government
	Oklahoma History
	Sociology
	World History
Special Education	Emotionally Disturbed
	Physically Handicapped
	Learning Disabilities
	Mentally Handicapped
	Visually Impaired

The following steps in the development process will be reviewed in the remainder of this chapter:

- Formation of Committees
- Content Outlines
- Objectives
- Job Analysis
- Selection of Final Objectives
- Item Development
- Content Validation
- Standard Setting
- Test Form Development
- First Administration
- Standard Error of Measurement
- Monitoring Activities

### Formation of committees

For each certification area (e.g., biology, typing, algebra), a committee of content experts was convened to guide and review the development of all content and all materials. The selection of the approximately 300 members of the advisory committees was a difficult and critical part of the development process. An attempt was made to have each committee reflect certain demographic characteristics. The characteristics included sex, position (college professor, public school teacher, or State Department curriculum expert), region within the state, race, and experience (years in the particular field). Since the committees were small (i.e., six to twelve), and there was a finite number of candidates in a given subject area, some adjustments had to be made, but on the whole, the committee membership was a fair representation of all content experts in the field.

A significant part of the claim that the tests are valid rests with the composition and activities of these committees. The committees discussed issues and revised materials throughout the test development process. While it is still possible that some point of bias (e.g., sex bias, racial bias) could escape detection, the likelihood was greatly reduced by assuring that all perspectives were given voice during the development and field test phases.

In addition to the advisory committees, the OSDE also selected individuals to participate in the job analysis and standard-setting phases of the project. The composition of these groups, and the selection procedures, are discussed below under the appropriate headings.

In all, over 10,000 Oklahomans contributed information and opinion during the design and construction of the tests. Figure 4.1 illustrates the relationship of the OSDE and NES to the advisory committees during the development process.

### Content outlines

During the fall of 1980, the NES staff developed content outlines for each of the 79 test areas. The outlines were based on a review of Oklahoma curriculum guides, textbooks, and other materials. The outlines were designed to provide an overview of all content knowledge for a specific field. The documents used standard outline format (i.e., I, A, 1) to divide the knowledge for a field into manageable increments that could ultimately translate into behavioral objectives. The first level of organization within an outline, represented by a Roman numeral, identified the major subdivisions within the field. When the final tests were developed, these headings would correspond to the subareas by which the test scores would be reported. For a test of general mathematics, for example, these headings might be algebra, geometry, calculus, etc.

The second level of organization within the outline, represented by a capital letter, indicated a level of specification appropriate for the development of behavioral objectives. On an American History test, for example, a topic at this level might be "persons and events of the Civil War," or "causes of the westward migration." The third level, indicated by an Arabic numeral, identified possible subelements of a topic.

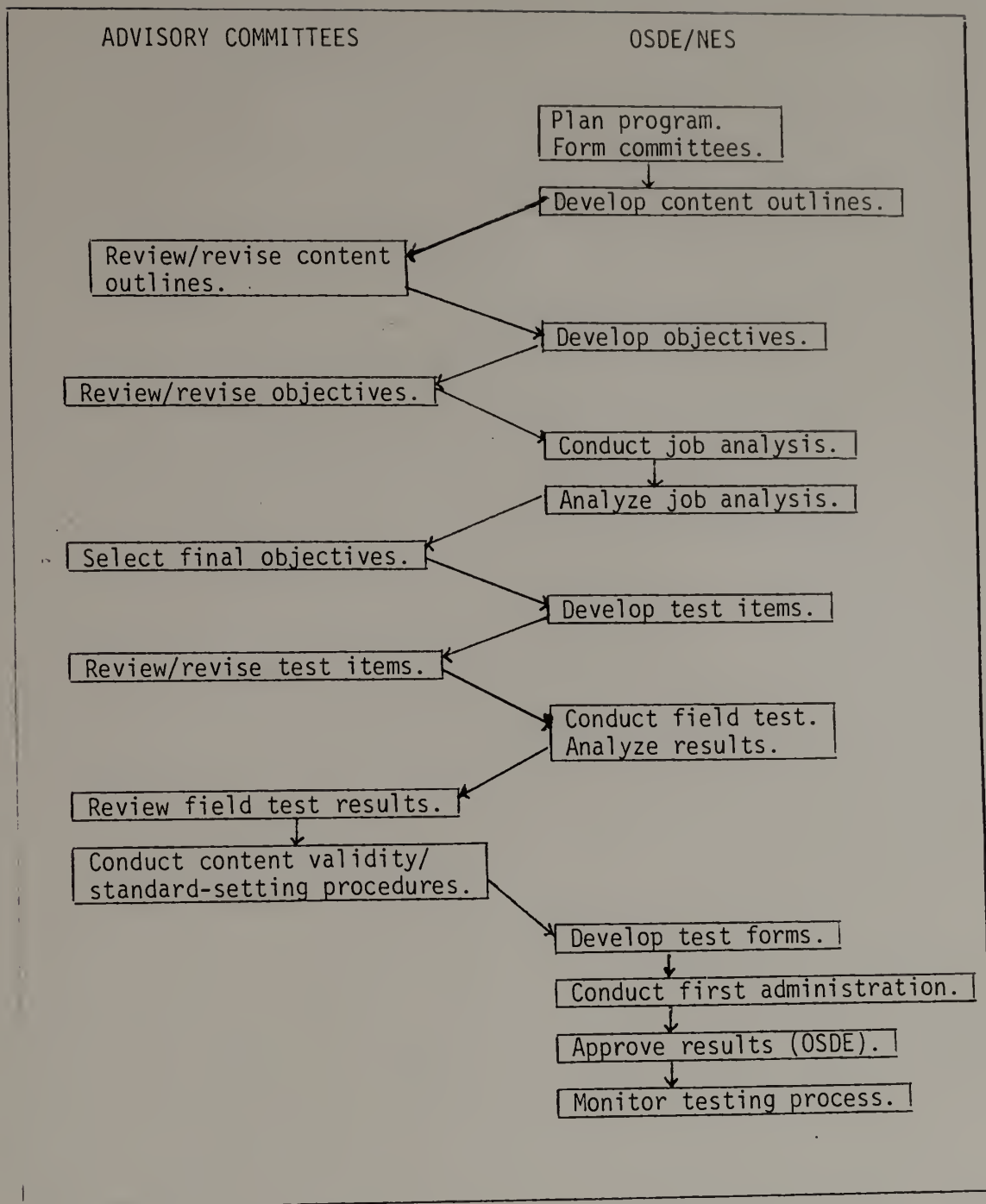


Figure 4.1. Test development process.

When the outlines were completed, NES staff met with the advisory committees. The committees were given instructions and criteria for reviewing the content outlines. The committees were instructed to review the entire outline for comprehensiveness. They were also instructed to review each element of the outline for accuracy, and to make revisions wherever necessary. Each of the 79 outlines received some modification. Committees discussed issues of accuracy, comprehensiveness, relative weight (e.g., should a heading be an "A" or a "1"), and clarity. Specific topics were deleted, added, or revised. In some cases, committees undertook a complete reorganization of the outline. At the end of the two-day conference, all 79 outlines were approved.

### Objectives

The second step in the development process was to convert the material from outline form into a set of measurable objectives. For each topic in the outline, identified by a capital letter, NES staff developed a behavioral objective. For a topic such as "causes of the Civil War," for example, the objective "Evaluate the causes of the Civil War (e.g., economic, social political)" might be appropriate. The objectives were grouped according to the major subdivisions within the test, and materials were prepared for advisory committee review. During the winter of 1981, the advisory committees met for a second two-day conference. As was the case with the outline review, the committees were given instructions and criteria for reviewing the objectives. The instructions included procedures for reviewing the



entire set of objectives for completeness of domain coverage and proportionality, and for examining each individual objective for accuracy, significance, bias, clarity, and taxonomic level.

In other words, the complete set of objectives had to cover the entire domain of knowledge and had to place appropriate emphasis on the most important parts of the domain. The individual objectives had to be clear and accurate. An objective had to measure a significant aspect of the domain. It had to be free of any racial, sex, ethnic, or regional bias. And, finally, it had to require performance at the appropriate taxonomic level. The verb, as in "Identify the branches of the federal government," determines the taxonomic level. The taxonomy refers to a classification by B.S. Bloom (1956) of the levels of the thinking process required by performance objectives in the cognitive domain. The six levels -- knowledge, comprehension, application, analysis, synthesis, and evaluation -- are hierarchical, i.e., each cognitive level subsumes the preceding one. The committee members might decide that teacher candidates should be able to analyze the content of one topic, while they should be expected to only identify (knowledge level) the content of another topic.

The committees reviewed the materials during a two-day conference. As with the content outlines, the objective review process involved adding, deleting, and revising objectives. At the end of the second day the 79 sets of objectives were approved.

### Job analysis

The job analysis procedure for the Oklahoma Teacher Certification Testing program is based on a rating of the approved objectives for each of the fields. NES staff, working with the OSDE, developed a stratified random sampling procedure for identifying approximately 150 incumbent teachers in each test area. In low-incidence fields, the sample was the entire population. The stratified sample represented the population of Oklahoma teachers in terms of school size and geographic subdivisions within the state.

Each incumbent teacher in the sample received a survey booklet that contained an explanation of the project, instructions, demographic questions, the complete set of objectives, a machine-scorable answer sheet, and a business-reply return envelope. The survey booklet instructed the teacher to examine each objective and indicate if he or she used the objective in the process of his or her work. If the teacher indicated "No," he or she was instructed to go on to the next objective. If the teacher indicated "Yes," he or she was asked to rate the objective on two five-point scales. The first scale concerned frequency (how often was the objective employed?). The second scale concerned essentiality (how important was the objective to success as a teacher?). Teachers were also asked to provide general comments about the objectives.

The analysis of the results involved developing a mean response for each of the two scales, for each objective. A scatterplot, showing the relative positions of all objectives for the field, was then pro-

duced. Next, a grand mean and standard deviation was calculated for the set of objectives. Finally, the objectives were classified as preferred (greater than 1 SD above the grand mean), acceptable (1 SD below to 1 SD above the grand mean), and least job-related (greater than 1 SD below the grand mean). Figure 4.2 illustrates the analysis process.

#### Selection of final objectives

Following the job analysis, the results were presented to the advisory committees for consideration in selecting the set of objectives that would form the basis of each test. NES and OSDE staff worked with each committee and guided the members through an interpretation of the data from the job analysis. Members reviewed the demographic characteristics of the sample for their field and read comments offered by sampled teachers. Then they examined the rating of objectives. This information was presented in the form of a scatterplot, a rank ordering, and descriptive statistics.

Each committee was given guidelines and criteria for selecting a specific number of objectives from all that appeared in the job analysis survey. A portion of the instructions was dedicated to the issue of proportionality. During the content outline and objective development phases of the project, the committee members had reviewed whole outlines and complete sets of objectives to assure that the entire domain was sampled and that the sampling was proportionate (i.e., the largest or most significant aspects of the domain received the greatest

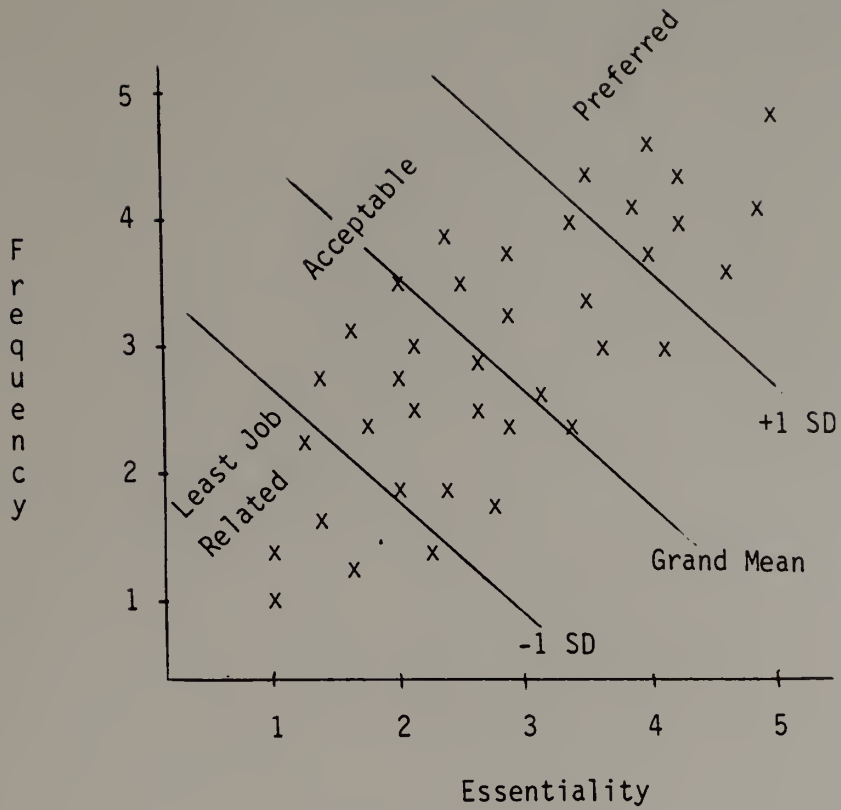


Figure 4.2. Job analysis results for a hypothetical field, showing the distribution and classification of objectives.

emphasis). This process resulted in the establishment of four to seven subareas within each general area test. The number of objectives within a particular subarea was an indication of the subarea's relative importance to the whole domain of knowledge. At the objective selection phase, the committees were given formulas for reducing the approximately 150 original objectives in each field to 50 objectives that would maintain the original proportionality. If, for example, one-fifth of an original 145 objectives, or 29, were found in subarea II of a particular field, then one-fifth of the final 50 objectives, or 10, should appear in that subarea.

A second important guideline concerned "emerging areas" within a field. It was explained to the committee members that the courts allow exceptions to a strict adherence to job analysis results when the content can be shown to concern a new or emerging area within the discipline. It may be, for example, that computers do not currently play an important role in the work of incumbent teachers, but that the state plans to make the use of computers an integral part of certain courses in the near future. In this case, there would be a compelling reason to want teacher candidates to have some knowledge of computers. The committee members were cautioned that this exception is just that: an exception. It was explained that emerging areas could not account for more than ten percent of a field, and that where they selected an objective that was rated low on the job analysis, they had to provide a rationale for why the data were "misleading."

At the end of the objective selection process, the committees had selected 50 objectives for each of the general tests, 40 objectives for each of the umbrella tests, and 25 objectives for each of the specific area tests. These selected objectives would form the basis for the item development phase of the project. Following the conference, all selected objectives plus all objectives not selected but ranked "preferred" or "acceptable" by the job analysis, were assembled into booklets. These booklets represented the range of job-related objectives, and they served as study guides for the examinees. While they could not identify which of the listed objectives would be measured by the test, students could study the whole set of objectives and be well prepared, not only for the examination, but also for the demands of their careers. These booklets were copied and distributed to schools of education throughout Oklahoma.

#### Item development

Based on the results of the job analysis survey and the objective selection process, NES staff developed a set of multiple-choice questions to measure each objective. The number of items written for each objective varied from one to four, depending on the objective's relative standing in the job analysis survey results.

The drafted item banks were taken to the advisory committees for their review during the fall of 1981. The committees were given instructions and criteria for review and, where necessary, revision of each item. The review criteria included an examination for "goodness

of fit," or "item/objective congruence." This refers to an effort to determine how well the particular item measured the objective to which it was matched. The item/objective congruence was assessed by having the committee members look at the principal verb in the objective (e.g., identify, evaluate, analyze) and determine whether or not the item was measuring the cognitive level represented by the verb. In addition, the committee members were asked to look at any specifications given in the objective. For example, if an objective said "Calculate the decimal equivalent of a fraction to the nearest hundredth," then the item matched to this objective would have to require the basic calculation to the nearest hundredth.

The committee members were also instructed to review each item for accuracy. Each item had to have one, and only one, correct answer. The distractors could not be correct or partially correct, but had to be plausible responses. All information contained in the directions, stimulus material, stem, or alternatives for an item had to be accurate.

The committee members also reviewed the set of items matched to an objective for the quality of their domain coverage. As the objectives for a field of knowledge had to sample the entire domain of knowledge, so the items selected to measure a particular objective had to sample the entire domain of the objective. And the item had to measure a significant aspect of that domain. The committee members had to ask themselves this question: "If we had only one item for this objective, would this item be significant enough to stand by itself as an item representative of the objective's content domain?" If the answer was

yes, then the item was significant.

All items were also carefully reviewed for any form of racial, sex, ethnic, or regional bias.

As with previous conferences, committee members performed their tasks vigorously. Parts of many test items were revised and, in some cases, whole new sets of items were written for some objectives. At the end of the two-day conference, the committees approved 79 item banks. Each item bank contained approximately 20-25% more items than the final test forms would require.

#### Field test

The next step in the development process was to submit the item banks to a field test. NES and OSDE staff worked with the various schools of education throughout Oklahoma to identify juniors and seniors who would be willing to participate in the field test of their particular discipline. In other words, future math teachers would help to field test the math items, and future physical education teachers would participate in the field test of the physical education test items. The incentive for participation was the fact that participation would provide a good exposure to the form and general content of the actual test. For many students, participation served the same purpose as taking a test preparation course.

For the most fields, between 50 and 100 students took part in the field test. The attempt was made to select students equally from each of the participating institutions. All subjects used machine-scorable



answer sheets that included standard demographic information such as sex, race, and region within the state.

Following the field test, an item statistics analysis was run for each set of answer sheets. The data were analyzed for reliability (Kuder-Richardson KR-20), item/objective correlation, difficulty (P-value), and distribution of response choices.

#### Content validation

Following the field test, NES and OSDE staff identified members of new committees charged with the responsibility of determining the content validity of the test items and setting the standards for the tests. The content validity and standard-setting committees were selected using the same criteria that were applied to the original advisory committee selection process. That is, the members were college professors, public school teachers, and OSDE curriculum experts, and they represented the race, sex, and geographic distribution of the total educator population. Each of the 79 fields was represented by approximately 10 people. Since some people were able to represent more than one field, the total number of committee members was between 300 and 400, rather than 790.

The conference opened with a review of the project to date. A summary of the advisory committee work, the job analysis, and the field test was presented. The committee members were then given instructions and criteria for the content validity process. The members were given sets of the selected objectives and the field test items, and were in-

structed to review each item for item/objective congruence, accuracy, significance, domain coverage, clarity, and bias. They were also instructed on how to use the statistical analysis of the field test results during their deliberations. The committee members' responses were recorded on machine-scorable answer sheets. Each member indicated, first, whether or not he or she was familiar with the content of the item. If he or she was not, the member so indicated and proceeded to the next item. If he or she was familiar with the content, then the member was instructed to apply the criteria and indicate that the item was valid or not valid.

#### Standard setting

The standard-setting process for the tests was conducted at the same time as the content validation procedure. The standard-setting procedure selected for the Oklahoma Teacher Certification Program was the Angoff (1971) method. Committee members were instructed to re-read the first test item that they had judged content valid. They were then instructed to imagine a group of 100 minimally competent teachers and to estimate how many of the hypothetical teachers could answer the item correctly (i.e., what percentage of minimally competent teachers?). Committee members indicated their estimates on a machine-scorable answer sheet.

Analysis for the content validity and standard-setting portions of the project consisted primarily of summing and averaging the data. An item was judged to be valid if there was significant agreement between

the judges. The significance of agreement was determined by comparing the number of "yes" responses with the probability tables for the binomial distribution. "Unfamiliar with content" ratings were not used in calculating any item, so that dichotomous ratings with different numbers of judges were generated. If the probability of receiving a given number of "yes" ratings (i.e., appropriate for minimum competence) was less than a chance of one in ten, the item was classified as an appropriate requirement for minimum competence. All items thus rated, along with their Angoff ratings, were stored on computer tapes for future test form assembly. All remaining test items were discarded.

#### Test form development

Working with an item bank that contained only content-valid items, NES staff assembled the 79 test forms for the first administration. For each field, a detailed test blueprint was developed. The blueprint indicated the number of items to be selected for each objective within the test. It also detailed rationales for selecting items of varying degrees of difficulty based on the assigned Angoff ratings.

In addition to the assembling of the 79 test forms, NES and OSDE staff developed and documented a detailed set of procedures for administering and scoring the tests.

#### First administration

The first administration of the Oklahoma Teacher Certification Testing Program took place in January 1982, at six sites throughout

Oklahoma. The administrations were conducted by NES and OSDE staff. Each of the tests was untimed. That is, an examinee could have as much time as he or she needed to complete the test. The average testing time varied from approximately one hour (specific area test) to two-and-one-half hours (general test). There were four four-hour testing sessions over a two-day period. If a student was applying for certification in all areas of business education or all branches of science, he or she had to take a total of 10 tests (one umbrella and nine specific area tests). As a result, a number of examinees attended all four sessions.

#### Standard error of measurement

Following the first administration, the test results were scored, and based on the Angoff rating of each scorable item (approximately 80% of the tests were scorable; 20% were experimental and did not contribute to the score), a preliminary cut score was produced for each test. The Angoff rating for each item was an average of the proportional ratings given by all judges. The ratings of the scorable items were averaged to produce the cut score for the test. Figure 4.3 illustrates this procedure.

The OSDE carefully reviewed all of the data for the first administration. Based on the performance of the initial examinees, and based on the interests and intents of the state certification process, the OSDE staff decided to lower the preliminary cut scores in all fields by 1 standard error of measurement (SEM). By so doing, they decreased the

All Content Valid Items		Items Selected for Administration		Scorable Items	
Item	Rating	Item	Rating	Item	Rating
1	.86	1	.86	1	Exp
2	.72				
3	.80	3	.80	3	.80
4	.47	4	.47	4	.47
5	.85				
6	.57	6	.57	6	.57
7	.60				
8	.83	8	.83	8	Exp
9	.70	9	.70	9	.70
10	.65	10	.65	10	.65
11	.84				
12	.27	12	.27	12	.27
13	.92				
14	.68				
15	.78	15	.78	15	Exp
16	.59				
17	.86	17	.86	17	.86
18	.81				
19	.79	19	.79	19	.79
20	.95	20	.95	20	.95
(Passing score = 81%, or 7.2 items)				AVERAGE .81	

Figure 4.3. Passing score determination based on hypothetical 20-item data bank.

likelihood of a type II error (i.e., failing to certify a competent but borderline candidate). NES staff then reanalyzed all data based on the revised cut scores and issued test result information to all examinees.

### Monitoring activities

Since the first administration, NES staff has continued to provide support services to the Oklahoma Teacher Certification Testing Program. These services have included the following.

Test updating. After a test form has been exposed to a prescribed number of examinees, a new test form is developed from the item bank. Test equating may occur after each administration for some high-incidence fields, or less than once a year for some low-incidence fields. The equating process is based on a combination of test form equating (i.e., the finalized forms are equivalent), and P-value equating (i.e., the new and old items are of equal difficulty).

Topicality review and job analysis. At prescribed intervals all sets of objectives and all item books are reviewed for topicality and accuracy. For this process, the OSDE again assembles advisory committees based on the same criteria used in the original advisory committee selection process. These committees review all materials to see if any content has become dated. For example, an objective on the school principal test might require familiarity with the details of a particular law. If a new law was subsequently passed, a revision to the objective and a new set of items might be required.

If the results of the topicality review reveal substantial changes in the field, plans are then developed for a new job analysis survey of that field.

Examinee preparation. In addition to distributing lists of all job-related objectives to schools of education throughout the state of Oklahoma, NES and OSDE staff have developed 79 study guides to help students prepare for the certification tests. These guides provide a listing of the objectives with sample questions for each objective. In addition, they provide general information about the form of the tests and about how to prepare for the tests.

C H A P T E R    V  
THE FUTURE OF TEACHER CERTIFICATION TESTING

Introduction

This paper has focused on three aspects of the teacher certification process: (1) Chapter II reviewed the legal parameters of the issue; (2) Chapter III explored the psychometrician's response to some of the legal problems; and (3) Chapter IV examined how a program was actually developed within a state. In order to ascertain the current state of affairs, and to explore probable future courses in these three areas, interviews were conducted with experts from the legal, psychometric, and programmatic areas.

The purpose of the surveys was to determine what is currently happening, and what is likely to happen in the immediate future. Because this type of insight is not likely to be garnered from a rigidly organized, broadly distributed survey, the decision was made to limit the number of people to be interviewed to those clearly identified as leaders in each area. Three individuals with a legal perspective, three with a psychometric expertise, and three with programmatic concerns were identified, and subsequently agreed to 30-to-45-minute telephone or personal interviews. The experts, all of whom were interviewed between January 15 and March 1, 1985, are as follows:



Legal

Michael Rebell, Esq.

Partner, Rebell, Kreiger, Fishbein, Olivieri

Mr. Rebell is Special Counsel to the New York State Assembly and has written extensively on teacher preparation and teacher credentialing.

Charles Coody, Esq.

Chief Counsel, Alabama State Department of Education

Mr. Coody has represented the Alabama State Board of Education and the Alabama State Department of Education since a class-action suit was filed against the teacher testing program in December 1981.

Dr. Bernard McKenna

Program Development Specialist, National Education Association

Dr. McKenna has studied legal issues at the NEA for the past 10 years. He has a particular interest in teacher certification testing.

Psychometric

Scott M. Elliot

Division Director, Licensing and Certification, National Evaluation Systems, Inc.

Mr. Elliot is responsible for all teacher certification tests being developed or administered by NES. Mr. Elliot's current projects include Alabama, Georgia, Oklahoma, West Virginia,

and Texas.

Dr. Ronald K. Hambleton

Professor, School of Education, University of Massachusetts,  
Amherst

Dr. Hambleton is an internationally-recognized authority on criterion-referenced tests used for measuring student or teacher achievement. He has written extensively on issues of reliability, validity, and standard setting.

Dr. Alan Seder

Project Director, California Teacher Certification Test,  
Educational Testing Service

Dr. Seder is responsible for the development of a criterion-referenced teacher certification test for the state of California. Dr. Seder is currently dealing with many of the issues raised in this paper.

#### Programmatic

Dr. Lester Soloman

Director, Georgia Teacher Certification Office, Georgia Department of Education

Dr. Soloman is one of the pioneers in criterion-referenced teacher certification. In his current position he supervised the development of the first such program in 1978.

Dr. C. C. Baker

Assistant Superintendent, Alabama State Department of Education

Dr. Baker not only supervised the development of the Alabama Initial Teacher Certification Testing program, but he also has been very much involved in analyzing this and similar programs in light of the pending Alabama litigation.

Dr. Joseph R. Weaver

Director, Teacher Education, Testing, and Staff Development,  
Oklahoma State Department of Education

Dr. Weaver supervised the development of the Oklahoma Teacher Certification Testing Program described in Chapter IV of this paper. He writes and speaks frequently on the programmatic issues involved in such a project.

In addition to restricting the number of individuals to be interviewed to those few at the forefront of this issue, the survey was also conducted in a manner that encouraged some flexibility of response. Each interviewee was encouraged to consider the six-to-eight questions as headings or starting points for discussion. The interviewees were encouraged to rephrase the question if that was expedient to providing a useful observation or piece of information. These two limitations -- restricting the sample to nine key experts, and encouraging them to reach beyond the question at hand -- proved worthwhile. The interviewees shared a number of insights that may be critical to practitioners now and in the future.

Responses to the three sets of questions are presented below.

### Legal Survey

1. A major legal (and moral) concern in the teacher certification process involves the collision of two basic rights: the right (or obligation) of the state to protect children from incompetent teachers, and the right of the teacher candidate to be protected from arbitrary and/or unfair employment practices. Do legal advisors to the state have a role in this issue?

All three respondents were somewhat uncomfortable with the wording of this question, and they objected to the implication that there was an inherent conflict. Mr. Coody pointed out that if the test was fair and rationally related to its purpose, both rights could be served. Mr. Rebell commented similarly, and then added that in this particular situation Title VII of the 1964 Civil Rights Act places the initial burden on the challenger (teacher candidate) to show that adverse impact has occurred. If adverse impact is indicated, the burden then shifts to the state and to the developers to show a compelling need for the test and to defend its technical merits.

Dr. McKenna added that this was an issue that was being dealt with in the courts when it should be handled by the public and by professional educators. He indicated that he would like to see tests like these developed under the auspices of the education profession. He offered the example that law boards are developed and controlled by the legal

profession. He felt that if the profession took responsibility for certification, with appropriate public input, the matter might never go to court.

2. Even after all efforts to eliminate or reduce bias have been undertaken, the pass rates for minorities are substantially lower than for the general population. Does the legal advisor have a role in dealing with adverse impact when it arises?

Mr. Coody indicated that it was "not the role of the court to make policy decisions." He said that this "might be appropriate if there has been a history of discrimination, and if using [a teacher certification test] would perpetuate discrimination."

Mr. Rebell saw evidence of movement in two opposite directions on this issue. He pointed out that Congress, through Title VII, had made adverse impact an issue, and that the scrutiny has been getting tougher over the years, in many cases "scaring off users." At the same time, though, he pointed out that "standards are looser and EEOC is talking about changes" that would make it easier for test users. He felt that one direction or the other would begin to prevail in the near future.

Taking a somewhat different approach, Dr. McKenna suggested that there should be a restriction such that the "courts only adjudicate differences." He was concerned that the courts are getting into areas where they are not experts. As an example, he offered the fact that there is a "judge writing curriculum in West Virginia" as a result of a court decision. The real solution to the problem of adverse impact, he

suggested, was "remediation, financial aid, and increased opportunity" for all those who aspire to teach.

3. Concerning the application of strict scrutiny vs. rational relationship to issues in employment of school personnel, where do the courts stand now? How will this change in the future? What will this mean for the developers and users (states) of tests?

Mr. Rebell felt that the courts are moving away from strict scrutiny and toward more relaxed standards. He offered *Washington v. Davis*<sup>1</sup> (see Chapter II) as an example. He noted an exception to this, however, in *Connecticut v. Teal*,<sup>2</sup> where the state tried to claim that "as long as the test as a whole did not have adverse impact, they [the users] did not have to defend the parts [of the test] ." The Court, however, found that each part of the testing program had to meet established standards.

Dr. McKenna also felt that the courts are returning to a rational relationship approach to employment testing.

Mr. Coody felt that the courts have found, and will stick to, a "middle ground" between strict scrutiny and rational relationship. As an example, he offered *U.S. v. South Carolina*,<sup>3</sup> where the state was obliged to show a rational relationship between the testing program and the state's legitimate needs, but where they were also required to show that the test was technically sound (i.e., valid and reliable).

<sup>1</sup>*Washington v. Davis*, 426 U.S. 250 (1976).

<sup>2</sup>*Connecticut v. Teal*, 457 U.S. 440 (1982).

<sup>3</sup>*United States v. South Carolina*, 445 F.Supp. 1094 (D.S.C. 1977).

4. What do states need to make the certification process in general, and the testing of teachers in particular, more legally defensible?

All three experts stressed the need to use multiple criteria in the certification process. Dr. McKenna suggested increased reliance on practicum evaluation. Mr. Rebell emphasized the need for a careful validation process focused on the content domain, and a thorough job analysis. Mr. Coody suggested attention to minority test results to reduce bias, and the use in the development process of independent panels to maintain objectivity.

5. In order to address apparent inequities and potential legal problems in the teacher certification process, states have developed more detailed and more rigorous requirements. Is it likely that these new standards will cause the courts to demand more from the states in future legal challenges?

The three experts felt that it was likely that this would occur. Mr. Rebell indicated that it is a part of basic due process that "if states adopt new standards, they will be held accountable." Mr. Coody cautioned, however, that this situation may vary, depending on the litigation. Dr. McKenna again pointed out that the courts "are not there to set standards," and that "the profession should set these standards."

6. What legal issues do you envision arising in future teacher certification litigation?

Both Mr. Rebell and Mr. Coody suggested that future litigation will probably involve the role of state-set curriculum standards in the teacher preparation institutions. They pointed out that, based on recent litigation at the secondary level, a state that held a student accountable for mastering a certain level of competence, was legally obliged to present the student with a clear explanation of the standards and a sufficient opportunity to prepare. One question that may be raised in future teacher certification litigation is "How well has the state met its obligation to help the teacher education institution and the student prepare for the test?" In future cases, states may be held liable for the instructional validity of their standards.

#### Psychometrist Survey

1. In the last five years, a number of states have added a testing component to their certification process. Do you think that this trend will continue?

There was general agreement that there would continue to be interest in teacher testing because of the rising demand for accountability and the concerns raised in A Nation at Risk (National Commission on Excellence in Education, 1983). Mr. Elliot suggested that "the pattern will follow the trend-setters like Texas, California, New York, and Florida." Dr. Hambleton cautioned that "we don't yet know how good these approaches are because we are still in a period of development."



He pointed out that we are still trying to develop measures of teacher competence and that we are still defining methods such as one-day simulations for performance testing.

2. Many states have developed customized, criterion-referenced tests, and Educational Testing Service has revised the National Teacher Examinations. Will CRTs push out NRTs (or vice versa)?

All of the respondents were in agreement on the idea that a balance would be struck. Some states would be interested in the diagnostic advantages of the CRT, while others may want to be able to make the national comparisons that an NRT allows. Cost was also mentioned as a factor because the CRTs require a costly development process. Dr. Hambleton said that he would like to see the new TCTs begin to establish some predictive validity. "We need, ultimately," he felt, "to be able to match a good teacher with a certain level of performance on the test." Mr. Elliot offered a list of benefits of the NRT and the CRT:

NRT - inexpensive

requires no development time

allows national comparisons

CRT - establishes a clear threshold standard

job-related

customized to the needs of state

involves groups within the state

more legally defensible

provides diagnostic and prescriptive information

3. What are state departments of education asking of psychometrists?

Dr. Seder said that he had noticed an increased concern with face validity and requests for help in addressing political concerns from legislators and unions. Dr. Hambleton cited new interest in critiquing existing practices and in researching new methods of item writing, validation, and analysis of technical data. Mr. Elliot added "the ability to match the test to various certification requirements surrounding certain subject areas, more diagnostic information, item response theory, multiple validity procedures, and more state department involvement and control" to the above list.

4. As a psychometrist, what do you see as the necessary next steps in research/development?

Both Dr. Seder and Dr. Hambleton predicted an increased emphasis on identifying the skills that define a good teacher. Dr. Hambleton and Mr. Elliot also mentioned clear validity, job analysis, and standard-setting procedures. Dr. Hambleton also saw a future for computer-assisted testing, performance testing, video-disk technology, and teacher assessment centers similar to those employed in business and industry.

5. Validity studies for TCTs have thus far relied primarily on content validity, and the courts have so far accepted this because of the difficulty of establishing predictive validity. Do you think that this will continue?

Mr. Elliot said that one can "make a good case that predictive validity is not an appropriate form for certification testing" and that "content validity is the appropriate method for determining validity in a test that has a minimum standard." Dr. Seder indicated that the courts have, thus far, been reluctant to demand predictive validity because of the lack of a clear dependent measure. He felt that with new research, the courts may want to see new forms of validity.

6. Which technical areas, in addition to validity, need to be addressed in the immediate future? What do you think will/should happen? What will be the result of these changes?

Dr. Seder suggested that "technical considerations are not going to come into play until we define the job of teaching." Mr. Elliot felt a number of concerns were likely to be addressed soon, including "minimizing error around the cut score, more use of item response theory for item selection, improved job analysis techniques, and improved methods for dealing with low-incidence fields." He also said that there are a number of questions that need to be addressed, including these: "What should a teacher education program do?" "What does a teacher do?" and "How does a particular trait improve education?"

7. A major legal (and moral) concern in the teacher certification process involves the collision of two basic rights: the right (or obligation) of the state to protect children from incompetent teachers, and the right of the teacher candidate to be protected from arbitrary and/or unfair employment practices. Does the psychometrist have a role in this

issue?

Both Dr. Seder and Mr. Elliot stated that the psychometrist had a very limited role in this issue. Dr. Seder noted that "not everyone has a right to serve in a profession, only to be judged fairly," and Mr. Elliot stated that "the psychometrist can only make sure that the test is job-related, properly prepared, and assessed in a fair and standardized manner." Dr. Hambleton raised the issue that the legal experts had brought up about this question -- that there does not have to be a conflict. He stated that "any assessment should be fair, but also of a quality to identify poor teachers -- both are important and possible. Tests do not have to be discriminatory."

8. Even after all efforts to eliminate or reduce bias have been undertaken, the pass rates for minorities are substantially lower than for the general population. Does the psychometrist have a role in dealing with the adverse impact?

Mr. Elliot stated that "the psychometrist has a responsibility to develop the strongest possible tests for bias, but bias is subjective and not always directly measurable. The psychometrist must give full attention to trying to identify bias, but he or she is not responsible for resolving all bias." Dr. Hambleton expressed a similar idea and said that there was a "serious problem, but not with the tests. Women tend to be shorter than men, but we don't revise the ruler. There is unequal education, and we need to improve the quality of education. The test is only the messenger."

State Department Survey

1. What was the impetus for adding a changing the test component to your certification requirements (legislature, state board of education, etc.)?

Dr. Solomon: "A statewide task force, the superintendent of schools, and the commissioner of education. The issue was accountability. We had had a long-range assessment of educational needs in Georgia, and the assessment recommended a performance-based certification process."

Dr. Baker: "In the mid-1970s, the State Board of Education revised the standards for completing a teacher education program of study. The certification test developed from the standards concerning the exit process for institutions."

Dr. Weaver: "The Oklahoma legislature passed H.B. 1706 under pressure from the teaching profession. The concerns were public accountability and adequate pay. Also, college deans wanted to improve the quality of teachers."

2. Why did you choose your particular approach to teacher testing (NTE or customized CRT)?

Dr. Solomon: "It was important that the test be customized to Georgia's needs. We needed objectives before the test and diagnostic information after it."

Dr. Baker: "The test needed to work for Alabama. After some research, we decided that the CRT was more fair and equitable."

Dr. Weaver: "We were not interested in national norms. We wanted to make sure that it met Oklahoma needs. We've had a philosophy that favors CRTs because we value the diagnostic information. We want to compare people to a standard, not each other."

3. How satisfied are you with the design and implementation of the program?

Dr. Solomon: "We're very satisfied with both the tests and our own performance assessments. We have 28 fields now, including leadership and service fields. The retake scores based on the performance of candidates who take the test more than once are getting higher. We now have a 67 percent retake pass rate. The Board of Regents is now putting an institution on probation if their overall pass rate falls below 70 percent, but even before this, institutions were making progress."

Dr. Baker: "We are pioneers in this movement. We like the CRT because it allows us to make adjustments at critical points in the process. We couldn't have developed a better system, and it's legally defensible."

Dr. Weaver: "It meets our needs well. We are accomplishing about everything we wanted. We are screening out some candidates, and the system assures that we meet standards."

4. Are there particular areas such as validity, administration,

or scoring that you think need more attention from the test developers? What are they?

Dr. Solomon said that he would like to see the number of items per objective expanded, a larger item pool, and better subarea reliability by collapsing small subareas. He also saw two sets of concerns: his breadth concerns included "a need for a basic skills test, a professional pedagogy test, and a general knowledge test"; his depth concerns were for a career ladder test. "Do you develop new higher-level tests or do you use the existing subject area tests using higher-level objectives or a newly-validated higher cut score?" Dr. Baker emphasized the need to improve multiple performance measures. Dr. Weaver said that he was concerned with possible adverse impact. "Although we couldn't have done anything differently, there is a low performance for minorities." Dr. Weaver pointed out that the state has been working with the colleges and universities to assure that the tests are valid and reliable. Since the program began Dr. Weaver has worked with the test developers (NES) to "change the composition of some tests, lengthen other tests, develop new tests for health education, accommodate students applying for 'minor' or 'major' certificates, and convert to a new system of certification endorsements."

5. A major legal (and moral) concern in the teacher certification process involves the collision of two basic rights: the right (or obligation) of the state to protect children from incompetent teachers, and the right of the teacher candidate to be protected from arbitrary and/or

unfair labor practices. What role or responsibility does the state have in this issue?

All three state department heads agreed that the state's responsibility is to assure that quality instruction occurs in the classroom. All three also agreed that the testing process had to be fair. According to their responses, fairness includes making the test job-related; limiting the test to specific subject-matter knowledge, rather than general knowledge; providing diagnostic information and objectives; allowing candidates to retake the exams; using multiple measures (e.g., grades, internship); and reducing the cut score by some reasonable standard error of measurement.

6. Even after all efforts to eliminate or reduce bias have been undertaken, the pass rates for minorities are substantially lower than for the general population as a whole. Does the state have a role in dealing with this issue?

In response to this question, all three state department heads went beyond the restatement that the tests had to be free of bias, and added that the state had to work with minorities and minority institutions to reduce adverse impact. Dr. Baker pointed to the two-year-old Tyson Amendment, which states a teacher education institution must maintain a certain overall pass rate in order to retain accreditation. He noted while some might think that this sounds punitive, the intent, and the way that the Department of Education is going about it, is to identify and help institutions that are not adequately preparing their



graduates. Dr. Baker said that the system is working, and that the situation is improving.

Dr. Weaver described a set of study guides that had been developed for, and with the cooperation of, the teacher education institutions in Oklahoma. He said that the Oklahoma State Department of Education has been working with the administrations of the minority institutions and that the joint efforts had produced a rise in minority pass rates from less than 40 percent, when the program began, to greater than 60 percent this year.

Dr. Solomon added that the opportunity to retake the tests an unlimited number of times provides an important means of combatting the adverse impact problem. In this system, which also exists in Oklahoma and Alabama, a candidate is never denied an opportunity for certification, and "allowing retake is a form of due process." Minority candidates who have not had equal access to education may fail the test the first time, but they are able to take the diagnostic information from their test results back to the institution and begin to make up identified deficiencies. Dr. Solomon pointed out that if one compares first-time test-takers, minorities have a significantly lower rate of passing the test, but that if one adds "retakes" into the analysis, the pass rates are almost even.

## Summary

### The issues

This paper has examined a number of the legal and psychometric forces that have guided the direction of the teacher certification testing movement over the past 20 years. The information has been organized and presented in an effort to help practitioners to understand the foundations of the present approaches to teacher testing, and to anticipate the challenges that test developers and licensing agencies are likely to face in the future.

Legal issues. Chapter II provided an overview of the major legal precedents that have been set between 1965 and 1985. The first major test of Title VII of the Civil Rights Act of 1964 came about in Griggs.<sup>4</sup> In this case, the U.S. Supreme Court ruled that the employer had failed to establish the job-relatedness of an employment test as required by Title VII. This case established the need to show the job-relatedness of a test, and it established the EEOC Guidelines as the source for judging the legal defensibility of employment tests. In Moody,<sup>5</sup> the Supreme Court reaffirmed the EEOC Guidelines as the test development benchmark, and in particular those sections that "draw upon and make reference to professional standards of test validation established by the American Psychological Association."<sup>6</sup>

<sup>4</sup>Griggs v. Duke Power Co., 401 U.S. 424 (1971).

<sup>5</sup>Albemarle Paper Co. v. Moody, 422 U.S. 405 (1975).

<sup>6</sup>422 U.S. 431 (1975).

Baker<sup>7</sup> represented the first Title VII challenge to a teacher examination. The district court of Mississippi concluded that the minimum composite score on the National Teacher Examinations of 1000 had been set in an arbitrary manner and bore no rational relationship to the classroom effectiveness of teachers. One outcome of this case was an increased attention to the use (e.g., passing score) of a test, as well as the professional development process.

The Chance<sup>8</sup> case moved the question of job-relatedness from the actual employment process to the area of certification. In Chance, the tests were found to have little relevance to the duties of the position (school supervisor), and an injunction was issued.

The significance of Davis<sup>9</sup> was the implication that Title VII was not always the relevant standard. The Supreme Court held that because the program that was using the test was not discriminatory (it was a "model" program), it did not have to adhere to the EEOC Guidelines but was free to follow the less demanding United States Civil Service standards. As a result, test developers now had not only to consider the test and its uses, but also to consider the user of the test.

The two direct challenges to teacher certification tests came about in North Carolina<sup>10</sup> and South Carolina.<sup>11</sup> The issues for the developers

---

<sup>8</sup>Baker v. Columbus Municipal Separate School District, 329 F.Supp. 706 (N.D. Miss. 1971).

<sup>9</sup>Chance v. Board of Examiners, 330 F.Supp. 203 (S.D.N.Y. 1971).

<sup>10</sup>Davis v. Washington, 348 Supp. 15 (D.C. 1972).

<sup>11</sup>U.S. v. State of North Carolina, Civil No. 4476 (E.D.N.C. 1975).

<sup>12</sup>U.S. v. State of South Carolina, 445 F.Supp. 1094 (D.S.C. 1977).

of teacher certification tests were clear. Where the use of the test was based on a proper validation study (South Carolina), the test was deemed fair and appropriate. Where the use and cutoff score were not validated (North Carolina), the tests were ruled illegal.

The most recent case presented here, Teal,<sup>12</sup> makes it clear that the "bottom line," that is, the percentage of minority members actually hired or promoted, cannot be used to justify a test that would, on its face, be judged unfair to minorities.

Taken together, the cases suggest that a test developer must:

- develop the test in accordance with professional standards;
- determine the job-relevance of the test; and
- determine the rational relationship of the use of the test (i.e., passing score) to the requirements of the job.

Psychometric issues. How have test developers attempted to address the legal concerns outlined in the landmark court cases? Chapter III presented a synopsis of the major psychometric advances that have occurred in the last 20 years. The job analysis procedures and post-development validation approaches implemented by National Evaluation Systems and Educational Testing Service exemplify the efforts that test developers are making to establish the job-relatedness of the tests. One area of this issue that has received direct attention from the courts and

<sup>12</sup>Connecticut v. Teal, 457 U.S. 440 (1982).

that has stimulated substantial research involves the identification of a passing score that is rationally related to the requirements of the position in question. The standard-setting procedures described in Chapter III offer an indication of the extent to which test developers are trying to identify practical and legally-defensible procedures for setting passing scores. Chapter III also showed that other technical areas such as reliability and error estimation have also been re-examined and, in some cases, modified in light of the need to possibly defend the final test in a court of law. Each of these issues is briefly reviewed.

#### The future

Chapter V presented the highlights of nine interviews with legal experts, test developers, and test users. Based on these interviews, less structured discourses with other professionals, and the experiences derived from working on the development and administration of four teacher certification tests, it is possible to suggest a short list of issues that will dominate the field of teacher certification testing in the immediate future.

- \* The courts will remain one of the major forces in the process of test development. Their role and their concerns, however, may be quite different. They may move away from the close attention to technical issues, and give test developers and test users more latitude, as suggested in

Davis,<sup>13</sup> or they may become more embroiled in the specifics of test construction and administration, as evidenced by the amount of pre-trial research that both the litigants and the defendants in the upcoming Alabama case have felt obliged to amass. The next few cases should begin to indicate the direction of this movement.

- \* Job-relatedness will remain the key to the legally-defensible test, but due process will probably take on more importance. In future cases litigants may claim that they did not have a fair opportunity to prepare for the test. States may have to justify teacher education programs that a state approved, but that are not able to prepare students who can pass the tests. They may also have to show that students were given adequate warning that the test was to be required. They may point to publication of study objectives and liberal re-examination policies as state efforts to honor the test-takers' right of due process.
- \* States will continue to select both criterion-referenced and norm-referenced tests, depending on their needs and resources. Neither will supplant the other in the near future.

---

<sup>13</sup>Washington v. Davis, 426 U.S. 250 (1976).

- \* Validation procedures will continue to be at the forefront of the technical demands that are made of test developers. Reliable validation procedures for subject areas with only a few test-takers (e.g., Latin), and predictive validation techniques will be the focuses of most attention. The development of some techniques, however, will have to await agreement within the education profession on some fundamental concepts: What are the parameters that define good teaching? What are the characteristics that make a good teacher?
- \* The uses of teacher certification tests will probably expand in the future. Arkansas has just instituted a test for incumbent teachers, and other states are looking for tests that are appropriate to both entry-level and promotional considerations. In a legally-defensible program, however, the tests will continue to be only one of a number of indicators of performance.

One model. The model outlined in Chapter IV goes a long way toward meeting the needs identified in this dissertation. While particular technical aspects of the Oklahoma model may be revised and updated in the future, it is hard to imagine improvements to the overall approach employed in this test development effort. The designers and the state worked closely together to develop a system that involved many groups at each step of the process. In the final analysis, this test develop-

ment attitude that the content of the test must come from the involved parties may prove to be the strongest claim for legal defensibility.



## SELECTED BIBLIOGRAPHY

- Alter, J. Why teachers fail. Newsweek, September 24, 1984.
- American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. Standards for educational and psychological tests and manuals. Washington, D.C.: American Psychological Association, 1974.
- Berk, R. Criterion-referenced measurement: the state of the art. Baltimore: Johns Hopkins University Press, 1980.
- Bersoff, D. Testing and the law. American Psychologist, 1983, 36, 1047-1056.
- Block, A.R.; Rebell, Michael A. The assessment of occupational competence. 5. Competence assessment and the courts: an overview of the state and the law. Unpublished report to the National Institute of Education, 1980.
- Bloom, B. S. (ed.) Taxonomy of educational objectives: cognitive domain. New York: David McKay, 1956.
- Cook, L. L.; Eignor, D. R.; Hutten, L. R. Considerations in the application of latent trait theory to objectives-based criterion-referenced tests. Laboratory of psychometric and evaluative research report. Amherst, MA: School of Education, University of Massachusetts, 1979.
- Ebel, R. L. Content standard test scores. Educational and Psychological Measurement, 1962, 22, 15-25.
- Elliot, S. M. Teacher certification testing, technical challenges: part II. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1982.
- Equal Employment Opportunity Commission. Guidelines on Employee Selection Procedures. Washington, D.C.: Author, August 24, 1966. (29 C.F.R., 1607).
- Equal Employment Opportunity Commission, Civil Service Commission, Department of Labor, and Department of Justice. Adoption by four agencies of uniform guidelines on employee selection procedures. Federal Register, 1978, 43, 38290-38315.
- Hambleton, R. K. Test score validity and standard-setting methods. In R. D. Berk (ed.) Criterion-referenced measurement: the state of the art. Baltimore, MD: Johns Hopkins University Press, 1980.

- Hambleton, R.; Eignor, D. A practitioner's guide to criterion-referenced test development, validation, and usage. Laboratory of psychometric and evaluative research report no. 70. Amherst, MA: School of Education, University of Massachusetts, 1979 (2nd ed.).
- Harward, L. E.; Hoetker, J. A brief review of recent court decisions related to the use of examinations for purposes of making personnel decisions. Unpublished addendum to materials prepared for the writing subtest of the Florida Teacher Competency Examination. Florida State Department of Education, 1979.
- Help! teacher can't teach! Time, June 16, 1980.
- Hively, W.; Patterson, H. L.; Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, 5, 275-90.
- Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-64.
- Jaeger, R. M. Measurement consequences of selected standard-setting models. Florida Journal of Educational Research, 1976, 5, 173-82.
- Kline, M. Why Johnny can't read. New York: St. Martin's Press, 1973.
- Leonard, G. Car pool. Esquire, May, 1983.
- Levitov, B. Licensing and accreditation in education: the law and the state interest. Lincoln, Nebraska: University of Nebraska, 1976.
- Maeroff, G. Questions on teachers' skills fuel debate over quality of education. New York Times, April 12, 1983.
- Marshall, J. L.; Haertel, E. H. The mean split-half coefficient of agreement: a single administration index of reliability for mastery tests. Manuscript, University of Wisconsin, 1976.
- Merz, W. R.; Grossen, N. E. An empirical investigation of six methods for examining test item bias (Final report grant NIE - 6-78-0067). Sacramento, CA: Foundation of California University, 1979.
- Montgomery, J. Can 'teach' teach? New York Times, May 14, 1979.
- Nassif, P. M. Standard-setting for criterion-referenced teacher licensing tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978.
- National Commission on Excellence in Education. A nation at risk: the imperative for educational reform. Washington, D.C.: U.S. Printing Office, 1983.

- National Commission on Excellence in Education. Teacher certification testing, technical challenges: part I. Paper presented at the annual meeting of the National Council on Measurement in Education, New York, 1982.
- National Evaluation Systems. The Georgia teaching field criterion-referenced test project: validation issues. Technical report, author, 1976.
- Nedelsky, L. 1954. Absolute grading standards for objective tests. Educational and Psychological Measurement, 1954, 14, 3-19.
- Novick, M. Federal guidelines and professional standards. American Psychologist, 1981, 36, 1036-1047.
- Nunnally, J. C., Psychometric Theory, New York: McGraw-Hill, 1978.
- Pittman, J. Actions taken by state departments of education in developing CBTE certification systems. Paper delivered at the Association of Teacher Educators Annual Conference, New Orleans, February, 1975.
- Popham, W. J. Educational Evaluation. Englewood Cliffs, N. J.: Prentice-Hall, 1975.
- Popham, J. Criterion-referenced measurement. Englewood Cliffs, N. J.: Prentice-Hall, 1978.
- Popham, J. Domain specification strategies. In Criterion-referenced measurement: the state of the art, ed. R. K. Berk. Baltimore, M.D.: Johns Hopkins University Press, 1980.
- Psychological Corporation. Summaries of Court Decisions on Employment Testing, 1968-1977; author, 1978.
- Rebell, M. The law, the courts, and teacher credentialing reform. In B. Levitov (ed.) Licensing and accreditation in education: the law and the state interest. Lincoln, Nebraska: University of Nebraska, 1976.
- Rubinstein, S.; McDonough, M.; Allan, R. The changing nature of teacher certification programs. Paper presented at the annual meeting of the American Educational Research Association, New York, 1982.
- Shimberg, B. Testing for licensing and certification. American Psychologist, 1981, 36, 1138-1146.
- Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-76.

- Subkoviak, M. J. Decision-consistency approaches. In R. D. Berk (ed.) Criterion-referenced measurement: the state of the art. Baltimore, M.D.: Johns Hopkins University Press, 1980.
- Swaminathan, H.; Hambleton, R. K.; Algina, J. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1984, 11, 263-67.
- Traub, J. Principals in action. Harper's, May, 1983.
- U. S. Department of Health, Education, and Welfare, Public Health Services. Credentialing health manpower (DHEW Publication No. 05 77-55057).

