1986

5. Minimum Competency Testing: Status and Potential

Ronald A. Berk
Johns Hopkins University, rberk1@jhu.edu

5 Minimum Competency Testing: Status and Potential

Ronald A. Berk
The Johns Hopkins University

INTRODUCTION

Competency becomes an issue when one seemingly encounters incompetence and its consequences. For example, suppose an automobile is taken to a dealer for brake repair. Once the repair has presumably been completed, the owner drives the car to the first intersection, one block from the dealer, and the brake light in the dashboard appears. This owner would probably begin to question the competence of the attending mechanic. As consumers, employers, or even students, we witness countless other examples of probable incompetence.

It is this questioning of competence that provided the impetus for the minimum competency testing movement in this country. The movement which began in the 1970s developed in two distinct but interrelated fields: education and occupational licensing. In education, the public seriously questioned the meaning of the high school diploma and, in essence, the competence of a high school graduate. Coterminously, thousands of consumer complaints against licensed and certified practitioners brought into question the quality of services rendered and the conditions for relicensure. Although many of the issues in education and licensure are quite similar, especially in regard to assessment, only the competency movement in education will be reviewed here in order to avoid redundancy in this chapter and with Kane's chapter in this volume.

Minimum Competence

Despite the recency of the minimum competency testing movement, the concept of minimum competence is not new. It has been an integral part of occupational licensing in the United States for more than 200 years. Licensing is "the process

by which an agency of government grants permission to an individual to engage in a given occupation upon finding that the applicant has attained the minimal degree of competency necessary 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). This concern for ''minimal competency'' or ''minimum qualifications'' for safe practice underlies the state regulation of more than 800 occupations and professions (Greene & Gay, 1980; Shimberg, 1981). Individuals seeking licensure, for example, physicians, pilots, electricians, lightning rod installers, or horseshoers (see Shimberg, 1982a, chap. 1), are usually required to pass an examination in order to demonstrate their competence. Shimberg (1982a) has listed three responsibilities of licensing boards using such an examination:

1. the examination is a satisfactory measure of competence;
2. it measures the critical and important knowledge, skills, and abilities prerequisite to performance of the job at the minimum level of competence deemed necessary for the public protection; and
3. it is capable of screening out those who lack the requisite level of competence. (p. 56)

The educational analogue of these characteristics will become apparent in subsequent sections of this chapter.

Another perspective on the concept considers minimum competence in the context of United States social policies. Cohen and Haney (1980) have pointed to the longstanding interest in having government promote minimum levels of social welfare. Examples include public health programs, social security, unemployment insurance, welfare programs, and, certainly, a free public education.

Throughout the relatively brief history of the competency testing movement the expression *minimum competence* has engendered a considerable amount of confusion among lay people and educators alike. In practice, it connotes both the type of competence to be measured and the performance standard that is specified to designate attainment of the competencies. A further discussion of this topic is given in the section on ''Definitions.''

Grass Roots Movement

Earlier, it was indicated that the origin of the competency testing movement in education was public questioning of the competence of high school graduates. Public support transformed into legislative action has also been the sustaining force behind its continuation. As Lerner (1981) observed, minimum competency testing is a genuine grass roots movement that is ''clearly being led, or pushed, by noneducators'' (Pipho, 1978, p. 586). To date, the 39 state mandates for minimum competency testing programs were instigated by either legislative action or state board of education action, not by professional educators. Since the first programs were mandated in 1971 (Florida and Georgia), the competency

testing movement has been viewed as an outgrowth of the increasing public clamor for accountability in the schools. Beard (1979) stated:

> [Minimum competency testing] has widespread popular appeal to citizens and politicians who see it as a way of holding schools accountable and forcing them "back to basics." These groups are convinced that the quality of public education has been eroded over a period of years and that high schools are graduating significant numbers of students who are unable to read and write, and consequently, unable to support themselves through gainful employment. (p. 9)

Several Gallup polls on public attitudes toward education over more than a decade furnish ample evidence of the massive public support for the movement (Gallup, 1978, 1979, 1980, 1981, 1982, 1983, 1984).

Evidence of Incompetence

One major purpose of minimum competency testing is to restore confidence in the high school diploma by requiring students to satisfy certain standards of basic competence. This focus stems from the accumulating evidence of incompetence in the 1970s. Declining SAT scores (College Entrance Examination Board, 1977) and increasing rates of illiteracy and semiliteracy among American teenagers (National Assessment of Educational Progress, 1976) led to widespread public disillusionment and dissatisfaction with the quality of the entire educational system. Complaints by employers that high school graduates were unable to complete job applications correctly were echoed with complaints by colleges and universities that the reading ability of a substantial number of incoming students was inadequate for college level work (Perkins, 1982), which necessitated the institution of remedial reading classes.

The 1983 report by the National Commission on Excellence in Education, titled "A Nation at Risk: The Imperative for Educational Reform," listed several indicators of risk that convey more dramatically the incompetence of American youth:

1. Some 73 million American adults are functionally illiterate by the simplest tests of everyday reading, writing and comprehension.
2. About 13% of all 17-year-olds in the United States can be considered functionally illiterate. Functional illiteracy among minority youth may run as high as 40%.
3. The College Board's Scholastic Aptitude Tests demonstrate a virtually unbroken decline from 1963 to 1980. Average verbal scores fell over 50 points and average mathematics scores dropped nearly 40 points.
4. Both the number and proportion of students demonstrating superior achievement on the SATs (i.e., those with scores of 650 or higher) have also dramatically declined.

5. Many 17-year-olds do not possess the "higher order" intellectual skills we should expect of them. Nearly 40% cannot draw inferences from written material; only one-fifth can write a persuasive essay; and only one-third can solve a mathematics problem requiring several steps.

6. Between 1975 and 1980, remedial mathematics courses in public four-year colleges increased by 72% and now constitute one-quarter of all mathematics courses taught in those institutions.

7. Business and military leaders complain that they are required to spend millions of dollars on costly remedial education and training programs in such basic skills as reading, writing, spelling, and computation. (pp. 8–9)

From these and many other indicators of risk cited in the report, the Commission concluded that more and more young people emerge from high school ready neither for college nor work. One recommendation was that state and local high school graduation requirements be strengthened and that, at a minimum, all students seeking a diploma be required to lay the foundations in the Five New Basics: English, mathematics, science, social studies, and computer science. The Commission stressed that whatever the student's educational or work objectives, knowledge of the New Basics is the foundation of success for the after-school years.

The pressing need for this re-emphasis on instruction and assessment of basic skills was expressed by the Commission:

> Individuals in our society who do not possess the levels of skill, literacy, and training essential to this new era will be effectively disenfranchised, not simply from the material rewards that accompany competent performance, but also from the chance to participate fully in our national life. (p. 7)

The seriousness of the consequences of incompetent high school graduates and meaningless diplomas was articulated by Lerner (1981):

> Functional literacy and/or numeracy is an essential prerequisite for the competent performance of almost all skilled jobs, blue-collar or white-collar, in the United States or in any other developed nation in the world today. (p. 1059)
>
> With or without diplomas, young Americans who leave school without basic skills face bleak futures. Some will manage to secure unskilled work on at least an intermittant basis. Many others will not, because without those basic skills, they are not just unemployed—they are for most practical purposes in today's economy, unemployable. (p. 1060)

Confronted with these facts, one must decide whether minimum competency testing programs can, at least, partially solve these educational problems or another approach will prove more effective. At present, there are no alternatives with the overwhelming public support accorded competency testing. More than

that, however, the testing technology exists and teaching and testing for competence (or mastery) have sound theoretical bases.

This chapter reviews the current status of minimum competency testing and the issues that must be addressed for its future success. Special attention is given to the most thorny technical problems in competency test construction and score analysis and use. Specific recommendations are also offered for the more promising measurement techniques.

CURRENT STATUS OF MINIMUM COMPETENCY TESTING

This section assesses the state of minimum competency testing practices in relation to three major topics: (1) definitions, (2) policy specifications, and, (3) pros and cons.

Definitions

The burgeoning literature on minimum competency testing over the past decade has produced numerous and diverse definitions of *competency, minimum competency, minimum competency test,* and *minimum competency testing program.* Although it is easy to conclude that "there is no consistent terminology for minimum competency testing in use in the testing field" (Gorth & Perkins, 1979a, p. 8), there are certain key characteristics of all of the testing programs in operation that render the differences between the most popular definitions as trivial. Several of these definitions are presented in Table 5.1. A close inspection of the definitions along with analyses of the results of a nationwide survey of 31 states and 20 local district minimum competency testing programs conducted by Gorth and Perkins (1979b) and of a similar survey by Pipho (1983) reveal the following common features:

1. There is an emphasis on the acquisition of minimum skills or competence, usually academic skills (e.g., reading, math, writing) and/or life skills (e.g., following directions, filling out a job application, balancing a checkbook)
2. An explicit performance standard for pass-fail decisions is set so that one can separate the competent from the incompetent
3. The test results are used to make important decisions about individual students such as a promotion to a higher grade (or retention at the same grade), awarding of a high school diploma or a certificate of special recognition (or awarding a certificate of school attendance), or assignment to remedial classes.

These features are reflected in the definition adopted by the widely publicized National Institute of Education sponsored adversary evaluation hearing on the topic held in Washington, D.C., in July, 1981:

TABLE 5.1
Definitions of Minimum Competency Testing (Listed Chronologically)

Source	Definition
Elford (1977)	Minimum competency testing involves: (1) the use of objective, criterion-referenced competency tests; (2) the assessment of reading and computation using "real life" or "life skill" items; (3) the requirement of a specialized mastery level for high school graduation; (4) the early introduction of such testing for purposes of identification and remediation.
American Friends Service Committee (1978)	[Minimum competency testing programs are] organized efforts to make sure public school students are able to demonstrate their mastery of certain *minimum* skills needed to perform tasks they will routinely confront in adult life.
Airasian, Pedulla, and Madaus (1978)	[Minimum competency testing is] a certification mechanism whereby a pupil must demonstrate that he/she has mastered certain minimal (sic) skills in order to receive a high school diploma.
Miller (1978)	Minimum competency tests are constructed to measure the acquisition of competency or skills to or beyond a certain defined standard.
National School Boards Association (1978)	[Minimum competency testing programs are] testing programs which attempt to learn whether each student is at least "minimally competent" by the time the student graduates from public school.
Beard (1979)	Minimum competency testing involves the administration of proficiency tests in order to certify that minimum competency or proficiency exists with regard to a well-defined set of knowledge or skills.
Cohen and Haney (1980)	Nearly all minimum competency testing programs seek to define minimum learning outcomes for students in a variety of academic areas and to insure that these standards are satisfied
Lerner (1981)	[T]he common core of all minimum competence programs is an insistence on the cardinal importance of the basics--reading, writing, and arithmetic--and an equal insistence on hard, objective test data to measure success or failure in the acquisition of those fundamental intellectual tools by school children of all races, classes, and backgrounds.

> Minimum competency testing refers to programs mandated by a state or local body which have the following characteristics: (1) All or almost all students of designated grades are required to take paper-and-pencil tests designed to measure basic academic skills, life or survival skills, or functional literacy; (2) a passing score or standard for acceptable levels of student performance has been established; and (3) test results may be used to certify students for grade promotion, graduation, or diploma award; to classify students in remedial or other special services; to allocate compensatory funds to districts; to evaluate or to certify schools or school districts; or to evaluate teachers.

Obviously the most meaningful and useful strategy for tackling the construct of minimum competency has been to define it operationally in terms of program characteristics. Although such an approach does not help clarify *competency* in contrast to knowledge or performance (see Chickering & Claxton, 1981; Gale & Pol, 1975; Klemp, 1979; Senior, 1976; Shimberg, 1982b), it does direct attention toward the most crucial elements of the problem so that solutions can at least be attempted. The polemics over "what is competence" will probably continue for decades.

Policy Specifications

In most states, the policy of a minimum competency testing program is initiated in one of two ways: legislative action or state board of education action. The legislature may pass one or more laws stipulating the components and requirements of the program and/or the agents responsible for implementation. Alternatively, the state board may be empowered by state law to pass a mandate to establish the program. In a few cases where neither the legislature nor state board instituted the testing program, the local board of education may pass a mandate to initiate a program in its district (e.g., Denver, CO, Gary, IN).

The Gorth and Perkins (1979b) survey of minimum competency testing programs cited previously gathered information on the policy specifications in 31 states. More recently, Pipho (1983) updated that information on 39 states. The type of initial mandate (legislature or state board), the date of that mandate, and the date of completed implementation are identified for each state in Table 5.2. State board action (26 states) was the more frequent source of mandate compared to legislative action (15 states). While most of the mandates were passed by 1979, the implementation of a majority of the programs (22 states) will not be completed until the 1980s.

In addition to the data on state mandates, Table 4.2 displays how the test results are to be used. Only 15 of the states require a student to pass the competency test in order to receive a high school diploma. Arizona, California, Florida, and Maryland also require satisfactory test performance to advance to a higher grade level. In Illinois and New Hampshire both uses of the test scores are optional. For students who pass the test in 20 states, either standard diplomas are

TABLE 5.2
Minimum Competency Testing Policy Specifications in 39 States[a]

State	Mandate			Use of Test Results[b]				
	Legislature	State Board	Date	Imple.	H.S. Diploma	Special Recog.	Grade Promotion	Remed.
Alabama		X	1977	1984	X	U		O
Arizona	X	X	1972/75	1976	X		X	
Arkansas	X		1979	1982				M
California	X		1976/81	1980	X		X	M
Colorado	X		1975		O			
Connecticut	X		1978	1980				M
Delaware		X	1978	1981	X			
Florida	X		1971/75/76	1983	X		X	M
Georgia		X	1971/76/77/78	1978				O
Idaho		X	1977	1982	O	X		
Illinois		X	1978	1980	O	O	O	O
Indiana		X	1978					O
Kansas	X		1978/81	1980				O
Kentucky	X		1978	1979				O
Louisiana	X		1979	1980			X	
Maine	X		1978	1980				O
Maryland		X	1976/77/78	1986	X		X	M
Massachusetts		X	1978	1981				O
Michigan		X	1969	1979				O
Missouri		X	1976/78	1979				O
Nebraska		X	1975	1975				O
Nevada	X		1977/79	1979	X			M
New Hampshire		X	1977	1984	O		O	
New Jersey	X	X	1976/79	1978/85	X			M
New Mexico		X	1976	1979		X		M
New York		X	1978/79	1981	X			M
North Carolina	X		1977	1979	X			M
Ohio		X	1982					
Oklahoma		X	1977	1978				
Oregon		X	1972/76/80	1978	X			O
Rhode Island		X	1978	1981				
South Carolina	X		1978	1985				M
Tennessee		X	1977/80/82	1982	X			M
Texas		X	1979	1978				M
Utah		X	1977	1980	X			M
Vermont		X	1977	1979	X			O
Virginia	X		1975	1980	X			M
Wisconsin	X		1982		O			M
Wyoming		X	1977/79	1980				M

[a]Sources: Adapted from Gorth and Perkins (1979b, p.2) and Pipho (1983) with permission of the authors.

[b]The specifications on the use of test results indicate characteristic of program (X), mandatory (M), optional/conditional (O), or undecided (U).

awarded or diplomas with an endorsement or certificate of competency achievement (special recognition) are given. In the 15 states that mandate the test for graduation, the students who fail receive only a certificate signifying completion of all other requirements (attendance or course credits). In Florida, Commissioner Ralph Turlington estimated that 1300 seniors across the state (about 2%) fell into this category in 1983 (Cody, 1983). Remediation for students who fail the test is mandatory in 17 states (including Florida) and optional in 13. Remediation may take the form of providing extra attention and special remedial materials in the regular classroom setting, remedial classes, or other methods to address the skill deficiencies.

Pros and Cons

From the characteristics of the testing programs and the movement described thus far, it should be apparent that minimum competency testing is politically motivated and educationally implemented. As such, it has become a hotly debated topic with growing numbers of proponents (e.g., Beard, 1979; Fisher, 1978; Fremer, 1978; Lerner, 1981; Popham, 1981a) and opponents (e.g., Airasian, Madaus, & Pedulla, 1979; Glass, 1978a, 1978b; Haney & Madaus, 1978; Jaeger, 1982; Lazarus, 1981; Linn, Madaus, & Pedulla, 1982; Madaus, 1981). The proponents argue for its potential benefits and the opponents argue about its potentially harmful effects. Since an extended discussion of pros and cons is beyond the scope of this chapter, only a brief summary of some of the major arguments is given. Interested readers should consult the above references and the transcripts or videotapes of the National Institute of Education hearing on minimum competency testing (National Institute of Education, 1981; Thurston & House, 1981) for a detailed account of the pro (Popham, 1981a) and con (Madaus, 1981) issues.

Perhaps the most up-to-date and comprehensive summary of arguments on both sides is the list of perceived benefits and perceived costs of minimum competency testing completed by Perkins (1982). It is reproduced here in Table 5.3. The benefits seem to be related to five key assertions (Gorth & Perkins, 1979a): "(1) restore confidence in the high school diploma, (2) involve the public in education, (3) improve teaching and learning, (4) serve a diagnostic, remedial function, and (5) provide a mechanism of accountability" (p. 12). The costs tend to concentrate on the harmful effects of the testing on students, teachers and administrators, the curriculum, and the control of education.

Inasmuch as the opposing arguments on minimum competency testing are irreconcilable at this time, although the representatives on each side are convinced that they are right, policy makers should weigh carefully the advantages and disadvantages and then decide for themselves what is the most appropriate course of action. The 50 arguments in Table 5.3 should help guide that informed decision.

TECHNICAL SPECIFICATIONS FOR MINIMUM COMPETENCY TESTS

Another area where "benefits and costs" may be applied is the minimum competence test itself and its technical specifications. There are several strategies now available for constructing a competency test and for assessing the validity and reliability of the scores. The testing technology is derived from the research on criterion-referenced measurement (Berk, 1980c, 1984b, 1984c; Hambleton, Swaminathan, Algina, & Coulson, 1978; Popham, 1978a).

TABLE 5.3[a]
Perceived Benefits and Costs of Minimum Competency Testing

Benefits	Costs
1. restores meaning to a high school diploma	1. emphasis on the practical will lead to an erosion of liberal education
2. reestablishes public confidence in the schools	2. causes less attention to be paid to difficult-to-measure learning outcomes
3. impels us to face squarely the question of "what is a high school education?"	3. promotes teaching to the test
4. sets meaningful standards for diploma award and grade promotion	4. will be the "deathknell for the inquiry approach to education"
5. challenges the validity of using seat time and course credits as basis for certifying student accomplishments	5. oversimplifies issues of defining competencies and standards and of granting credentials to students
6. certifies that students have specific minimum competencies	6. promotes confusion as to the meaning of the high school diploma when competency definition is left to local districts
7. involves the public and local educators in defining educational standards and goals	7. fails to adequately consider community disagreement over the nature and difficulty of competencies
8. focuses the resources of a school district on a clear set of goals	8. will exclude more children from schools and further stigmatize underachievers
9. defines more precisely what skills must be taught and learned for students, parents, and teachers	9. will cause "minimums" to become "maximums," thus failing to provide enough instructional challenge in school
10. promotes carefully organized teaching and carefully designed sequential learning	10. may unfairly label students and cause more of the "less able" to be retained
11. reemphasizes basic skills instruction	11. may cause an increase in dropouts, depending on the minimum that is set
12. helps promote competencies of life after school	12. provides no recognition of the "average" student

13. broadens educational alternatives and options	13. fails to provide alternatives that can "inspire" average students to excel in some areas
14. motivates students to master basic reading, mathematics, and writing skills	14. ignores the special needs of gifted students, giving them less opportunity to be challenged and to expand their horizons
15. stimulates teachers and students to put forth their best efforts	15. may have adverse impact on a student's future career as a result of a withheld diploma
16. identifies students lacking basic skills at an early stage	16. may promote bias against racial, ethnic, and/or special needs groups
17. encourages revision of courses to correct identifies skill deficiencies	17. places the burden of "failure" on the student
18. ensures that schools help those students who have the greatest educational need	18. causes educators to be held unfairly accountable
19. can bring about cohesiveness in teacher training	19. intensifies the conflict for educators between humaneness and accountability
20. can truly individualize instruction	20. increases the record-keeping burden for administrators
21. shifts priorities from process to product	21. does not assure that students will receive effective remediation
22. holds schools accountable for educational process	22. does not assure that all the perceived needs and benefits will be met and realized
23. furnishes information to the public about performance of educational institutions	23. promotes the power of the state at the expense of local district autonomy
24. provides students with an opportunity to remedy the effects of discrimination	24. can be costly, expecially where implementation and remediation are concerned
25. provides greater holding power for students in the senior year	
26. provides for easier allocation of resources	

[a]Source: Reprinted with permission of the author and the National Council on Measurement in Education from Perkins, Marcy. "Minimum Competency Testing: What? Why? Why Not" *Educational Measurement: Issues and Practice*, Winter, 1982, pp. 5-9 (Tables II and III, pp. 7-8). Copyright 1982, National Council on Measurement in Education, Washington, DC.

Before proceeding with an examination of the pertinent technical topics, the following definition is proposed:

> *A minimum competency test is designed to determine a student's performance level with respect to a well-defined domain of competencies and prespecified performance standard.*

Two elements in this definition, competencies and standards, are consistent with most definitions of minimum competency testing (see Table 5.1) and the common features of existing programs which were described previously. In addition, the definition focuses on an individual student's score. The uses of the score for decisions about graduation or promotion are reflected in the types of validity and reliability evidence gathered as the test is developed.

Technical Standards for Competency Test Construction and Score Use

For the construction and use of norm-referenced tests, the *Standards for Educational and Psychological Tests* (APA/AERA/NCME Joint Committee, 1974) has served as the guide and, in essence, the "bible" of acceptable measurement practices. No single set of standards established by a joint committee of national experts is available for minimum competency tests. While it is possible to search through the *Standards* and glean some standards relevant to competency or certification tests, the product of this effort will be far from adequate. The fourth draft (February 1984) of the next edition of the *Standards,* titled the *Joint Technical Standards for Educational and Psychological Testing,* which is to be published in 1985, suggests that a separate section will be devoted exclusively to standards for certification testing in elementary and secondary education.[1] These standards and the collection of standards devised by Hambleton (1982; Hambleton & Eignor, 1978) for criterion-referenced tests will provide the foundation for the technical specifications and issues discussed hereafter.

Among the steps in the development of a minimum competency test (see Hambleton & Eignor, 1980), five are particularly troublesome: (1) defining the domain of competencies, (2) setting the performance standard, (3) gathering appropriate validity evidence, (4) estimating the reliability of the scores and decisions, and (5) equating the scores on different test forms. They are troublesome because there are many methods or statistical procedures one can use at each step and there is no consensus on any best method. Reviews of these technical areas by Hambleton and Eignor (1980), Shepard (1980b), and Jaeger (in press) furnish a few guidelines that may be helpful. Their recommendations will be integrated into this presentation.

[1]There will also be a section on professional and occupational licensure and certification.

Defining the Domain of Competencies

The first step in the construction of a minimum competency test is the specification of what the test is to measure. If one cannot define clearly what the test measures, then the resulting scores will be virtually meaningless. From the standpoint of score interpretation, a score *must* be referenced to the domain of competencies prior to any other type of referencing. For example, a teacher might say that Joanna's score on the test tells that she acquired 80% of the functional reading skills in the areas of survival signs, directional vocabulary, map symbols, and simple forms.

Academic Skills or Life Skills

The types of competencies typically measured by minimum competency tests are academic skills and/or life skills. The academic skills are those which have been traditionally taught in school, usually reading, mathematics, and writing. The life or survival skills often involve the transfer and application of the academic skills to practical "life-like" situations. In the preceding example, a simple form such as a bicycle registration form might be used to test the application of reading skills to a real situation that the student would encounter outside of the school environment.

The Gorth and Perkins (1979b) and Pipho (1983) surveys indicated that among 39 states, 18 states assess both academic and life skills in their minimum competency testing programs. The results are summarized in Table 5.4. All of the other states, with the exception of Georgia, emphasize academic skills only. In other words, almost every state includes academic skills and almost half of the states also test life skills. As shown in the table, the primary subject area coverage of the academic skills is the basic skills or "three Rs." Only six states have measured speaking and/or listening skills. Just what skills comprise the domain of competencies is determined most often by a special state level committee. Parents and citiziens either serve as members of this committee or are surveyed for their reaction to the domain definition.

Traditional Approach

The rigor and precision with which the domain is defined can enhance or diminish the score interpretation. The interpretation may be vague or explicit. Since the 1970s, the leading proponents of criterion-referenced and minimum competency tests have argued that the traditional approach to defining a content domain, which includes a content outline, a list of objectives, and a table of specifications or similar "blueprint," tends to provide an *ambiguous* domain definition. The arguments focus on the subjectivity involved in composing those specifications. That is, the selection of competencies and objectives is quite arbitrary, typically representing only one conceptualization of the domain, that adopted by the state level committee. Such specifications would be open to different interpretations by different teachers, administrators, students, and par-

TABLE 5.4
Types of Competencies Measured by Minimum Competency Testing Programs in 39 States[a]

	Emphasis			Subject Area[b]					
State	Academic Skills	Life Skills	Both	Reading	Math	Writing	Speaking	Listening	Other
Alabama			X	X	X	X			
Arizona	X			X	X	X			
Arkansas	X			X	X				
Calif.			X	X	X	X			X
Colorado			O	O	O	O			
Conn.			X	X	X	X			X
Delaware	X			X	X	X			
Florida			X	X	X	X			
Georgia		X		X	X				X
Idaho	X			X	X	X			X
Illinois			X	X	X	X	X	X	X
Indiana	X			X	X	X			X
Kansas			X	X	X				
Kentucky	X			X	X	X			X
Louisiana	X			X	X	X			X
Maine	X			X	X	X			
Maryland			X	X	X	X			X
Mass.			X	X	X	X	X	X	
Michigan	X			X	X	X		X	X
Missouri			X	X	X				X
Nebraska	X			X	X	X			O
Nevada	X			X	X	X			
New Hamp.	X			X	X	X			X
New Jer.	X			X	X	X			
New Mex.			X	X	X	X			
New York	X			X	X	X			
N. Caro.			X	X	X				
Ohio	X			X	X	X			
Oklahoma			X	X					
Oregon			X	X	X	X	X	X	X
Rhode Is.			X	X	X				
S. Carol.			X	X	X	X			X
Tennessee	X			X	X				X
Texas	X			X	X	X			X
Utah			X	X	X	X	X	X	X
Vermont			X	X	X	X	X	X	X
Virginia			X	X	X				X
Wisconsin	X			X	X				X
Wyoming	X			X	X	X			X

a Sources: Adapted from Gorth and Perkins (1979b, p. 3) and Pipho (1983) with permission of the authors.

b The specifications on the subject area indicate characteristic of program (X) or optional/conditional (O).

ents. Even if the academic and/or life skills are selected and reviewed by these groups, the ambiguity still remains, particularly when one attempts to identify those skills needed by adults to "function" or to "survive" in society. Haney and Madaus (1978) noted:

> People function differently in society, and some do it in ways offensive to others. Are we interested in the "essential skills" of the librarian or the lawyer, the bureaucrat or the baker, the con artist or the congressman? Would prisoners be considered as "functioning" in society? Or pardoned politicians? Even if we could reach agreement on what constitutes success (for example, functioning at a high level of competency) and what constitutes minimum functioning in society, their determinants are simply not very well understood. We do know, for example, that success in school seems not to be a very good predictor of success in later life—at least as measured by social scientists. (p. 465)

For the present, it appears that consensus on what constitutes the domain may be attained by some group which is representative of the lay public and professional educators in a given state, but the inherent ambiguity in an objectives-based definition can not be removed.

Coupled with this criticism is the charge that traditional item construction procedures are also ambiguous inasmuch as they can result in a set of items that manifest the biases and idiosyncrasies of each test maker. Consequently, different test makers would probably develop different items from the same specifications.

Domain Specification Strategies

Although the traditional approach is predominant in minimum competency testing programs, perhaps for practical reasons, several new strategies have been devised in order to overcome the aforementioned deficiencies: (1) amplified objectives, (2) IOX test specifications, (3) item transformations, (4) item forms, (5) algorithms, and (6) mapping sentences (for details, see Millman, 1980; Popham, 1981b, 1984; Roid, 1984; Roid & Haladyna, 1982). The characteristics of these strategies are outlined in Table 5.5. Clearly, the applications have been restricted to reading or mathematics, or both. The first two objectives-based strategies are more adaptable than item transformations, item forms, and algorithms. Although mapping sentences can be applied to most any domain, there are few examples of its utility. As the project applications suggest, life skills is a relatively unexplored domain. Recent developments on other approaches such as Tiemann and Markle's (1983) system derived from the research on concept and rule learning are also rather limited.

One underlying purpose of these strategies is to provide an *unambiguous* definition of a domain by implicitly or explicitly delineating sets of rules for generating test items, such that any two test makers would construct identical items from the same specifications. However, the extent to which the strategies can actually supply an unambiguous link between a domain of competencies and the corresponding test items varies markedly from one strategy to the other (Berk, 1980a). For building minimum competency tests, the ambiguity in defining what are "basic," "essential," "functional," or "survival" skills is still problematic. Regardless of how objective or mechanical the strategies in Table 5.5 operate in producing the items, the subjectivity used to arrive at the competencies remains.

Setting the Performance Standard

Setting the standard for minimum competence is the most important and the stickiest technical topic in minimum competency testing. Since the standard is the point of decision making and the basis for inferences about individual competency, the state department of education must be concerned about whether it provides a foundation for accurate, fair, and equitable decisions. If that founda-

TABLE 5.5
General and Technical Characteristics of Six Strategies for Defining the Domain of Competencies[a]

Strategy	Developer(s)	Project Application/ Domain(s)	Rule Structure	Major Components	Item Domain[c]
Amplified Objectives	Baker (1974) Popham (1974)	IOX Test Development/ Reading, Language, Mathematics, Social Studies	Traditional Item Construction Rules	Objective Sample Item Testing Situation Response Alternatives Criterion for Correctness	Infinite
IOX Test Specifications	Popham (1978a, 1984)	IOX Test Development/ Reading, Language, Mathematics, Social Studies, Affective, Psychomotor	Traditional Item Construction Rules	General Description Sample Item Stimulus Attributes Response Attributes Specification Supplement	Infinite
Item Transformations	Anderson (1972) Bormuth (1970) Conoley & O'Neil (1979) Finn (1975)	UCLA Readability Project/ Reading	Transformational Rules	Base and Derived Sentence	Finite
Item Forms	Hively et al. (1973) Osburn (1968)	MINNEMAST Project/ Mathematics	Generation Rules	Shell Replacement Sets	Finite
Algorithms	Scandura (1973, 1977)	MERG Projects[b]/Mathematics, Reading	Rules of Competence (Higher and lower order)	Equivalence Classes	Finite
Mapping Sentences	Berk (1978) Castro & Jordan (1977) Guttman (1970) Jordan (1978) Schlesinger (1978)	Cross Cultural Research/ Attitudes	Facet Design and Item Construction Rules	Facets Facet Elements Semantic Profiles	Infinite

[a]Source: Adapted from Berk (1980a, p. 51) by permission of Educational Technology Publications, Englewood Cliffs, NJ.
[b]The Mathematics Education Research Group (MERG) Projects provided the bases for most of the empirical research on algorithms.
[c]While an item domain may be viewed theoretically as either infinite or finite regardless of the particular strategy, the distinction between the two types of domain is intended to draw attention to the relative precision of the strategies and the need to consider that characteristic in their application.

tion is shaky, one will inevitably confront the consequences of inaccurate, unfair, or inequitable decisions in the school board room or courtroom.

The performance standard can be expressed as a number (24 out of 30 items), as a percentage (80%), or as a proportion (.80) of the items an individual must answer correctly.[2] The number which is based on the specific item sample measuring a single objective or a cluster of objectives (e.g., total test) is commonly referred to as the cutoff or passing score. It is the score that cuts the score distribution in two mutually exclusive categories: one category containing the scores from which "competency" is inferred and a second category containing the scores from which "incompetency" is inferred. Individuals who are labeled competent must score at or above the cutoff score; those who are labeled incompetent score below the cutoff score.

Although the percentage and proportion correct have been used interchangeably with the term cutoff score, they should be reserved more appropriately for the standard of performance in the item domain.[3] That is, if an individual can answer correctly 24 items in the 30-item sample, it is expected that he or she should be able to answer correctly about 80% of the items in the domain. If the domain happens to consist of 150 items, than 120 items or more should be answered correctly.

The responsibility for setting the standard on a minimum competency test resides with the state in about 80% of the cases (Gorth & Perkins, 1979b). In most other cases, the local districts set the standard. Very often the standard is specified for each subject area measured on the test and for each subset of items comprising a subject area. Only Connecticut and Tennessee are required to set a passing score for the total test.

While the polemics over certain issues in standard setting are far from over, at present there seems to be consensus among the experts on standard setting on at least one issue—*all of the methods involve some form of human judgment.* A completely objective, scientifically precise method does not exist (see Rowley, 1982). Regardless of how complex and statistically sophisticated a method might be, judgment plays a role in the determination of the cutoff score and/or in setting acceptable classification error rates. However, when a legislature sets a standard such as 80% without any foundation or reason, the judgment is capricious. This is the weakest and least defensible approach to standard setting. For its lack of any logical, experiential, or empirical justification, it has been characterized as the "cardiac approach" (Berk, 1979b, 1983), i.e., I know in my heart that she is competent and he is incompetent.

[2]Alternatively, a standard may be expressed as the number or percentage of competencies mastered. Multiple standards or cutoffs may also be used. However, these interpretations are less frequent in minimum competency testing programs than the number of items a student must answer correctly.

[3]The observed percentage correct is not necessarily the best estimate of the domain percentage correct.

Deficiencies of the "Cardiac Approach"

The deficiencies or problems associated with this approach are numerous. In order to appreciate the serious ramifications of decisions made on the basis of that type of standard, a few of the problems are specified below in terms of competency testing practices:

1. *An individual's pass-fail performance on the test has no meaning.* If an individual passes the test, there is no way of knowing whether he or she truly possesses the necessary skills (e.g., academic, life, survival, job-specific). The relationship between the performance standard and competence-incompetence on the actual skills is indeterminable. If an individual does well or poorly, there is no way to explain why.
2. *The percentage of individuals passing the test has no meaning.* This information which is simply an aggregate of individual performance data is supposed to indicate the overall competency of the group (e.g., the percentage that can be certified) and often the effectiveness of the instructional program as well. For example, if 70% of the 11th graders passed a minimum competency test as a requirement for graduation, no explanation of this percentage in terms of competence is possible. Certainly anyone can attach any meanings that they wish; negative inferences would be as unjustified and unfounded as positive ones.
3. *The standard does not reflect the difficulty or complexity of the items* measuring a single objective, a collection of different objectives or the total construct. Given the probable variability in item difficulty levels, an 80% standard may be easily attainable in some objectives or tests and highly unrealistic or unattainable in others.
4. Coupled with this insensitivity to difficulty is *the unavailability of any performance data on how individuals who are judged to be competent* (by their teacher or immediate supervisor) *in their position actually score on the test.* This information is essential to assess whether the standard is too high or too low. It would also provide a means of linking the standard to competency on particular skills.
5. Probably the most unfortunate consequences of using a completely arbitrary standard are the *incorrect, unfair, and inequitable decisions that could be made in individual promotions and graduation certification.* The cardiac approach precludes the estimation of decision accuracy, fairness, and equity. For example, the incorrect decision of denying a high school diploma to an individual who is truly competent (false incompetency error) suggests not only that the individual may be labeled as a failure, but also that the competency test failure may eliminate many potential opportunities and jobs for which that individual might otherwise be qualified. The seriousness of this problem becomes accentuated when one considers that the approach does not permit the decision maker even to estimate how many individuals have been mistakenly promoted or

certified or how many have unjustifiably been denied promotion or certification based on their competency test performance.

Clearly any standard setting method that is recommended as a substitute for the "cardiac approach" must address these problems. Specific criteria by which one can appraise the adequacy of a method will be delineated shortly.

On the spectrum of practicability, ranging from the simplest "cardiac approach" to the most complex Bayesian models, there are more than 30 different standard setting methods (Berk, 1985). Several extensive reviews of these methods have been conducted by Hambleton and his colleagues (Hambleton, 1980; Hambleton & Eignor, 1980; Hambleton & Powell, 1983), Meskauskas (1976), Shepard (1980a, 1980b, 1983, 1984), and Berk (1985). A few summaries, more limited in scope, have also been presented by Berk (1980d), Popham (1978b, 1981b, chap. 16), Livingston & Zieky (1982), and Jaeger (in press). The review of standard setting methods which follows will build on the structure, content, and insights proffered by these earlier works. In order to expedite a more perceptive selection of standard setting methods and to increase the use of the better methods by competency test makers, *criteria* for judging their quality and a *framework* for choosing the most appropriate method need to be developed. The next two sections are devoted to these considerations.

Criteria for a Defensible Standard Setting Method

In view of the aforestated deficiencies of the most popular standard setting method and the requirements of current competency testing programs, a defensible standard setting method should ultimately satisfy the following criteria:

1. Given the variation in the difficulty and complexity of the skills measured by competency tests, the method should be sensitive to the different difficulty levels;
2. Given the variation in the lengths of the tests and their component subtests, the method should be flexible for application to different test lengths:
3. Given the design and overall intent of competency tests, the method should be directly linkable to the performance of individuals who use the skills that are measured by the test in school or on the job;
4. Given the types of decisions for which the competency tests are used, the method should produce classifications of competence and incompetence for the different score continua;
5. Given the need for evidence to defend the accuracy of the decisions based on the standard, the method should provide estimates of probabilities of correct classification decisions and decision errors for any score point in the different score continua;
6. Given the various professional educators and lay people who will need to defend the method and to interpret the results on individuals and programs, the

method should be intuitively sound and conceptually simple, and the results should be easily interpretable;

7. Given the typical practical problems and constraints in educational settings, the method should be practicable in terms of execution and available resources and should be computationally simple.

Recent court decisions pertaining to the choice of a performance standard for a teacher certification test (National Teacher Examination) indicated that in order for the standard to be judged valid, it must be logical and be related to a specific level of job performance (see *Georgia Association of Educators* v. *Nix,* 1976; *United States* v. *State of North Carolina,* 1975, 1977; *United States* v. *State of South Carolina,* 1977). The implications of those decisions for setting minimum competency standards are expressed in criteria 3, 4, and 5. Criteria 1 and 2 focus on the sensitivity of a standard setting method to technical characteristics of competency tests (e.g., difficulty level, test length). The last two criteria stress the utility and practicability of a method. While it may be difficult for any single standard setting method to satisfy all of the criteria, certain criteria should be met so that the method might be defensible legally. Primary emphasis should be placed on criteria 3 and 4 (cf. Bernknopf, Curry, & Bashaw, 1979), and secondary weight should be assigned to criterion 5. The evidence gathered in support of decision accuracy, however, would be highly desirable, where possible.

A Framework for Standard Setting Methods

Numerous classification schemes have been devised to facilitate the study, interpretation, and use of cutoff score methods. From these schemes and the characteristics of the methods, Berk (1980d) derived a rather simple bilevel framework for classifying most available approaches. The first level, adopted from Meskauskas' (1976) review, partitions the methods into two major categories based on their assumptions about the acquisition of the underlying trait or ability: *state models* and *continuum models*. The second level, adopted in part from Hambleton's (1980) review, classifies the methods according to whether they are based purely on judgment or incorporate both judgmental and empirical information: *judgmental methods* and *judgmental-empirical methods/models* (see also Berk, 1985, for an extension of this classification). There are certainly other features that test makers need to consider, such as the definition of the internal or external criterion variable, the type of data, the distribution assumptions, and the specifications of a loss function (utility analysis). However, in the interest of parsimony, the bilevel framework should prove adequate for an analysis of the major methodological issues and to guide the selection of the type of method appropriate for decisions of grade level promotion and high school graduation certification.

The first step toward deducing which standard setting method is best for a particular competency test and decision application is to determine which general

standard setting category is most appropriate: state or continuum. The key factor in this determination is the assumption regarding the acquistion of the underlying ability.

State models assume that competence or true-score performance is an all-or-nothing state; the standard is set at 100%. Deviations from this true state are presumed attributable to "intrusion" (false competency) and/or "omission" (false incompetency) errors. After a consideration of these errors, the standard is adjusted to values less than 100%. Glass (1978c) referred to these models as "counting backwards from 100%" (p. 244). Unfortunately, this all-or-nothing assumption is implausible, unrealistic, or difficult to apply to the academic and life skill domains measured by minimum competency tests. Competence is usually conceptualized in "degrees" such that it could be defined at any number of points on a test score continuum.

Continuum models assume that competence is a continuously distributed ability that can be viewed as an interval on a continuum, i.e., an area at the upper end of the continuum circumscribes the boundaries for competence. This conceptualization appears to fit the design and intent of most competency tests.

Given this initial assessment of the two standard setting categories in terms of current practices in competency testing, only a brief description of some state models and a more extensive description of those continuum models with the greatest potential for addressing the standard setting problem will be presented in the succeeding sections.

State Models of Standard Setting

Although a considerable amount of research has accumulated on standard setting, state models have received relatively little attention. Macready and Dayton (1980) have provided the most comprehensive survey of state models to date. The sources for these models are listed in Table 5.6. Although they claim that the models are nonjudgmental in nature, those models possess many of the

TABLE 5.6
Primary Sources for Selected State and Continuum Models of Standard Setting
(Listed Alphabetically by Category)

STATE	CONTINUUM	
	Judgmental	Judgmental-Empirical
Judgmental-Empirical		
Bergan, Cancelli, and Luiten (1980)	Angoff (1971)	Berk (1976)
Emrick (1971)	Ebel (1979, chap. 17)	Block (1972)
Knapp (1977)	Jaeger (1978)	Huynh (1976b)
Macready and Dayton (1977, 1980)	Nedelsky (1954)	Kriewall (1972)
Roudabush (1974)		Livingston (1975)
Wilcox (1977a, 1977b)		Livingston (1980)
		Livingston and Zieky (1982)
		Novick and Lewis (1974)
		van der Linden and Mellenbergh (1977)
		Wilcox (1979a)

same judgmental and empirical characteristics of the decision-theoretic approaches for continuum competency models. A further discussion of this point follows.

The various models employ decision rules to identify the cutoff score that minimizes expected loss due to classification errors. Examples of these models include Emrick's (1971) mastery testing evaluation model, Roudabush's (1974) true score model, and Macready and Dayton's (1977, 1980) latent state models (see also Bergan, Cancelli, & Luiten, 1980). The decision rules require judgment in designating the loss ratio. The subjectivity involved in this process has been described at length by Shepard (1980a). Macready and Dayton 1980) indicate that all decision making must incorporate implicitly or explicitly a weighting of losses. Yet they also note that this judgmental component can be eliminated by setting the loss ratio equal to 1.0. In addition, they recommend a judgmental assessment of parameter estimates in conjunction with the absolute and relative statistical assessments of model fit. Clearly, judgment is an integral part of the decision-theoretic state models.

There are several specific limitations of the models that render them less compatible with competency testing programs than the continuum models. One limitation is that some of the models (e.g., Knapp, 1977; Roudabush, 1974; Wilcox, 1977a, 1977b) are based on mastery of only one or two items. Decisions at the item level would be appropriate, for example, in the context of algorithmic testing as in Scandura's (1977) structural learning theoretic approach. The use of a single item to measure attainment of an objective, however, is extremely restrictive in view of the structure and imprecision of most domain specifications. Coupled with this limitation is the requisite homogenity of the domain. Only discrete pieces of information (facts, terminology, etc.) or skills where perfection is essential would produce an adequate model fit. This restriction constrains the application of the models to low-level cognitive skills and ultra-specific objectives. The third limitation pertains to the requisite homogeneity of the student population that is tested. The models assume that competents answer all items correctly and they have an equal chance of incurring an inappropriate response (omission error) to an item. The converse assumptions exist for incompetents. Intact classes, schools, and school districts are more heterogeneous than these assumptions would permit. Probably the composition of certain specially formed groups of students would provide the necessary homogeneity. Finally, many of the models are theoretically and statistically complex. This factor alone will limit their application and usefulness.

Continuum Models of Standard Setting

The bulk of the research on standard setting has concentrated on continuum models. In fact, the majority of the cutoff score methods developed within the past decade fall into this category, and consequently the reviews cited previously have focused primarily on these methods. Table 5.6 presents the sources for the methods according to the judgmental and judgmental-empirical classifications.

The *judgmental methods* are based on judgments of the probability that minimally competent persons would select particular distractors in a multiple-choice item (Nedelsky, 1954) or the probability that they would answer the item correctly (Angoff, 1971; Ebel, 1979, chap. 17; Jaeger, 1978). The subjectivity of these item content decisions used to arrive at an overall cutoff score was expressed succinctly by Shepard (1980a): judges have the sense that they are "pulling the probabilities from thin air" (p. 453). This problem is reflected in the variability among judgments within a single method and also across methods (see Berk, 1985; Jaeger, in press). Recent empirical comparisons of the Angoff, Ebel, and Nedelsky methods have found that they produce different cutoff scores and the Nedelsky method yields consistently lower cutoffs than the others (Andrew & Hecht, 1976; Behuniak, Archambault, & Gable, 1982; Brennan & Lockwood, 1980; Colton & Hecht, 1981; Halpin, Sigmon, & Halpin, 1983; Kleinke, 1980; Koffler, 1980; Poggio, Glasnapp, & Eros, 1981; Saunders, Ryan, & Huynh, 1981; Skakun & Kling, 1980). Van der Linden (1982) even identified three possible sources of arbitrariness in the Angoff and Nedelsky techniques: (1) different conceptions of mastery underlying the technique, (2) different interpretations of learning objectives, and (3) intrajudge inconsistency.

This imprecision and the methods' strong dependence on judgments that are relatively unsystematic and arbitrary render these approaches less desirable than the judgmental-empirical methods for use with minimum competency tests. The Angoff method, in fact, does appear to satisfy six of the seven criteria for a defensible standard setting method specified previously; criterion five requires empirical information.

All the remaining standard setting methods not mentioned in the preceding sections can be lumped into the *judgmental-empirical category*. These methods are based on some type of judgment and actual or simulated data, judgmental data, and/or distribution assumptions. To clarify this point and to justify this classification, the specific judgmental and empirical components in 10 continuum methods that have been given wide visibility in the research literature are defined in Table 5.7. They appear to be the primary candidates for resolving the standard setting problem in many competency testing programs. Just how many nominations a method receives will depend largely on how well it meets the seven criteria.

As one examines these methods, the role of judgment should not be underestimated. While the majority of the judgmental-empirical methods are statistically sophisticated, that does not necessarily imply that they are scientifically precise. The judgmental component of each method furnishes the foundation for much of the statistical estimation of probabilities of correct classification decisions and false competency/false incompetency decision errors.

The judgmental-empirical methods differ according to other characteristics as well: (a) overall purpose, (b) type of empirical information, (c) definition of internal or external criterion variable, (d) distribution assumptions, (e) consideration of utilities, (f) statistical sophistication, and (g) practicability. Perhaps the

TABLE 5.7
Judgmental and Empirical Components of Continuum Methods for Setting Cutoff Scores and/or Estimating Error Rates[a]
(Listed in Order of Increasing Overall Complexity)

Method	Source	Judgmental Component	Empirical Component		
			Actual Data	Judgmental Data	Distribution Assumptions
Educational consequences	Block (1972)	Selection of criterion variable	X		
Criterion groups	Berk (1976)	Selection of intact criterion groups	X		
Contrasting groups/Borderline groups	Livingston and Zieky (1982)	Selection of individuals to compose comparison groups	X		
Binomial model	Kriewall (1972)	Setting boundaries for mastery and nonmastery ranges			X
Utility based	Livingston (1975)	Selection of criterion variable; assignment of benefits/costs	X	X	
Linear loss function	van der Linden and Mellenbergh (1977)	Selection of cutoff for latent variable; assignment of losses	X	X	X
Stochastic approximation	Livingston (1980)	Selection of performance criterion	X	X	
Control comparison	Wilcox (1979a)	Selection of control by panel of judges	X		X
Beta-binomial model (Empirical Bayesian)	Huynh (1976b), Huynh and Saunders (1979), Wilcox (1979b)	Selection of referral task	X		X
Bayesian decision model	Novick and Lewis (1974), Schoon, Gullion, and Ferrara (1979), Swaminathan, Hambleton, and Algina (1975)	Setting prior probabilities and loss ratio	X	X	X

[a] Source: Reprinted by permission from Berk (1980d, Table 1, p. 568), *Applied Psychological Measurement*, 4(4), Fall 1980, edited by David J. Weiss, Copyright 1980, West Publishing Company. All rights reserved.

most important and basic distinction between these methods, however, pertains to their purposes. Only the Berk (1976) and Livingston and Zieky (1982) approaches are intended to *select a cutoff scores*; all of the remaining methods *presume a standard already exists* on a criterion or latent variable. This standard is then translated into a cutoff score for the test, and decision error rates based on various assumptions are estimated. In some cases those rates can be used to adjust the cutoff. In fact, van der Linden (1980, p. 470) emphasized that even the most complex decision-theoretic models are not techniques for setting standards or optimizing competency decisions; they *are* techniques for minimizing the consequences of measurement and sampling errors once the true cutoff has already been chosen.

Inter alia, the general *unavailability of an acceptable criterion measure* of present or future individual competency makes it extremely difficult to apply the majority of the methods in Table 5.7 to minimum competency tests. Their other deficiencies have been mentioned elsewhere (Glass, 1978c; Hambleton & Eignor, 1980; Shepard, 1984).

Among the remaining methods, Kriewall's (1972) binomial model utilizes an indifference zone instead of a true cutoff score to differentiate between competents and incompetents and has a restricting distribution assumption. While an indifference zone or region of no-decision may be meaningful in sequential mastery testing at the classroom level, an exact point for the dichotomous classification of all individuals is required for most competency test decisions.

The Bayesian decision models permit the incorporation of a loss ratio, prior information on the distribution of domain scores, current information on the person's domain score, and the degree of certainty that a person's domain score exceeds the cutoff score (Schoon, Gullion, & Ferrara, 1979). Unfortunately, those models possess at least three disadvantages pertinent to the seven criteria: (1) they constitute a rather circuitous solution by augmenting as opposed to actually determining a cutoff score; (2) they are theoretically and statistically complex, and (3) their execution would be unwieldy and the results would be difficult to explain given the dimensions and constraints associated with competency test development by school districts and state departments of education.

Recommendations

It would appear as though the original list of potential methods has now been reduced to include only the criterion- and contrasting-groups methods. Despite the fact that no other alternatives are apparent at this time and these two methods are far from perfect (see Berk, 1984e, chap. 6), they do provide a best fit to the criteria for a defensible method. Probably an amalgam of both methods plus some extensions are necessary to address all aspects of the standard setting problem in minimum competency testing.

The method that seems to hold the most promise for competency tests in education can be derived from the construct validation models proposed by Berk

(1976) and Livingston and Zieky (1982) and the variety of statistical techniques suggested by Berk (1976) and Koffler (1980) that can be used in conjunction with those models. The statistical techniques are especially valuable for selecting the optimal cutoff score based upon estimates of correct and incorrect classification probabilities and the weighted cutoff score based upon probabilities that have been adjusted after a cost-benefit utility analysis.

The judgmental component of this approach consists of operationally defining competence in terms of the actual test performance of individuals who have been judged by their teachers, immediate supervisors, or similar persons as competent on an appropriate collection of skills (e.g., Christie & Casey, 1983). Teacher nominations of masters and nonmasters of academic skill objectives in reading, mathematics, and writing could be used effectively. For survival level skills, occupational groups of unskilled and service workers could be compared with unemployed adults or junior high school students. The competency groups are frequently accessible through the coordinators of work-study programs in local districts.

The process of identifying "competent" or "minimally competent" individuals for inclusion in one of the criterion groups represents the Achilles heel of the approach. Regardless of the rigor imposed on the specification of selection criteria and the systematic and standardized procedures used with each teacher or supervisor, there is no known strategy for objectifying the judgments. Interpretations of what is "competent" in relation to a well-defined list of skills may be diverse or comparatively narrow. There is no way to verify either. One must accept this scientific imprecision in the context of the state of the art and proceed to the next steps. If this judgmental process is not credible or intuitively convincing to the decision makers, the empirical component that follows from that premise will be meaningless. The explicit steps for setting the cutoff score have been outlined in the references cited previously (see also Berk, 1984e, chap. 6).

Unless a deliberate and conscientious attempt is made to obtain estimates of how "survivors" in different occupational categories perform on a minimum competency test, decision makers will be hard-pressed to assign meaning to the passing score and to the diploma (Berk, 1983). Only by testing individuals who have been judged competent can one ascertain the validity of the standard and of the decisions based on that standard.

Gathering Appropriate Validity Evidence

Validity is the degree to which a test achieves the purposes for which it was designed. That is, it relates to the intent or purposes of the test. For if a test does not perform its intended functions satisfactorily, why use it? This definition suggests that validity is

1. inferred from the way in which the test scores are used and interpreted;
2. specific to a particular score use;
3. determined ultimately by judgment;
4. expressed by degree.

The three traditional components of validity—content, criterion-related, and construct—are applicable to minimum competency tests. Only the emphases are different from those of norm-referenced tests due to the first three considerations listed above. In fact, the emphases have given rise to some new types of validity which are peculiar to competency testing. There are a few relatively recent discussions of validity for criterion-referenced and minimum competency tests by Hambleton (1980, 1984; Hambleton & Eignor, 1980), Jaeger (in press), Linn (1979b, 1980), Madaus (1983), Millman (1979), and Shepard (1980b). Some of the key issues related to content, curricular, and instructional validity, sex, racial, and ethnic bias, and criterion-related validity are examined here.

Content Validity

Content validity refers to the extent to which the items on a test constitute a representative sample of the domain of items the test is intended to measure. The adequate sampling of the domain of competencies via explicit content specifications is necessary to assure clarity and meaning in test score interpretation. Several procedures for assessing the match between the items and the objectives and the representativeness of the item sample have been suggested by Berk (1984a) and Hambleton (1984).

Unfortunately, the validity evidence gathered by such procedures is not sufficient for minimum competency tests, according to the ruling of the Fifth Circuit Court of Appeals in the trial of *Debra P.* v. *Turlington* (1981). In *Debra P.*, student plaintiffs challenged Florida's functional literacy test as the requirement to receive a standard high school diploma. Functional literacy was defined as "the satisfactory application of basic skills in reading, writing, and arithmetic, to problems and tasks of a practical nature as encountered in everyday life" (p. 259). Experts for the plaintiffs argued that the students should have received instruction on the domain tested if the certification test was to be valid. The Fifth Circuit Court ruled that "the state must demonstrate that the material on the test was actually taught in the state's classrooms in order to establish the requisite 'content validity'" (Citron, 1982, p. 11).

Much of the testimony concentrated on *curricular validity* and *instructional validity*, and the court failed to distinguish between those types of validity and content validity. The confusion in defining these terms is expressed by Madaus (1983): "The court's description of *content* validity—including as it does a reference to *curricular* validity—in fact implicitly incorporates McClung's (1978, 1979) earlier descriptions of *instructional* validity" (p. 25).

Curricular validity refers to the extent to which the items on the minimum competency test measure the content of a local curriculum (cf. McClung, 1979, p. 682). While conceptually similar to content validity (Madaus, 1983; Schmidt, Porter, Schwille, Floden, & Freeman, 1983) and even viewed as synonymous with content validity (Cureton, 1951; Hopkins & Stanley, 1981, chap. 4; Madaus, Airasian, Hambleton, Consalvo, & Orlandi, 1982), curricular validity is operationally very different. In the case of minimum competency tests, it does not always focus on the domain of academic and/or life skills the test was designed to measure; it deals with a specific domain to which the test is applied. The relevance of the test in a specific application is being evaluated. For basic skills, which are typically included in all curricula, this issue of relevance is not a problem. It is the domain of life or survival skills which is not usually part of the curriculum that is troublesome.

Evidence of curricular validity is obtained by determining the degree of incongruence or mismatch. This is based on a systematic, judgmental review of the test against the curricular objectives or materials by content experts. These experts may be classroom teachers or curriculum specialists; they are the only professionals in a position to judge curricular validity. The review can vary as a function of the following: (a) single grade versus cumulative grade content, (b) specificity of objectives or content/process matrix, (c) internal versus external determination, and (d) curricular materials versus actual classroom activities (for details, see Schmidt, 1983a, 1983b; Schmidt et al., 1983). What emerges from this process are several estimates of content overlap, including the amount of content in common, the percentage of the local curriculum measured by the test, and the percentage of items on the test not covered by the curriculum. The second estimate in particular can furnish evidence of the curricular validity of the test.

While curricular validity is an important characteristic, the most crucial legal question deals with whether minimum competency tests measure *what is actually taught* in the schools. Very often it is simply assumed or implied that evidence of curricular validity means that the objectives guided the instruction and the curricular materials were used in the classroom. This does not necessarily follow, as several studies have demonstrated (Hardy, 1983; Leinhardt & Seewald, 1981; Leinhardt, Zigmond, & Cooley, 1981; Poynor, 1978; Schmidt et al., 1983). What is measured by the test is not always the same as what is taught, especially with regard to life or survival skills on minimum competency tests. Hence, a distinction has been made between these different domains to which the test items can be referenced (Schmidt et al., 1983). When the domain is the instruction actually delivered, a "measure of whether schools are providing students with instruction in the knowledge and skills measured by the test" (McClung, 1979, p. 683) is called instructional validity.

Instructional validity refers to the extent to which the items on the test measure the content actually taught to the students. The requirement that minimum competency tests must be instructionally valid strongly suggests that either life

skills be taught in the schools as a standard component of the curriculum or those skills should not be tested. If state departments of education tend to choose the latter, in time, the testing programs will probably drift back to the basics and only academic skills may be measured.

Several techniques have been proposed for assessing the overlap between the test and the instruction. Popham (1983) has identified four data-sources for describing whether students have received instruction that would enable them to perform satisfactorily on a test: (1) observation of classroom transactions, (2) analyses of instructional materials, (3) instructor self-reports, and (4) student self-reports. Although he views these sources as methods for determining the adequacy of test preparation (Yalow & Popham, 1983), they can be considered as techniques for gathering evidence of instructional validity. Unfortunately, Popham's (1983) evaluation of those techniques indicates that the process of estimating the percentage of a minimum competency test that has been covered by teaching is fraught with difficulties. Most of these are methodological problems in executing the data-gathering procedures, so as to provide *adequate* evidence (see Leinhardt, 1983; Schmidt et al., 1983). They stem, in large part, from the variability of instructional content, not only among different classes, but within a single classroom.

Despite the conclusion about how instructional validity evidence should be obtained, two recent court rulings revealed that sufficient evidence could be expressed in very different forms. In *Anderson* v. *Banks* (1982) the trial court accepted a Georgia school district's proof of instructional validity based on expert testimony that tested material was covered in their schools' curriculum, and on teacher testimony that that curriculum was actually taught. At the other extreme, in the latest phase of *Debra P.* (1983), Florida conducted an extensive study of instructional validity to amass voluminous evidence that the material covered on the test was indeed taught in the state's classrooms. The study consisted of six components (Fisher, 1983): (1) principals' disemination of the State Student Assessment Test, Part II (SSAT-II) skills, (2) a student remediation study to determine the status of students who failed the test on their first try, (3) a district-by-district analysis of content in the curriculum of the 67 school districts based on self-report, (4) a survey of approximately 65,000 teachers in the state to ascertain whether they taught the SSAT-II skills sufficiently to enable students to master the skills if they applied reasonable effort, (5) on-site visitations of a sample of schools in every district to verify the accuracy of the self-report and to determine if there was evidence of instruction on the SSAT-II skills, and (6) a survey of about 5,000 students asking them whether they had been taught the test material (see also Citron, 1983a). The court concluded that "although the instruction offered in all the classrooms of all the districts might not be ideal, students are nevertheless afforded an adequate opportunity to learn the skills tested on the SSAT-II before it is used as a diploma sanction" (*Debra P.* v. *Turlington,* 1983, p. 186).

Sex, Racial, and Ethnic Bias

Another aspect of validity that must be addressed in the context of minimum competency testing is sex, racial, and ethnic bias. The research and discourse on bias are organized in terms of validity issues and, in fact, reflect the traditional trinary scheme mentioned previously: content, criterion-related, and construct. Bias in the content of the test has been investigated judgmentally and statistically. A *judgmental review* or logical analysis (Shepard, 1982) is intended to detect stereotypic, culture-specific, and offensive language and to assure fair representation in the work roles and life styles of sex, racial, and ethnic groups (Tittle, 1982). The *statistical analysis* based on an appropriate experimental design (Schmeiser, 1982) is conducted to detect discrepancies in item performance between specific groups (e.g., males and females, blacks and whites, Hispanics and whites). When such discrepancies are found, an *a posteriori (judgmental) analysis* is employed to discern whether true item bias is present and, if it is, to deduce explanations for why it occurred and consider procedures for eliminating it (Scheuneman, 1982).

An item is biased if individuals with the same ability have an unequal probability of answering the item correctly as a function of their group membership. This definition is similar to those proposed by Pine (1977) and Scheuneman (1979). Operationally, bias is inferred from differences in performance between groups. The differences are computed using one or more statistical methods (see review by Angoff, 1982, and Ironson, 1982); these methods have been examined in several studies (Burrill, 1982).

Interestingly, item bias has been the predominant form of bias investigation undertaken by publishers of ability and achievement tests, but item bias has not received attention in minimum competency testing until recently (e.g., Christie & Casey, 1983). Initially, the content or behaviors that a test measures is an integral part of all score inferences, and since the item is the most fundamental level of content analysis and the foundation for these inferences, *item bias studies are necessary for all tests.* However, they are not sufficient for all test score inferences and uses. For example, additional studies would be required if the scores are used to make predictions about future performance, which is implied in the construct of life skills. Second, charges of bias from numerous sources frequently include a citation of specific items that are claimed to be biased against a minority population. These sources can be public or professional organizations such as Parents in Action on Special Education (PASE), the National Education Association, and the Association of Black Psychologists (Jackson, 1975; Williams, 1970, 1971), or individual citizens and organizations who take legal action on specific claims of bias (e.g., *Armstead et al.* v. *Starkville, Mississippi Municipal Separate School District,* 1972; *Larry P. et al.* v. *Wilson Riles et al.,* 1979, 1984; *PASE et al.* v. *Joseph P. Hannon et al.,* 1980). Third, the results of bias studies at the subtest and total test levels do not preclude the

presence of bias at the item level. For example, a predictive bias study that finds no sex bias does not rule out the possibility that specific items on the test may be biased against females. Fourth, item bias studies can be incorporated into the early stages of test construction and item analysis to minimize the chances of bias accusations arising later. Finally, the elimination of item bias may decrease the likelihood of test bias, although research evidence is needed to verify this relationship.

The *test bias* literature has focused almost exclusively on intelligence and aptitude tests (Jensen, 1980). The studies have dealt with predictive and construct validity issues. *Predictive bias* may be defined as follows:

> Bias exists in regard to predictive validity when there is systematic error in the prediction of the criterion score as a function of group membership.

This definition is a less technical version of the definitions proffered by Cleary (1968), Cleary, Humphreys, Kendrick, and Wesman (1975), and Reynolds (1982a). A slight restatement of Reynolds' (1982b, p. 194). definition of *construct bias* is presented below:

> Bias exists in regard to construct validity when a test measures different psychological constructs as a function of group membership or measures the same construct but with differing degrees of accuracy.

The statistical methods used to detect these two types of bias are no less numerous and varied than those employed in item bias studies (see review by Reynolds, 1982a). The indices which result are intended to signal possible bias and indicate, for example, whether a test predicts the criterion with greater accuracy for whites than for blacks or whether the constructs measured by the test are different for these groups.

Where bias is inferred, the minimum competency test scores for the group in question should be reported by the state. The nature of the bias should be fully explained. Indeed, all pertinent research evidence should accompany any presentation of scores partitioned by sex, racial, or ethnic subpopulations. Test scores may not be validly used without taking account of group differences. In view of the political and social implications of these distinctions, the decision maker should be very cautious in interpreting differential validity evidence.

While the bias literature has concentrated very heavily in the areas of item bias, predictive bias, and construct bias, many other types of bias have been described in relation to minority group populations (Baca & Chinn, 1982; Gonzales, 1982; Oakland, 1980; Oakland & Matuszek, 1977; Reschly, 1979; Samuda, 1975; Sattler, 1982, chap. 19; Ysseldyke, 1979). Examples are atmosphere bias, linguistic bias, examiner bias, and decision-making bias. The descriptions of these various sources of invalidity are usually couched in the con-

text of the litigation involving charges of racial or ethnic bias (see reviews by Bersoff, 1979, 1982a, 1982b; Jensen, 1980, chap. 2; Oakland & Laosa, 1977; Reschly, 1979) or the Public Law 94–142 (1975) mandate for *nondiscriminatory evaluation.*

Criterion-Related Validity

Criterion-related validity refers to the extent to which test performance is related to some criterion measure of performance. For minimum competency tests measuring academic skills, the mastery criterion must be defined operationally in terms of master and nonmaster students. These students can be selected using the criterion- or contrasting-groups procedures described previously. A *concurrent validity* study could then be conducted by correlating competency test performance and the criterion master-nonmaster classification. Alternatively, the test can be correlated with other achievement tests assessing the same content areas (e.g., Christie & Casey, 1983). A *predictive validity* study is appropriate to predict future performance related to life or survival skills. Since it is often impractical to wait several years to obtain criterion performance data on a current group of test takers, one can instead administer the competency test to adults in the community who by their occupation and/or supervisor's evaluation may be judged at a minimum level of survival or higher. Occupational groups of professional, managerial, sales, skilled, and clerical workers can be employed to establish a hierarchy of competency performance. Unskilled and service workers (e.g., cooks, custodians, truck drivers) can comprise a minimum competency (survival) group. Unemployed adults who are actively seeking employment can serve as an incompetent (nonsurvival) group. Correlations between the minimum competency test scores of these adults and their criterion occupational classification can furnish evidence of predictive validity.

One type of criterion-related validity especially important for minimum competency tests is decision validity. *Decision validity* refers to the extent to which a test can yield accurate decisions according to a criterion classification (Hambleton, 1980, 1984). This may be perceived as analogous to concurrent validity. The principal difference lies in what is being studied: the *decisions* reached on the basis of test scores or just the *test scores.* An investigation of decision validity examines the relationship between the decisions made using a specific test and the decisions made using a criterion procedure. In other words, two dichotomous variables are being compared: the pass-fail status on the minimum competency test and the competent-incompetent classification of the persons tested.

The effectiveness of a minimum competency test resides ultimately in the degree to which it can distinguish competent from incompetent students, that is, the accuracy of competent-incompetent classification decisions. Decision validity evidence is usually expressed as probabilities of correct and incorrect classifications, sensitivity and specificity indices, and validity coefficients (for details,

see Berk, 1976, 1984e, chap. 6). Essentially, the value or usefulness of a minimum competency test is contingent on the nature of this evidence. For example, if 93% of the students are correctly labeled competent and incompetent on the mathematics subtest, then that subtest may be judged effective in accomplishing what it was designed to do. However, if only 74% of the students are correctly classified with 18% false incompetency and 8% false competency errors on the writing subtest, it is less effective and, depending on the loss function adopted, the cutoff score may be lowered to reduce the 18% error rate.

Such evidence is also crucial in attempting to justify the selection of the performance standard using the criterion-groups and contrasting-groups approaches. Furthermore, without decision validity evidence related to the cutoff score, it seems pointless even to compute an index of decision consistency (see next section on "Reliability"). Certainly one can compute an index based on any performance standard. However, if it is not known whether the decisions based on the cutoff score will be accurate, then one possible interpretation of a high index of decision consistency might be that the *test can consistently classify students into the wrong groups*. Consistent decision making without accurate decision making has questionable value.

The groups of mastery and nonmastery students described in the preceding section on concurrent validity and in conjunction with the recommended standard setting procedure can be used in a decision validity study of the academic skills areas. Also, different occupational groups of competent and incompetent adults can supply the data for the life skills subtest. It is possible, in fact, to employ the same criterion groups for both the standard setting and criterion-related validity (concurrent, predictive, decision) analyses.

Estimating the Reliability of the Scores and Decisions

Reliability refers to the degree of consistency between two or more measurements of the same thing. It may be the individual scores or decisions based on those scores that are analyzed over repeated measurements using a single test or parallel test forms. This meaning of reliability should be viewed in the context of the following points. Reliability is

1. a necessary but not sufficient condition for validity;
2. inferred from the way in which the test scores are used and interpreted;
3. specific to a particular type of consistency;
4. determined ultimately by judgment;
5. expressed by degree.

There are numerous types of reliability that account for different sources of error in the test scores. Several summaries and critiques of reliability statistics recommended for criterion-referenced and minimum competency tests have been

conducted by Berk (1980b, 1984d), Hambleton et al. (1978), Linn (1979a), Millman (1979), Shepard (1980b), and Traub and Rowley (1980). In-depth presentations of two major categories of reliability have also been given by Subkoviak (1984) and Brennan (1984). This review concentrates on three components of reliability that are particularly important for minimum competency tests: parallel forms reliability, interscorer consistency, and decision consistency.

Parallel Forms Reliability

The development of parallel forms of a minimum competency test is essential for one or more of the following reasons. First, in this era of test disclosure (e.g., La Valle Act in New York), the public and the students may wish to scrutinize the test items and the answer key. Second, the ever present problem of test security can be reduced when several test forms are used. And third, students who are given multiple opportunities to pass a minimum competency test should not receive the same test each time.

These circumstances suggest that two or more test forms should be generated. The parallel forms reliability must then be estimated, and, finally, the scores on the different forms must be equated. The procedures for equating will be discussed in a subsequent section of the chapter.

Parallel forms reliability is estimated using two separate but equivalent, parallel, or alternate forms of a test. The forms are constructed systematically from the same competency specifications so that, at least from a judgmental perspective initially, they both appear to measure the same material. This can be accomplished by drawing two random samples of items from the domain of items developed from the specifications or by building the two forms item by item according to content and difficulty level. The former method results in *randomly parallel forms*; the latter produces *classically parallel forms*. The item sampling approach is often preferable because the reliability coefficient derived from the classical approach does not take into account item sampling error.

The test forms are then administered to the same group of students in close succession with no intervening time. Frequently the items from the two forms are included in one test, where Form A items may be even-numbered and Form B items odd-numbered. This procedure is intended to minimize the effects of certain factors that could lower the degree of equivalence. For example, fatigue at the end of the test should theoretically influence performance equally on items from both forms when the items alternate (A, B, A, B, etc.); if Form A items were administered first and Form B items administered second, only Form B would be affected.

The two administrations produce two sets of scores, one from each form. These scores can then be correlated to determine the degree to which the items on each form measure the same construct, an academic skill or life skills. A correla-

tion coefficient of .90 or above is required to adequately demonstrate equivalence.

In addition to the correlation coefficient, other statistics need to be computed in order to assess the equivalence of classically parallel forms. These are the mean, variance, and the item analysis results (i.e., difficulty, discrimination, and interitem correlation matrix) for each test form.

Interscorer Consistency

Most minimum competency tests currently in use typically employ an objective item format, such as multiple choice (Gorth & Perkins, 1979b). This characteristic facilitates either manual or computer scoring which cannot be influenced by individual judgment; that is, the scoring is totally objective, not subjective. In certain academic skills, for example, writing and speaking (e.g., Illinois, Massachusetts, Oregon), and in performance-based life skills, such as using a telephone in a simulated emergency situation, where the behaviors must be observed directly, objectivity is not easily achieved. The individuals who score an essay test or record specific behaviors may allow their own judgments, biases, and/or opinions to contaminate the results. This is possible whenever writing samples or essays are required or behavioral checklists or rating scales are used.

The problem is that if scores vary markedly from one scorer to another, how can one discern the true score. This fluctuation or inconsistency between scorers, judges, observers, or raters must be minimized in order to provide useful data. The most effective strategies for achieving interscorer consistency are to delineate very specific, operational criteria for scoring (or recording), and then to train the persons involved so that their tasks can be executed as objectively as possible.

One method to measure the degree of objectivity attained and, in essence, the effectiveness of those strategies is to estimate interscorer consistency. Over the past 30 years more than 20 different statistical indices have been recommended (see review by Berk, 1979a). Among the various indices, the correlation coefficient used to express the previous types of reliability can also be applied here. Two sets of scores/ratings by two independent scorers/observers are obtained on one group of students at the same point in time. The results are then correlated to estimate the scoring consistency. In this case, the index, referred to as an *interclass correlation,* assesses the amount of error in the scores due to the person(s) who did the scoring. No other source of error is considered.

The criterion for an adequate level of interscorer consistency may vary as a function of the skills or behaviors being measured, the particular scoring procedures followed, and the index used. Very often, as scorers/observers are being trained, several reliability checks are conducted, so that by the completion of training (and sometimes retraining), a near perfect level of consistency is attained. When coefficients are finally estimated, they usually fall in the .90s. For

minimum competency writing tests and other performance tests, interclass correlations in that vicinity are required to assure dependable individual decisions.

Decision Consistency

The type of reliability that reflects the purpose and the characteristics of a minimum competency test as well as the decisions for which the scores are used is decision consistency. It deals with the consistency of competency-incompetency classification decisions based on the performance standard.

There are two indices of decision consistency: p_o, the percentage of students consistently classified as competent and incompetent across repeated measures with one test or classically parallel test forms, and κ the percentage of students consistently classified beyond that expected by chance. They are derived from the threshold loss function that assumes (a) a dichotomous, qualitative classification of students as competent and incompetent based on a threshold or cutoff score and (b) the losses associated with all false competency and false incompetency classification errors are equally serious regardless of their size.

The selection of p_o or κ is a function of the method for setting the cutoff score (relative or absolute) and the conclusions reached from an analysis of the disadvantages of each index (see Berk, 1984d). The p_o index should be used where an absolute standard is chosen and for minimum competency tests that contain short subtests and/or yield low score variance. The κ index may be the preferred index of agreement where relative cutoff scores are set according to the consequences of passing or failing a particular proportion of the population, as in the case of some minimum competency tests where the cutoff score is adjusted according to the political, economic, social, and/or instructional consequences of not graduating or promoting a certain proportion of the students in the school district. The problems associated with κ, however, render it less useful then p_o.

In regard to estimating p_o or κ for minimum competency tests, the Hambleton and Novick (1973) and Swaminathan, Hambleton, and Algina (1974) two-administration procedures are recommended using classically parallel test forms. These procedures make it possible to measure both *stability and equivalence*. That is, $\hat{p}_o$ and $\hat{\kappa}$ will estimate the stability of the competency-incompetency decisions over time *and* the equivalence of the scores on the two item samples (test forms). Alternatively, when only one test form is available, Huynh's (1976a) single-administration approach or Peng and Subkoviak's (1980) approximation can be employed.

Equating the Scores on Different Test Forms

When parallel forms of a minimum competency test or two different levels of the test (e.g., 9th grade and 12th grade) are developed, score equating is necessary to assure fair and valid decisions based on the individual scores from those forms. A parallel forms reliability coefficient provides evidence only of the degree of

equivalence; even when this equivalence is perfect (1.0) and the forms are tau-equivalent, individual scores will differ on the two tests. For example, one form of a minimum competency test, Form B, may be easier than another form, Form A. If no adjustment in the scores were made to account for those differences in difficulty, a passing score, of say, 60, on each form would mean something different. It would be harder to attain that score on Form A. The student taking Form B would have an unfair advantage over the student who was administered Form A. For this student, the consequences of not equating the scores would be failing the test and not graduating. All scores must be equated across Forms A and B, especially the cutoff score and those scores close to the cutoff, in order to adjust for these differences and to establish their comparability (see, for example, Bernknopf, 1980).

Although the need for test score equating has existed for some time, the La Valle Act, effective January 1980, in New York, added a legal impetus. This law required test disclosure—providing students the opportunity to see the test questions used in obtaining their scores on admission tests. Once the questions were released, new test forms had to be generated. Equating the scores on these different forms became essential if the decisions about test takers were to be fair and valid (Berk, 1983).

Horizontal and Vertical Equating

There are two types of equating: horizontal and vertical. *Horizontal equating* involves equating test forms that are developed to measure the same content at the same level for the same population, as in the preceding example of parallel forms (A and B) of a minimum competency test. *Vertical equating* is the process of equating tests that differ in difficulty so that they are roughly "exchangeable," i.e., converting to a common scale the scores on forms of a test designed for populations at different grade levels (Slinde & Linn, 1977, p. 23). This equating is applicable to states where two or more levels of a minimum competency test are constructed. For example, a 9th grade preliminary (practice) or diagnostic version of the test may be administered prior to the 11th or 12th grade version used for graduation certification. (*Note*: This strategy is similar to the administration of the PSAT and SAT.) Equating scores at adjacent grade levels has been accomplished satisfactorily (see, for example, Slinde & Linn, 1979); equating tests that differ more drastically in difficulty, say two or three grade levels apart, is troublesome.

There are three major approaches frequently used to equate test scores: linear, equipercentile, and logistic or item response theory. The first two methods are traditional; they have been applied for more than three decades and are, by far, the most popular (Angoff, 1971; Flanagan, 1951). The logistic or latent trait models constitute a relatively recent innovation in the field (Holland & Rubin, 1982; Marco, 1981). One- , two- , and three-parameter models have been studied extensively, and variations of those models have also been examined (Phillips,

1983). The empirical research over the past 5 years that has compared the precision of these various models suggests, in general, that similar results are found across methods for tests of approximately equal difficulty (horizontal equating), but substantially different results occur for tests of unequal difficulty (vertical equating) (see Arter, 1982; Butera & Raffeld, 1979, Jaeger, 1981; Kolen, 1981; Kolen & Whitney, 1982; Linn, 1981).

The net effect of all of this research on test score equating is that it is now possible to translate the raw scores on parallel forms or different levels of a minimum competency test into one scale. The resulting scores are often called scaled scores, which are usually assumed to constitute an equal-interval scale. Although there are systematic equating errors associated with the scaled scores (Hoover, 1982), they are typically less serious than the unfair and invalid decisions that can result from not equating the scores on different forms of a minimum competency test.

CRUCIAL ISSUES IN MINIMUM COMPETENCY TESTING

Embedded throughout the preceding description of the technical specifications are the major issues confronting minimum competency test makers. Since most state departments of education have chosen to construct their own tests and the technical analyses are conducted using in-house expertise (the alternative is to contract the work to an external agency)[4] (Gorth & Perkins, 1979b), the settlement of some of the issues may be contingent more on the commitment of resources than on psychometric research. Practical constraints and available resources will probably dictate what can be done. Hopefully this will closely approximate what should be done.

According to the latest edition of the *Standards* (AERA/APA/NCME Joint Committee, in preparation) and the methodological recommendations given previously, minimum competency testing practices must meet certain "minimum" standards; that is, the tests should be psychometrically as well as legally defensible. The issues that appear to be most critical to the success of a minimum competency testing program along with suggestions for their settlement are listed below:

1. *Can the domain of minimum competencies be defined objectively*? The choice of what competencies should be tested involves the judgments of professional educators and the lay public. While basic academic skills in reading, mathematics, and writing have a concrete educational foundation in the school curricula, the selection of the most important skills for the purpose of testing in

[4]It is also possible to split the effort between in-house expertise and outside contractors.

high school is highly subjective. The definition of life or survival skills which lack such a foundation tends to be even more subjective. There is no objective method for defining the domain of competencies or any other domain. The choices at each step rest on value judgments. Acknowledging this subjectivity in the process means that the task is to obtain the consensus of all interested parties so that the definition is meaningful and credible. Imposing "objective" procedures on the process will not remove the subjectivity.

2. *Is there a "most effective" strategy for defining the domain*? For the specification of academic skills, the strategies listed in Table 5.5 represent tradeoffs between precision and practicability. Once an outline of the skills has been developed and reviewed, perhaps one of the objectives-based schemes such as amplified objectives, IOX test specifications, or mapping sentences (Berk, 1978) offers a reasonable compromise (Berk, 1980a). Since none of the strategies has been applied extensively to life skills and some of them have been tested only in reading or mathematics, the most adaptable objectives-based approaches again seem worthy of recommendation.

3. *Are standardized test administrations essential*? Standardized procedures for administering a minimum competency test must be documented in a test administration manual and then followed precisely by the person who administers the test. Strict adherence to administration instructions, time limits, test presentation, item response mode, and similar specifications is essential to ensure comparability of test scores and fairness for all students. In addition, certain efforts should be made to maintain test security and to eliminate opportunities for cheating. These efforts might include monitoring the testing process, simultaneous administration to all individuals taking the same test form, and requiring particular seating arrangements (e.g., with adequate space between seats). Irregularities in any of these administration procedures can render the test results invalid. The meaning of scaled scores on multiple test forms and the passing score on the test is contingent on the observation of standardized administration procedures. If some students are given more than the designated time to complete the test or there were "minor" variations in the test taking instructions, the interpretation of their scores must necessarily be different from the interpretation of all other scores. Their scores, in fact, should be judged invalid; those students experienced an unfair advantage over other students, and the scaled scores and the passing score can not be applied.

4. *Are performance tests necessary*? Paper-and-pencil multiple-choice tests have many advantages in the measurement of certain academic skills. However, they are inadequate tools to assess writing, listening and speaking, and several application level life or survival skills. Alternative item and test formats must be employed in order to measure those areas validly. State departments should consider essay formats (restricted and extended response), performance tests such as work samples, situational tests, in-baskets, and trainability tests (see Berk, in press), and behavioral checklists. Certainly, impracticability has been a

drawback of these techniques in large-scale assessments. Recently, however, their popularity has increased and some states have already incorporated performance-based methods in their minimum competency testing programs (e.g., Maryland, Nebraska, Nevada, South Carolina, Texas).

5. *Is their a defensible approach to setting a standard for minimum competence*? Given the judgmental limitations of all of the methods reviewed, there are three options: (1) use a judgmental method such as Angoff (1971), (2) use a judgmental-empirical method such as Livingston and Zieky's (1982) contrasting groups, or (3) use a combination of judgmental and judgmental-empirical methods. The combination approach which has been recommended by Hambleton (1980), Koffler (1980), Shepard (1984), and others has the advantage of capitalizing on the strengths of different methods and the disadvantage of reconciling conflicting results from those methods. A judgmental approach by itself, while politically appealing, is actually a systematic way to "objectify arbitrary input" on what the standard should be. In view of the state of the art, the most defensible course of action seems to be to use a data-based method. The contrasting groups approach has numerous advantages over the judgmental methods, plus it is relatively easy to implement. The primary difficulties with the approach relate to the selection of competent and incompetent persons. Such difficulties are not insurmountable. They are worth tackling, for it is the performance of those groups that gives meaning to the standard.

6. *Is instructional validity evidence necessary for a minimum competency test*? In the *Debra P.* case, the Fifth Circuit Court ruled that the state was required to demonstrate that the material on the test was *actually* taught in the classrooms. Although referred to as content validity in the decision, this evidence of instructional validity (McClung, 1979) must be obtained. The appellate decision offered no advice on how a state was to gather such proof. Popham (1983) has identified four data sources for measuring instructional validity. Unfortunately, at present there are major methodological problems in executing the data gathering procedures, although evidence can be obtained (see Fisher, 1983). If direct measurement is not possible, then the state has two options: (1) either incorporate the skills being tested into the curricular documents and instruction or (2) do not test those skills not being taught formally in the schools. In other words, life skills either should be taught or not tested. Testimony on the teaching of the academic skills should prove adequate (e.g., *Anderson* v. *Banks,* 1982).

7. *Can teaching the test improve instructional validity*? Teaching the specific items on the test or very similar items can destroy the value of the test as a representative sample from the domain of academic or life skills. Such a practice will also invalidate the test scores. The match between the test content and what is actually taught can be improved by teaching from the *objectives* that the test items measure. Teaching to the test or the test itself can only lead to invalidity.

8. *Can miminum competency tests be biased against females and minorities*? Any achievement test can be biased against a particular sex, racial, or ethnic

subpopulation of students as well as groups from different geographic regions within a state. Precautions should be taken during the construction of the test to eliminate stereotypic, culture-specific, region-specific, and offensive language and to assure fair representation in the work roles and life styles of all groups. Furthermore, statistical analyses of item and test bias (see Berk, 1982; Selkow, 1984) should be conducted to furnish evidence that the test scores can be used validly with different groups (Citron, 1983b).

9. *What types of validity evidence are most important for minimum competency tests*? Considering the traditional categories of validity evidence and issues 6 and 8, the most important type of evidence pertains to decision validity. It addresses directly the purpose of a minimum competency test and the use of the scores. Decision validity evidence indicates the degree to which a test can differentiate accurately between competent and incompetent students, and therefore, reveals whether the test is effective and useful. Such evidence can also be used to justify or defend the choice of the performance standard. Concurrent and predictive validity evidence should follow.

10. *What types of reliability evidence are most important for minimum competency tests*? Despite the continued reliance on Kuder-Richardson Formula 20 and alpha coefficients for minimum competency tests, a pool of reliability indices exists that relate to the specific design of the tests and the score uses. Perhaps most important is decision consistency evidence. Once an acceptable level of accuracy in competency-incompetency classification decisions has been attained (decision validity), the dependability of those decisions needs to be assessed. The recommended agreement indices (p_o or κ) provide evidence of the stability of the decisions and the equivalence of item samples based on classically parallel test forms. Single administration estimates are also available (Huynh, 1976a; Peng & Subkoviak, 1980). If parallel forms of the test are constructed or sampled, an equivalence coefficient should also be computed. Finally, if performance tests (or subtests) which require judgmental scoring or direct observation are used, estimates of interscorer reliability are essential.

11. *Do the scores on different forms of a minimum competency test have to be equated*? Score equating is necessary only when the different forms are used for the same decision. If parallel test forms are administered to different students the same year or in different years and passing either form is required to receive a high school diploma, then the scores must be equated onto a common scale so that adjustments in test difficulty can be made. The passing score and each score on the scale should have the same meaning regardless of which form is used. Equating is one method to assure fair and valid individual decisions irrespective of test form (assuming, of course, there are no other sources of unfairness or invalidity).

12. *Should handicapped students be required to pass a minimum competency test to receive a regular high school diploma*? According to a survey of state competency testing programs completed by the National Association of State

Directors of Special Education (1979), 19 states currently have some form of competency testing for the handicapped, 6 states require handicapped students to take the tests, and 7 states are either providing or are in the process of developing special testing procedures for the handicapped population (see also Wiederholt, Cronin, & Stubbs, 1980). Of special significance, however, is the fact that 31 states issue regular diplomas to handicapped students and 17 states leave that decision to the local school board's discretion. Few states issue special diplomas.

The relationship between minimum competency testing and the requirements of Public Law 94–142 (The Education for All Handicapped Children Act of 1975) suggests a set of separate issues that must be tackled (McCarthy, 1980). Four provisions of the law which are directly relevant to competency testing programs are nondiscriminatory testing, the Individualized Education Program (IEP), procedural and placement safeguards, and free appropriate public education. Much of the literature on the topic has addressed these provisions, especially the IEP (e.g., Amos, 1980; Baratz, 1978; Ewing & Smith, 1981; Gillespie & Lieberman, 1983; Lewis, 1979; Linde & Olsen, 1980; McClung & Pullen, 1978; Olsen, 1980; Rosewater, 1979, Ross & Weintraub, 1980; Safer, 1980; Serow & O'Brien, 1983; Smith & Jenkins, 1980).

The first problem that needs attention is the definition of "handicapped." At present, the U.S. Department of Education (1980) has identified nine categories of handicapping condition: speech impaired, learning disabled, mentally retarded, emotionally disturbed, deaf and hard of hearing, visually handicapped, multihandicapped, deaf and blind, and other health impaired. The classification of students into many of these categories is imprecise, for example, learning disabled (Berk, 1984e, chap. 1), and individuals can vary markedly in the severity of their condition.

Once this definitional issue has been settled and the benefits and costs of testing handicapped students have been weighed, it is not unreasonable to conclude that all students should be required to pass the minimum competency test to receive a regular diploma. As McCarthy (1980) observed:

> The use of a single standard for the awarding of the diploma does not imply that the preparation process for all children must be the same. The IEP is a means to an end and should be individualized, while the diploma is an end itself and can be based on universal criteria. (p. 172)

Certainly there are alternatives to this conclusion, such as awarding certificates of attendance and special diplomas (Grise, 1980; Ross & Weintraub, 1980). These alternatives have been upheld by several recent appellate court decisions (e.g., *Board of Education of Northport-East Northport* v. *Ambach,* 1982). Policy makers should examine carefully the alternatives and the anticipated impact on handicapped students before reaching their own conclusion.

THE FUTURE OF MINIMUM COMPETENCY TESTING

It is very risky to predict the success or even the direction of most politico-educational movements. (Actually the only danger is being wrong.) While the minimum competency testing movement was politically instigated, the momentum for change in the schools now rests with the professional educators. More than a decade has passed since a state legislature mandated the first minimum competency testing program. At present, nearly 40 states have mandated such programs, a number large enough to ratify an amendment to the U.S. Constitution. Any ideas proferred here regarding the success of these programs are merely conjectural at this time.

First, the public's dissatisfaction with the "rising tide of incompetents" or the "regression toward mediocrity" and the mounting evidence of increasing rates of illiteracy and incompetent high school graduates has demonstrated that "a serious and substantial educational problem faces the country today" (Lerner, 1981, p. 1062). The National Commission on Excellence in Education (1983) recently emphasized the scope of the problem. The minimum competency testing movement is the public's response to this problem, its best hope for at least a partial solution when no superior alternative is available.

Second, the success of minimum competency testing programs will probably hinge on the credibility and technical quality of the test and on the extent to which the program can be executed effectively. These goals will require the galvanized efforts of educators at all levels—a strong commitment to make the program work. The goals are not within the purview of legislators. The design of the testing program and, particularly, the setting of competency standards are the responsibilities of testing experts with the approval of the public.

The testing technology exists to develop minimum competency tests that are both psychometrically and legally defensible. The dozen issues discussed in the preceding section must be confronted and tackled if a program is to succeed. Despite the role of judgment and subjectivity in all of the procedures, from defining the domain of competencies to equating the scores on different test forms, there are sufficient precedents in other fields of competency testing to suggest that such procedures will survive legal scrutiny. These precedents take the form of specifications to guide competency testing practices in Section 430 of the 1978 Civil Service Reform Act, in the U.S. Equal Employment Opportunity Commission et al.'s (1978) *Uniform Guidelines on Employee Selection,* and in the *Principles for the Validation and Use of Personnel Selection Procedures* (APA, 1980), as well as in the *Standards for Educational and Psychological Tests* (APA/AERA/NCME Joint Committee, 1974). Furthermore, competency test applications in occupational licensing and certification and in the performance appraisal of employees have a history of litigation in the 1970s that has implications for minimum competency testing practices in education (e.g., *Al-*

bermarle Paper Company v. *Moody,* 1975; *Brito* v. *Zia Company,* 1973; *Dickerson* v. *U.S. Steel,* 1978; *Griggs* v. *Duke Power,* 1971; *Wade* v. *Mississippi Cooperative Extension Service,* 1974).

Third, a testing program is just the first step toward solving the incompetency problem. It furnishes only the means of certification or the mechanism for accountability. No test can improve competency levels; it just measures them. The test must be augmented with a competency-based education program to teach the competencies (Goldhammer & Weitzel, 1981; Spady, 1977). Descriptions of 13 exemplary programs throughout the country have been presented by McClure and Leigh (1981). They represent a variety of approaches that may concentrate upon classroom organization, curriculum development, teacher responsibility, learning packages, or integrated tasks (see Lasser & Olson, 1977; Schalock, 1976). As Nickse (1981) points out, however:

> Whatever versions ultimately predominate, and it seems certain that there will continue to be several, the competency-based approach to instruction will serve as a powerful management tool for formal and informal education both within and outside traditional institutions. (p. 223)

These trends in minimum competency testing and competency-based education during the past decade strongly indicate that public pressure for results and educator response to that pressure will continue and probably intensify in the 1990s. The state mandates for educational change demand immediate action and long-term planning, at least until the discontent over incompetence has abated and the meaning of the high school diploma has been restored.

ACKNOWLEDGMENT

I express my appreciation to Stanley Bernknopf, Ronald K. Hambleton, Stephen L. Koffler, and Robert L. Linn for their very useful suggestions on an earlier version of this chapter.

REFERENCES

AERA/APA/NCME Joint Committee. (in preparation). *Joint technical standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Airasian, P. W., Madaus, G. F., & Pedulla, J. J. (1979). *Minimal competence testing.* Englewood Cliffs, NJ: Educational Technology Publications.

Airasian, P. W., Pedulla, J. J., & Madaus, G. F. (1978). *Policy issues in minimal competency testing and a comparison of implementation models.* Boston: Heuristics.

Albemarle Paper Company v. Moody, U.S. Supreme Court Nos. 74–389 and 74–428, 10 FEP Cases 1181, 1975.

American Friends Service Committee. (1978). *A citizen's introduction to minimum competency programs for students.* Columbia, SC: Southeastern Public Education Program.

Amos, K. M. (1980). Competency testing: Will the LD student be included? *Exceptional Children, 47,* 194–197.

Anderson, R. C. (1972). How to construct achievement tests to assess comprehension. *Review of Educational Research, 42,* 145–170.

Anderson v. Banks. 540 F. Supp. 761 (S.D. Ga. 1982).

Andrew, B. J., & Hecht, J. T. (1976). A preliminary investigation of two procedures for setting examination standards. *Educational and Psychological Measurement, 36,* 35–50.

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Angoff, W. H. (1982). Use of difficulty and discrimination indices for detecting item bais. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 96–116). Baltimore, MD: Johns Hopkins University Press.

APA/AERA/NCME Joint Committee. (1974). *Standards for educational and psychological tests* (rev. ed.). Washington DC: American Psychological Association.

APA (Division of Industrial and Organizational Psychology). (1980). *Principles for the validation and use of personnel selection procedures* (2nd ed.). Berkeley, CA: Author.

Armstead et al. v. Starkville Mississippi Municipal Separate School District., No. EC 70–51–5 (N. D. Miss., 1972).

Arter, J. A. (1982, March). *Out-of-level versus in-level testing: When should we recommend each?* Paper presented at the annual meeting of the American Educational Research Association, New York.

Baca, L., & Chinn, P. C. (1982). Coming to grips with cultural diversity. *Exceptional Education Quarterly, 2,* 33–45.

Baker, E. L. (1974). Beyond objectives: Domain-referenced tests for evaluation and instructional improvement. *Educational Technology, 14*(6), 10–16.

Baratz, J. (1978, July). *In setting minimum standards, have we abandoned concerns for equity and access?* Washington, DC: Educational Testing Service.

Beard, J. G. (1979). Minimum competency testing: A proponent's view. *Educational Horizons, 58,* 9–13.

Behuniak, P., Jr., Archambault, F. X., & Gable, R. K. (1982). Angoff and Nedelsky standard setting procedures: Implications for the validity of proficiency test score interpretation. *Educational and Psychological Measurement, 42,* 247–255.

Bergan, J. R., Cancelli, A. A., & Luiten, J. W. (1980). Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. *Journal of Educational Statistics, 5,* 65–81.

Berk, R. A. (1976). Determination of optimal cutting scores in criterion-referenced measurement. *Journal of Experimental Education, 45,* 4–9.

Berk, R. A. (1978). The application of structural facet theory to achievement test construction. *Educational Research Quarterly, 3,* 62–72.

Berk, R. A. (1979a). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83,* 460–472.

Berk, R. A. (1979b, April). *Technical issues in setting standards for minimum competency testing.* Invited address at the Regional Minimum Competency Testing Conference sponsored by the Basic Skills Assessment Consortium and Educational Testing Service, San Francisco.

Berk, R. A. (1980a). A comparison of six content domain specification strategies for criterion-referenced tests. *Educational Technology, 20*(9), 49–52.

Berk, R. A. (1980b). A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement, 17,* 323–349; Erratum, 1981, *18,* 131.

Berk, R. A. (Ed.). (1980c). *Criterion-referenced measurement: The state of the art.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1980d). A framework for methodological advances in criterion-referenced testing. *Applied Psychological Measurement, 4,* 563–573.

Berk, R. A. (Ed.) (1982). *Handbook of methods for detecting test bias.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1983, April). *Advances in the field of measurement: Job analysis, passing scores, and test score equating.* Invited address at the Educational Testing Service Seminar on New Frontiers in Assessment for the Medical and Health Professions, Washington, DC.

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

Berk, R. A. (1984b). Criterion-referenced tests. In T. Husén & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies.* Oxford, England: Pergamon Press.

Berk, R. A. (Ed.) (1984c). *A guide to criterion-referenced test construction.* Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1984d). Selecting the index of reliability. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 231–266). Baltimore, MD: Johns Hopkins University Press.

Berk, R. A. (1984e). *Screening and diagnosis of children with learning disabilities.* Springfield, IL: Charles C Thomas.

Berk, R. A. (1985, April). *A consumers' guide to setting performance standards on criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Berk, R. A. (Ed.). (in press). *Performance assessment: Methods and applications.* Baltimore, MD: Johns Hopkins University Press.

Bernknopf, S. (1980, April). *Cut-scores and alternate forms: A new frontier or back to the trenches.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Bernknopf, S., Curry, A., & Bashaw, W. L. (1979, April). *A defensible model for determining a minimal cut-off score for criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Bersoff, D. N. (1979). Regarding psychologists testily: Legal regulation of psychological assessment in the public schools. *Maryland Law Review, 39,* 27–120.

Bersoff, D. N. (1982a). *Larry P.* and *PASE*: Judicial report cards on the validity of individual intelligence tests. In T. R. Kratochwill (Ed.), *Advances in school psychology* (Vol. II) (pp. 61–95). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bersoff, D. N. (1982b). The legal regulation of school psychology. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 1043–1074). New York: Wiley.

Block, J. (1972). Student learning and the setting of mastery performance standards. *Educational Horizons, 50,* 183–190.

Board of Education of Northport-East Northport v. Ambach, 458 N.Y.S. 2d 680 (A.D. 1982).

Bormuth, J. R. (1970). *On the theory of achievement test items.* Chicago: University of Chicago Press.

Brennan, R. L. (1984). Estimating the dependability of the scores. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 292–334). Baltimore, MD: Johns Hopkins University Press.

Brennan, R. D., & Lockwood, R. E. (1980). A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement, 4,* 219–240.

Brito v. Zia Company, 478 F.2d. 1200 (1973).

Burrill, L. E. (1982). Comparative studies of item bias methods. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 161–179). Baltimore, MD: Johns Hopkins University Press.

Butera, B. J., & Raffeld, P. C. (1979, April). *Investigation of vertical scaling for out-of-level*

testing. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Castro, J. G., & Jordan, J. E. (1977). Facet theory attitude research. *Educational Researcher, 6* (11), 7–11.

Chickering, A., & Claxton, C. (1981). What is competence? In R. Nickse & L. McClure (Eds.), *Comptency-based education: Beyond minimum competency testing* (pp. 5–41). New York: Teachers College Press.

Christie, S. G., & Casey, J. A. (1983). Heading off legal challenges to local minimum competency programs. *Educational Evaluation and Policy Analysis, 5,* 31–42.

Citron, C. H. (1982). Competency testing: Emerging principles. *Educational Measurement: Issues and Practice, 1*(4), 10–11.

Citron, C. H. (1983a). Courts provide insight on content validity requirements. *Educational Measurement: Issues and Practice, 2*(4), 6–7.

Citron, C. H. (1983b, March). *Legal rules for student competency testing* (ECS Issuegram No. 36). Denver, CO: Education Commission of the States.

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement, 5,* 115–124.

Cleary, T. A., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. *American Psychologist, 30,* 15–41.

Cody, E. (1983, May 22). Diplomaless Florida "graduate" shocked. *The Washington Post,* p. A5.

Cohen, D., & Haney, W. (1980). Minimums, competency testing, and social policy. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency achievement testing: Motives, models, measures, and consequences* (pp. 5–22). Berkeley, CA: McCutchan.

College Entrance Examination Board. (1977). *On further examination: Report of the advisory panel on the SAT decline.* New York: Author.

Colton, D. A., & Hecht, J. T. (1981, April). *A preliminary report on a study of three techniques for setting minimum passing scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, Los Angeles.

Conoley, J. C., & O'Neil, H. F., Jr. (1979). A primer for developing test items. In H. F. O'Neil, Jr. (Ed.), *Procedures for instructional systems development* (pp. 95–127). New York: Academic Press.

Cureton, E. E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 621–694). Washington, DC: American Council on Education.

Debra P. v. Turlington, 474 F. Supp. 244, 265 (M.D. Fla. 1979).

Debra P. v. Turlington, 644 F.2d 397, 404 (5th Cir. 1981).

Debra P. v. Turlington, 564 F. Supp. 177 (M.D. Fla. 1983), appeal pending.

Dickerson v. United States Steel Corp., 582 F.2d. 827, 17 FEP 1589 (3d Cir. 1978).

Ebel, R. L. (1979). *Essentials of educational measurement* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Elford, G. (1977, May). *A review of policy issues related to competency testing for high school graduation.* Paper presented at the annual meeting of the New England Educational Research Association, Manchester, NH.

Emrick, J. A. (1971). An evaluation model for mastery testing. *Journal of Educational Measurement, 8,* 321–326.

Ewing, N. J., & Smith, J. E., Jr. (1981). Minimum competency testing and the handicapped. *Exceptional Children, 47,* 523–524.

Finn, P. J. (1975). A question writing algorithm. *Journal of Reading Behavior, 7,* 341–367.

Fisher, T. H. (1978). Florida's approach to competency testing. *Phi Delta Kappan, 59,* 599–602.

Fisher, T. H. (1983). Implementing an instructional validity study of the Florida high school graduation test. *Educational Measurement: Issues and Practice, 2*(4), 8–9.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (pp. 695–763). Washington, DC: American Council on Education.

Fremer, J. (1978). In response to Gene Glass. *Phi Delta Kappan, 59,* 605–606, 625.

Gale, L. E., & Pol, G. (1975). Competence: A definition and conceptual scheme. *Educational Technology, 15*(6), 19–25.

Gallup, G. H. (1978). *A decade of Gallup polls of attitudes toward education, 1969–1978.* Bloomington, IN: Phi Delta Kappan.

Gallup, G. H. (1979). The 11th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 61,* 33–46.

Gallup, G. H. (1980). The 12th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 62,* 33–46.

Gallup, G. H. (1981). The 13th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 63,* 33–47.

Gallup, G. H. (1982). The 14th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 64,* 37–50.

Gallup, G. H. (1983). The 15th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 65,* 33–47.

Gallup, G. H. (1984). The 16th annual Gallup poll of the public's attitudes toward the public schools. *Phi Delta Kappan, 66,* 23–38.

Georgia Association of Educators v. Nix, 407 F. Supp. 1102 (1976).

Gillespie, E. B., & Lieberman, L. M. (1983). Individualizing minimum competency testing for learning-disabled students. *Journal of Learning Disabilities, 16,* 565–566.

Glass, G. V. (1978a). Matthew Arnold and minimum competence. *Educational Forum, 42,* 139–144.

Glass, G. V. (1978b). Minimum competence and incompetencies in Florida. *Phi Delta Kappan, 59,* 602–605.

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

Goldhammer, K., & Weitzel, B. (1981). What is competency-based education? In R. Nickse & L. McClure (Eds.), *Competency-based education: Beyond minimum competency testing* (pp. 42–61). New York: Teachers College Press.

Gonzales, E. (1982). Issues in the assessment of minorities. In H. L. Swanson & B. L. Watson, *Educational and psychological assessment of exceptional children: Theories, strategies, and applications* (pp. 375–389). St. Louis: Mosby.

Gorth, W. P., & Perkins, M. R. (1979a, December). *A study of minimum competency testing programs* (Final program development resource document). Amherst, MA: National Evaluation Systems.

Gorth, W. P., & Perkins, M. R. (1979b). *A study of minimum competency testing programs* (Final typology report). Amherst, MA: National Evaluation Systems.

Greene, K., & Gay, R. (1980). *Occupational regulation in the United States.* Washington, DC: Employment and Training Administration, U.S. Department of Labor.

Griggs v. Duke Power Company, 401 U.S. (1971), 3 EPD 8137.

Grise, P. J. (1980). Florida's minimum competency testing program for handicapped students. *Exceptional Children, 47,* 186–191.

Guttman, L. (1970). Integration of test design and analysis. In *Proceedings of the 1969 Invitational Conference on Testing Problems.* Princeton, NJ: Educational Testing Service.

Halpin, G., Sigmon, G., & Halpin, G. (1983). Minimum competency standards set by three divergent groups of raters using three judgmental procedures: Implications for validity. *Educational and Psychological Measurement, 43,* 185–196.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80–123). Baltimore, MD: Johns Hopkins University Press.

Hambleton, R. K. (1982). Advances in criterion-referenced testing technology. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 351–379). New York: Wiley.

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

Hambleton, R. K., & Eignor, D. R. (1978). Guidelines for evaluating criterion-referenced tests and test manuals. *Journal of Educational Measurement, 15,* 321–327.

Hambleton, R. K., & Eignor, D. R. (1980). Competency test development, validation, and standard setting. In R. M. Jaeger & C. K. Tittle (Eds.), *Minimum competency testing: Motives, models, measures, and consequences* (pp. 367–396). Berkeley, CA: McCutchan.

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

Hambleton, R. K., & Powell, S. (1983). A framework for viewing the process of standard setting. *Evaluation and the Health Professions, 6,* 3–24.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. (1978). Criterion-referenced testing and measurement: A review of technical issues and developments. *Review of Educational Research, 48,* 1–47.

Haney, W., & Madaus, G. F. (1978). Making sense of the competency testing movement. *Harvard Educational Review, 48,* 462–484.

Hardy, R. (1983, April). *Measuring instructional validity: A report of an instructional validity study for the Alabama High School Graduation Examination.* Paper presented at the annual meeting of the American Educational Research Association, Montreal.

Hively, W., Maxwell, G., Rabehl, G., Sension, E., & Lundin, S. (1973). *Domain-referenced curriculum evaluation: A technical handbook and a case study from the MINNEMAST Project* (CSE Monograph Series in Evaluation, No. 1). Los Angeles: Center for the Study of Evaluation, University of California.

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating.* New York: Academic Press.

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

Hoover, H. D. (1982, March). *Some comments on vertical equating using item response theory.* Symposium paper presented at the American Educational Research Association, New York.

Huynh, H. (1976a). On the reliability of decisions in domain-referenced testing. *Journal of Educational Measurement, 13,* 253–264.

Huynh, H. (1976b). Statistical consideration of mastery scores. *Psychometrika, 41,* 65–78.

Huynh, H., & Saunders, J. C. (1979, April). *Bayesian and empirical Bayes approaches to setting passing scores on mastery tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Ironson, G. H. (1982). Use of chi-square and latent trait approaches for detecting item bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 117–160). Baltimore, MD: Johns Hopkins University Press.

Jackson, C. D. (1975). On the report of the Ad Hoc Committee on Educational Uses of Tests with Disadvantaged Students: Another psychological view from the Association of Black Psychologists. *American Psychologist, 30,* 86–90.

Jaeger, R. M. (1978). *A proposal for setting a standard on the North Carolina High School Competency Test.* Paper presented at the annual meeting of the North Carolina Association for Research in Education, Chapel Hill.

Jaeger, R. M. (1981). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement, 18,* 23–38.

Jaeger, R. M. (1982). The final hurdle: Minimum competency achievement testing. In G. R. Austin & H. Garber (Eds.), *The rise and fall of national test scores* (pp. 223–246). New York: Academic Press.

Jaeger, R. M. (in press). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed.). Washington, DC: American Council on Education and National Council on Measurement in Education.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Jordan, J. E. (1978). Facet theory and the study of behavior. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 192–209). San Francisco, CA: Jossey-Bass.

Kleinke, D. J. (1980, April). *Applying the Angoff and Nedelsky techniques to the National Licensing Examination in Landscape Architecture.* Paper presented at the annual meeting of the National Council on Measurement in Education, Boston.

Klemp, G. O., Jr. (1979). Identifying, measuring, and integrating competence. In P. S. Pottinger & J. Goldsmith (Eds.), *New directions for experiential learning (No. 3): Defining and measuring competence* (pp. 41–52). San Francisco, CA: Jossey-Bass.

Knapp, T. R. (1977). The reliability of a dichotomous test item: A "correlationless" approach. *Journal of Educational Measurement, 14,* 237–252.

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17,* 167–178.

Kolen, M. J. (1981). Comparison of traditional and item response theory methods for equating tests. *Journal of Educational Measurement, 18,* 1–11.

Kolen, M. J., & Whitney, D. R. (1982). Comparison of four procedures for equating the Tests of General Educational Development. *Journal of Educational Measurement, 19,* 279–293.

Kriewall, T. E. (1972). *Aspects and applications of criterion-referenced tests* (IER Technical Paper No. 103). Downers Grove, IL: Institute for Educational Research.

Larry P. et al. v. Wilson Riles, Superintendent of Public Instruction for the State of California, et al., No. C–71–2270 (N. D. Cal.) (1979, Oct 11).

Larry P. et al. v. Wilson Riles, Superintendent of Public Instruction for the State of California, et al., Appeal No. 80–4027 (9th Cir) (1984, Jan 28).

Lasser, B. R., & Olson, A. L. (1977, April). *Strategies for implementation of competency based education programs.* Salem, OR: Oregon Competency Based Education Program, Northwest Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED 147 950).

Lazarus, M. (1981). *Goodbye to excellence: A critical look at minimum competency testing.* Boulder, CO: Westview Press.

Leinhardt, G. (1983). Overlap: Testing whether it is taught. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 153–170). Hingham, MA: Kluwer- Nijhoff.

Leinhardt, G., & Seewald, A. M. (1981). Overlap: What's tested, what's taught? *Journal of Educational Measurement, 18,* 85–96.

Leinhardt, G., Zigmond, N., & Cooley, W. W. (1981). Reading instruction and its effects. *American Educational Research Journal, 18,* 343–361.

Lerner, B. (1981). The minimum competence testing movement: Social, scientific, and legal implications. *American Psychologist, 36,* 1057–1066.

Lewis, D. (1979). Certifying functional literacy: Competency testing and implications for due process and equal educational opportunity. *Journal of Law and Education, 8,* 145–148.

Linde, J., & Olsen, K. R. (1980). *Minimum competency testing and handicapped students.* Lexington, KY: Mid-South Regional Resource Center.

Linn, R. L. (1979a). Issues of reliability in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 90–107). Washington, DC: National Council on Measurement in Education.

Linn, R. L. (1979b). Issues of validity in measurement for competency-based programs. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement* (pp. 108–123). Washington, DC: National Council on measurement in Education.

Linn, R. L. (1980). Issues of validity for criterion-referenced measures. *Applied Psychological Measurement, 4,* 547–561.

Linn, R. L. (1981). Measuring pretest-posttest performance changes. In R. A. Berk (Ed.), *Educational evaluation methodology: The state of the art* (pp. 84–109). Baltimore, MD: Johns Hopkins University Press.

Linn, R. L., Madaus, G. F., & Pedulla, J. J. (1982). Minimum competency testing: Cautions on the state of the art. *American Journal of Educaton, 91*(1), 1–35.

Livingston, S. A. (1975). *A utility-based approach to the evaluation of pass/fail testing decision procedures* (Report No. COPA-75–01). Princeton, NJ: Center for Occupational and Professional Assessment, Educational Testing Service.

Livingston, S. A. (1980). Choosing minimum passing scores by stochastic approximation techniques. *Educational and Psychological Measurement, 40,* 859–873.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Macready, G. B., & Dayton, C. M. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2,* 99–120.

Macready, G. B., & Dayton, C. M. (1980). The nature and use of state mastery models. *Applied Psychological Measurement, 4,* 493–516.

Madaus, G. F. (1981). NIE clarification hearing: The negative team's case. *Phi Delta Kappan, 63,* 92–94.

Madaus, G. F. (1983). Minimum competency testing for certification: The evolution and evaluation of test validity. In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 21–61). Hingham, MA: Kluwer-Nijhoff.

Madaus, G. P., Airasian, P. W., Hambleton, R. K., Consalvo, R. W., & Orlandi, L. R. (1982). Development and application of criteria for screening commercial, standardized tests. *Educational Evaluation and Policy Analysis, 4,* 401–415.

Marco, G. L. (1981). Equating tests in an era of test disclosure. In B. F. Green (Ed.), *New directions for testing and measurement (No. 11)—Issues in testing: Coaching, disclosure, and ethnic bias* (pp. 105–122). San Francisco, CA: Jossey-Bass.

McCarthy, M. M. (1980). Minimum competency testing and handicapped students. *Exceptional Children, 47,* 166–173.

McClung, M. S. (1978). Are competency testing programs Fair? Legal? *Phi Delta Kappan, 59,* 397–400.

McClung, M. S. (1979). Competency testing programs: Legal and educational issues. *Fordham Law Review, 47,* 651–712.

McClung, M. S., & Pullin, D. (1978). Competency testing and handicapped students, *Clearinghouse Review, 11,* 922–927.

McClure, L., & Leigh, J. (1981). A sampler of competency-based education at its best. In R. Nickse & L. McClure (Eds.), *Competency-based education: Beyond minimum competency testing* (pp. 81–147). New York: Teachers College Press.

Meskauskas, J. A. (1976). Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research, 46,* 133–158.

Miller, B. S. (Ed.). (1978). *Minimum competency testing: A report of four regional conferences.* St. Louis: CEMREL.

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

Millman, J. (1980). Computer-based item generation. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 32–43). Baltimore, MD: Johns Hopkins University Press.

National Assessment of Educational Progress. (1976). *Functional literacy: Basic reading performance.* Denver, CO: Author.

National Association of State Directors of Special Education. (1979). *Competency testing, special education, and the awarding of diplomas: A report of survey information.* Washington, DC: Author.

National Commission on Excellence in Education. (1983, April). *A nation at risk: The imperative for educational reform.* Washington, DC: U.S. Government Printing Office.

National Institute of Education. (1981). *Transcript of minimum competency testing: Clarification hearings, Washington, DC, July 8–10, 1981.* Washington, DC: Alderson Reporting Company.

National School Boards Association. (1978). *Minimum competency. A research report.* Denver, CO: Author.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14,* 3–19.

Nickse, R. (1981). Conclusion. In R. Nickse & L. McClure (Eds.), *Competency-based education: Beyond minimum competency testing* (pp. 220–223). New York: Teachers College Press.

Novick, M. R., & Lewis, C. (1974). Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3, pp. 139–158). Los Angeles: Center for the Study of Evaluation, University of California.

Oakland, T. D. (1980). Nonbiased assessment of minority group children. *Exceptional Education Quarterly, 1,* 31–46.

Oakland, T. D., & Laosa, L. M. (1977). Professional, legislative, and judicial influences on psychoeducational assessment practices in schools. In T. D. Oakland (Ed.), *Psychological and educational assessment of minority children* (pp. 21–51). New York: Brunner/Mazel.

Oakland, T. D., & Matuszek, P. (1977). Using tests in nondiscriminatory assessment. In T. D. Oakland (Ed.), *Psychological and educational assessment of minority children* (pp. 52–69). New York: Brunner/Mazel.

Olsen, K. R. (1980). Minimum competency testing and the IEP process. *Exceptional Children, 47,* 176–183.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and Psychological Measurement, 28,* 95–104.

Parents in Action on Special Education (PASE) et al. v. Joseph P. Hannon, General Superintendent of Schools in Chicago, et al., No. 74 C 3586 (N.D. Ill., July 7, 1980).

Peng, C-Y. J., & Subkoviak, M. J. (1980). A note on Huynh's normal approximation procedure for estimating criterion-referenced reliability. *Journal of Educational Measurement, 17,* 359–368.

Perkins, M. R. (1982). Minimum competency testing: What? Why? Why not? *Educational Measurement: Issues and Practice, 1*(4), 5–9, 26.

Phillips, S. W. (1983). Comparison of equipercentile and item response theory equating when the scaling test method is applied to a multilevel achievement battery. *Applied Psychological Measurement, 7,* 267–281.

Pine, S. M. (1977, March). Applications of item characteristic curve theory to the problem of test bias. In D. J. Weiss (Ed.), *Applications of computerized adaptive testing* (RR 77–1) (pp. 37–43). Minneapolis: Department of Psychology, Psychometric Methods Program, University of Minnesota.

Pipho, C. (1978). Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan, 59,* 585–588.

Pipho, C. (1983, March). *State activity: Minimum competency testing* (unpublished table). Denver, CO: Education Commission of the States.

Poggio, J. P., Glasnapp, D. R., & Eros, D. S. (1981, April). *An empirical investigation of the Angoff, Ebel, and Nedelsky standard-setting methods.* Paper presented at the annual meeting of the American Educational Research Association, Los Angeles.

Popham, W. J. (1974). Selecting objectives and generating test items for objectives-based tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3, pp. 13–25). Los Angeles: Center for the Study of Evaluation, University of California.

Popham. W. J. (1978a). *Criterion-referenced measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1978b). *Setting performance standards.* Los Angeles: Instructional Objectives Exchange.

Popham, W. J. (1981a). The case for minimum competency testing. *Phi Delta Kappan, 63,* 89–91.

Popham, W. J. (1981b). *Modern educational measurement.* Englewood Cliffs, NJ: Prentice-Hall.

Popham, W. J. (1983, April). *Issues in determining adequacy-of-preparation.* Symposium paper presented at the annual meeting of the American Educational Research Association, Montreal.

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

Poynor, L. (1978, March). *Instructional dimensions study data management procedures as exemplified by curriculum analysis.* Paper presented at the annual meeting of the American Educational Research Association, Toronto.

Public Law 94–142. Education for All Handicapped Children Act, S.6, 94th Congress [Sec 613(a)(4)] 1st session, June 1975. Report No. 94–168.

Reschly, D. J. (1979). Nonbiased assessment. In G. D. Phye & D. J. Reschly (Eds.), *School psychology: Perspectives and issues* (pp. 215–253). New York: Academic Press.

Reynolds, C. R. (1982a). Methods for detecting construct and predictive bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 199–277). Baltimore, MD: Johns Hopkins University Press.

Reynolds, C. R. (1982b). The problem of bias in psychological assessment. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 178–208). New York: Wiley.

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

Roid, G. H., & Haladyna, T. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Rosewater, A. (1979). *Minimum competency testing programs and handicapped students: Perspectives on policy and practice.* Washington, DC: Institute for Educational Leadership.

Ross, J. W., & Weintraub, F. J. (1980). Policy approaches regarding the impact of graduation requirements on handicapped students. *Exceptional Children, 47,* 200–203.

Roudabush, G. E. (1974, April). *Models for a beginning theory of criterion-referenced tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago.

Rowley, G. L. (1982). Historical antecedents of the standard-setting debate: An inside account of the minimal-beardedness controversy. *Journal of Educational Measurement, 19,* 87–95.

Safer, N. (1980). Implications of minimum competency standards and testing for handicapped students. *Exceptional Children, 46,* 288–290.

Samuda, R. S. (1975). *Psychological testing of American minorities: Issues and consequences.* New York: Harper & Row.

Sattler, J. M. (1982). *Assessment of children's intelligence and special abilities* (2nd ed.). Boston: Allyn & Bacon.

Saunders, J. C., Ryan, J. P., & Huynh, H. (1981). A comparison of two approaches to setting passing scores based on the Nedelsky procedure. *Applied Psychological Measurement, 5,* 209–217.

Scandura, J. M. (1973). *Structural learning I: Theory and research.* New York: Gordon & Breach.

Scandura, J. M. (1977). *Problem solving: A structural/process approach with instructional implications.* New York: Academic Press.

Schalock, H. D. (1976, October). *Alternative models of competency based education* (2nd ed.). Salem, OR: Oregon Competency Based Education Program, Northwest Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED 147 951)

Scheuneman, J. D. (1979). A new method of assessing bias in test items. *Journal of Educational Measurement, 16,* 143–152.

Scheuneman, J. D. (1982). A posteriori analyses of biased items. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 180–198). Baltimore, MD: Johns Hopkins University Press.

Schlesinger, I. M. (1978). On some properties of mapping sentences. In S. Shye (Ed.), *Theory construction and data analysis in the behavioral sciences* (pp. 181–191). San Francisco, CA: Jossey-Bass.

Schmeiser, C. B. (1982). Use of experimental design in statistical item bias studies. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 64–95). Baltimore, MD: Johns Hopkins University Press.

Schmidt, W. H. (1983a). Content biases in achievement tests. *Journal of Educational Measurement, 20,* 165–178.

Schmidt, W. H. (1983b, April). *Methods of examining mismatch.* Symposium paper presented at the annual meeting of the National Council on Measurement in Education, Montreal.

Schmidt, W. H., Porter, A. C., Schwille, J. R., Floden, R. E., & Freeman, D. J. (1983). Validity as a variable: Can the same certification test be valid for all students? In G. F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 133–151). Hingham, MA: Kluwer-Nijhoff.

Schoon, C. G., Gullion, C. M., & Ferrara, P. (1979). Bayesian statistics, credentialing examinations, and the determination of passing points. *Evaluation and the Health Professions, 2,* 181–201.

Selkow, P. (1984). *Assessing sex bias in testing: A review of the issues and evaluations of 74 psychological and educational tests.* Westport, CN: Greenwood Press.

Senior, J. R. (1976). *Toward the measurement of competencies in medicine* (Report of the Computer-Based Examination Project). Philadelphia, PA: National Board of Medical Examiners and American Board of Internal Medicine.

Serow, R. C., & O'Brien, K. (1983). Performance of handicapped students in a competency testing program. *Journal of Special Education, 17,* 149–155.

Shepard, L. A. (1980a). Standard setting issues and methods. *Applied Psychological Measurement, 4,* 447–467.

Shepard, L. A. (1980b). Technical issues in minimum competency testing. In D. C. Berliner (Ed.), *Review of research in education* (Vol. 8, pp. 30–82). Washington, DC: American Educational Research Association.

Shepard, L. A. (1982). Definitions of bias. In R. A. Berk (Ed.), *Handbook of methods for detecting test bias* (pp. 9–30). Baltimore, MD: Johns Hopkins University Press.

Shepard, L. A. (1983). Standards for placement and certification. In S. B. Anderson & J. S. Helmick (Eds.), *On educational testing* (pp. 61–90). San Francisco, CA: Jossey-Bass.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169–198). Baltimore, MD: Johns Hopkins University Press.

Shimberg, B. (1981). Testing for licensure and certification. *American Psychologist, 36,* 1136–1146.

Shimberg, B. (1982a). *Occupational licensing: A public perspective.* Princeton, NJ: Educational Testing Service.

Shimberg, B. (1982b). What is competence? How can it be assessed? In M. R. Stern (Ed.), *Power and conflict in continuing professional education* (pp. 17–37). Belmont, CA: Wadsworth.

Skakun, E. N., & Kling, S. (1980). Comparability of methods for setting standards. *Journal of Educational Measurement, 17,* 229–235.

Slinde, J. A., & Linn, R. L. (1977). Vertically equated tests: Fact or phantom? *Journal of Educational Measurement, 14,* 23–32.

Slinde, J. A., & Linn, R. L. (1979). A note on vertical equating via the Rasch model for groups of quite different ability and tests of quite different difficulty. *Journal of Educational Measurement, 16,* 159–165.

Smith, J., & Jenkins, D. (1980). Minimum competency testing and handicapped students. *Exceptional Children, 46,* 440–443.

Spady, W. G. (1977). Competency based education: A bandwagon in search of a definition. *Educational Researcher, 6*(1), 9–14.

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

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement, 11,* 263–267.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1975). A Bayesian decision-theoretic procedure for use with criterion-referenced tests. *Journal of Educational Measurement, 12,* 87–98.

Thurston, P., & House, E. R. (1981). The NIE adversary hearing on minimum competency testing. *Phi Delta Kappan, 63,* 87–89.

Tiemann, P. W., & Markle, S. M. (1983). *Analyzing instructional content: A guide to instruction and evaluation* (2nd ed.). Champaign, IL: Stipes.

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

Traub, R. E., & Rowley, G. L. (1980). Reliability of test scores and decisions. *Applied Psychological Measurement, 4,* 517–545.

United States v. State of North Carolina, 400 F. Supp. 343 (E.D.N.C. 1975).

United States v. State of North Carolina, 425 F. Supp. 789 (E.D.N.C. 1977).

United States v. State of South Carolina, 15 FEP Cases 1196 (D.C.S.C. 1977).

U.S. Department of Education. (1980). *To assure the free appropriate public education of all handicapped children* (Second Annual Report to Congress on Implementation of Public Law 94–142: The Education for All Handicapped Children Act). Washington, DC: Author.

U.S. Department of Health, Education and Welfare, Public Health Service. (1977, July). *Credentialing health manpower* (DHEW Publication No. (05) 77–50057). Washington, DC: Author.

U.S. Equal Employment Opportunity Commission, U.S. Civil Service Commission, U.S. Department of Labor, & U.S. Department of Justice. (1978). Uniform guidelines on employee selection procedures. *Federal Register, 43* (166), 38290–38309.

van der Linden, W. J. (1980). Decision models for use with criterion-referenced tests. *Applied Psychological Measurement, 4,* 469–492.

van der Linden, W. J. (1982). A latent trait method for determining intrajudge inconsistency in the Angoff and Nedelsky techniques of standard setting. *Journal of Educational Measurement, 19,* 295–308.

van der Linden, W. J., & Mellenbergh, G. J. (1977). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 1,* 593–599.

Wade v. Mississippi Cooperative Extension Service, 372 F. Supp. 126 (1974), 7EPD 9186.

Wiederholt, J. L., Cronin, M. E., & Stubbs, V. (1980). Measurement of functional competencies and the handicapped: Constructs, assessments, and recommendations. *Exceptional Education Quarterly, 1,* 59–73.

Wilcox, R. R. (1977a). New methods for studying equivalence. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox (Eds.), *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation, No. 6, pp. 66–76). Los Angeles: Center for the Study of Evaluation, University of California.

Wilcox, R. R. (1977b). New methods for studying stability. In C. W. Harris, A. P. Pearlman, & R. R. Wilcox (Eds.), *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation, No. 6, pp. 45–65). Los Angeles: Center for the Study of Evaluation, University of California.

Wilcox, R. R. (1979a). Comparing examinees to a control. *Psychometrika, 44,* 55–68.

Wilcox, R. R. (1979b). A lower bound to the probability of choosing the optimal passing score for a mastery test when there is an external criterion. *Psychometrika, 44,* 245–249.

Williams, R. L. (1970). Danger: Testing and dehumanizing Black children. *Clinical Child Psychology Newsletter, 9,* 5–6.

Williams, R. L. (1971). Abuses and misuses in testing black children. *Counseling Psychologist, 2,* 62–77.

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

Ysseldyke, J. E. (1979). Issues in psychoeducational assessment. In G. D. Phye & D. J. Reschly (Eds.), *School psychology: Perspectives and issues* (pp. 87–121). New York: Academic Press.