Shepard, L.A. (1980). "Technical Issues in Minimum Competency Testing." In D.C. Berliner (Ed.), Review of Research in Education, Vol. 8. Itasca, IL: F.E. Peacock Publishers.

Chapter 2

Technical Issues in Minimum Competency Testing

LORRIE SHEPARD
University of Colorado

Minimum competency tests are intended to return meaning to the high school diploma by requiring that students meet standards of basic competence. Many testing programs of this type also involve testing in early grades to prevent social promotion of children who do not possess the necessary skills. Minimum competency testing takes its rationale from the psychology of competency-based education and its technology from criterion-referenced testing. The most important features of this view of school learning are the explicitness of instructional goals and the use of tests to ensure their accomplishment. The purpose of the introductory sections of this chapter is to set minimum competency testing in the context of the current issues and technology of educational measurement, and to identify the special characteristics of minimum competency testing that distinguish it from similar measurement techniques. Minimum competency testing has political purposes and does not serve day-to-day classroom decisions; as a result it is fundamentally unlike the instructionally oriented measurement practices from which it takes its form. Therefore, the identification of similarities and differences between minimum competency testing and criterion-referenced testing (already well reviewed elsewhere) is a guiding theme throughout this chapter.

ANTECEDENTS AND DEFINITIONS

Individually Prescribed Instruction and Mastery Learning

Individually prescribed instruction (Glaser, 1968; Lindvall & Bolvin, 1967) and mastery learning (Block, 1971a, 1974; Bloom, 1968, 1971, 1976) are two different approaches to individualized instruction which have contributed to the rhetoric of minimum competency testing. Although they

Robert L. Linn, University of Illinois, and Robert Glaser, University of Pittsburgh, were the editorial consultants for this chapter.
The author also thanks Gene V Glass and Ronald K. Hambleton for their helpful comments.

reflect different views about whether the outcomes of education should be the same for all learners and about the allocation of resources for fast and slow learners, they have in common several features that increase the likelihood of successful instruction: learning tasks are clearly defined, individual assignments are made on the basis of what a student apparently knows and does not know, and tests keyed to the curriculum are used to determine when a student has mastered a particular skill and should pass on to the next. The importance of testing as an essential part of the teaching/learning process, rather than just a distant measure of its outcomes, is a hallmark of objectives-based instructional systems that have shaped both the philosophy and technology of minimum competency testing. Simplistically, tests are expected to improve learning by giving teachers and students clearer targets.

Glaser's (1963) individually prescribed instruction (IPI) is a relevant antecedent of minimum competency testing, primarily because it exemplifies the instructional context he had in mind when he introduced criterion-referenced testing. In his description of the Oakleaf Project, Glaser (1968) stressed the importance of a clearly defined continuum of educational objectives and tests that accurately report a student's level of performance. Glaser was not, however, so interested in individualizing instruction to ensure equal achievement for all pupils. In this regard, he and others seeking to tailor curricula and teaching for individuals (Cronbach, 1967; Suppes, 1966; Talmage, 1975) have not fostered the current emphasis on the same minimum for all students. Rather, this focus on common attainments has come from the mastery-learning paradigm and the public mood. In a more recent work, Glaser (1977) presented his view of adaptive education, a type of teaching which adjusts to individual differences by providing different environments and different means for seeking different goals. Glaser's notion of individualization would be realized if every pupil achieved as much as he or she possibly could.

Mastery learning has a slightly different emphasis, viz., to ensure that all pupils acquire what is presently learned by, say, only the top 25 percent of students. In elaborating the theory of mastery learning, Block (1971a, p. 5) specifically rejected programmed instructional curricula exemplified by Individually Prescribed Instruction and Stanford's Computer Assisted Instruction project (Atkinson, 1968; Suppes, 1966). Although these approaches improved upon older versions of programmed instruction by tailoring learning units to fit the needs of individual students, they were not engineered to guarantee that all of the students would master all of the units. Instead Bloom (1968, 1971) and Block (1971a, 1974) were inspired by Carroll's (1963) model of school learning to build a mastery-learning theory.

Carroll (1963, 1970) observed that aptitude for school learning could be conceptualized as differential learning rate. Able children are distinguished

from less able children primarily by the amount of time it takes them to learn new material. Although some tasks might be too difficult to learn even without time limits, Carroll concluded, optimistically, that nearly all children in school (95 percent) can learn nearly all that one would want to teach them if the time allowed for instruction were more consistent with each student's needs. Carroll also acknowledged that quality of instruction and the student's ability to understand instruction interact to influence the amount of time taken to achieve criterion performance.

Time and absolute standards of performance are the key variables in the mastery-learning model. Rather than allocating the resources of schooling equally, Bloom (1976) argued for differential attention so that each child would achieve the criterion level. This would be accomplished by a tutorial approach in which each student would be taught, tested, and continuously remediated. The resource of time is then differentially allocated to match individual learning rates. Although Block (1971b) claimed some efficiencies in learning by this model, which will lead to increases in the amount learned for all students, Barr and Dreeben (1977) concluded that the tutorial model is likely to neglect students who reach mastery quickly. Interestingly, minimum competency testing has inherited from mastery learning both the belief that learning will improve if standards are zealously adhered to and pursued and the accompanying problem that minimums may become maximums if excellent students are not urged on to further accomplishments.

Competency-based education (CBE) is a generic label applied to many different versions of individualized instruction and programmed learning. Some approaches have the same general behaviorist origins as Glaser's IPI (Keller, 1968); other approaches have specifically adopted the logic and philosophy of the mastery model (Kulik & Kulik, 1976; Spady, 1977). CBE has been most widely attempted in college teaching (Robin, 1976; Trivett, 1975) and in teacher training programs (Dickson, 1975). The only additional meaning attached to CBE by some is the life-role definition of competencies given by Spady (1977). Rather than equating competencies with academic skills, more applied behaviors or performances are implied. Therefore, in addition to the characteristics of individually paced progress and unit-by-unit mastery, competency-based education is often distinguished by instructional goals that are tied to success in adult life.

Criterion-referenced Testing

Glaser (1963) introduced criterion-referenced tests as more appropriate measures, not only for monitoring the progress of students in objectives-based or programmed instructional systems, but also for evaluating the effectiveness of instruction. Unlike existing norm-referenced tests, which

only report an individual's relative standing in a group, criterion-referenced tests were more carefully keyed to a performance continuum. Test scores supported by this type of referencing provide more information about exactly what a student knows and does not know. This information may be useful in the day-to-day planning of instruction aimed at the next-most difficult topic a student is to learn.

The benefits and attributes of criterion-referenced tests have been further explicated by Popham (1975, 1978b) and Millman (1974). Although the field has suffered some lexicographic meanderings recounted by Glass (1978b) and Popham (1975), there is now some consensus supporting Popham's more recent definition: "A criterion-referenced test is used to ascertain an individual's status with respect to a well-defined behavior domain" (1975, p. 130). This is the definition adopted by Hambleton, Swaminathan, Algina, and Coulson (1978) in their extensive review of technical issues and is essentially synonymous with domain-referenced testing (Hively, 1974; Hively, Maxwell, Rabehl, Sension, & Lunden, 1973; Millman, 1974).

By this definition, criterion-referenced tests have well-specified universes of generalizations, and items are sampled or selected to allow accurate estimation of domain scores. An incidental feature of criterion-referenced tests, which also distinguishes them from traditional norm-referenced survey tests, is that they typically have many more items; better measurement is ensured simply by being more thorough as well as by carefully representing the content domain.

Norm-referenced tests have many important uses, especially when an overview of achievement is desired and when comparative data are needed to judge the merits of outcomes (Ebel, 1978b; Shepard, 1979a). Norms might even be built for criterion-referenced tests since domain scores do not carry with them any information about the goodness or badness of attainments (Popham, 1976). Nevertheless, criterion-referenced tests are the preferred measures for day-to-day instructional decisions (National Academy of Education, 1978). They are also the obvious source for the technology of minimum competency tests because both testing purposes require measurement of how much of the intended content a student knows rather than his or her relative standing in a group.

Mastery Testing and Competency Testing

Mastery tests are intended to be used to separate the competent from the incompetent. Unlike a criterion-referenced test, which locates an individual along a well articulated performance *continuum*, mastery tests require a cut-score (on that continuum) to distinguish acceptable from unacceptable performance. Much of the controversy over the definitions of criterion-referenced tests has been about whether a cut-score or standard is essential.

Some early proponents of an absolute rather than relative standard spoke of the "criterion" in Glaser's term criterion-referenced as if it were a standard or cut-score (Popham & Husek, 1969). Now the intended behavioral referencing discussed above is more fully appreciated.

The debate and reclarification of the definition of criterion-referenced tests have served to separate careful content specification inherent in domain-referenced testing from the cut-score problem. Authors of major reviews (Hambleton et al., 1978; Subkoviak & Baker, 1977) identify two purposes for criterion-referenced tests: to estimate an individual's proportion-correct score, or to make mastery-nonmastery decisions. Only in the second case is a cut-score needed. One of the reasons that the meaning of the term criterion might have been confused earlier is that in instructional settings for which criterion-referenced tests were originally intended cut-scores are clearly needed. In fact, Nitko (1974) described Glaser's proposal for criterion-referenced tests as a combination of minimum goals set for individuals (Flanagan, 1951) and standard content domains (Ebel, 1962). Anytime that a test is used to make black-and-white decisions about the placement of students (e.g., remediate or not, repeat sixth-grade or not), a cut-score is required. Obviously, minimum competency tests require cut-scores to fail those who are not minimally competent. It would, therefore, be a digression from the topic of minimum competency testing to pursue statistical methods for domain-score estimation and those test uses that do not require standards. However, it is useful to reiterate that accurate articulation of a test with a behavioral domain and standard setting are separate problems (both must be addressed to undertake minimum competency testing).

Minimum Competency Testing

Minimum competency tests are mastery tests intended to sort examinees into two categories, masters and nonmasters. The competencies are the substance the test measures, either performances or knowledge. Given some definitions of competence, the term minimum is redundant since the competencies themselves are usually identified because they are considered essential or mandatory for whatever level of education is being undertaken. Nevertheless, because mastery learning and competency-based approaches can be applied in advanced subject areas, such as college calculus or honors English, the term minimum is used to emphasize that the competencies are only those considered absolutely necessary to pass the gate guarded by the test—to leave high school or to enter adult life. This definition corresponds to a representative quotation from the report of four regional conferences on minimum competency testing, "Minimum competencies are the basic

proficiencies in skills and knowledge needed to perform successfully in real life activities" (Miller, 1978, p. 13).

The above definition is inconsistent, however, with that offered by Hambleton and Eignor (1979). They equated competency tests with criterion-referenced tests, both being described by a well-defined behavior domain. *Minimum* competency tests, in their terminology, refer to the second use of criterion-referenced tests, which is to make dichotomous classifications of masters and nonmasters. While this is of course the purpose of minimum competency testing, it is somewhat misleading to suggest that the word minimum means that a small amount of the competency must be learned. Rather, folkways make it customary to expect that large portions (usually 70 percent or higher) of each low-level or minimal competency be correct to attain mastery. High standards are consistent with the rationale for mastery learning (Block, 1971a) and the 85 percent criterion originally set in Individually Prescribed Instruction (Glaser, 1968). Instead of inferring that the term minimum refers either to low standards or to the existence of cut-scores not found in other competency tests, general usage suggests that minimum is used to distinguish competencies that are essential, such as basic addition facts, from those that are not, such as being able to bisect an angle with a compass and ruler. That there remains some ambiguity on this point, however, is illustrated by Brickell's (1978) query whether there should be more than one minimum, that is, more than one cut-score for students of different abilities.

Minimum competency testing is part of the back-to-basics movement and is believed by many to be the necessary solution to the decline in test scores (Ebel, 1978a; Rickover, 1978). Pessimistic proponents believe it will give meaning to the high school diploma by denying the credential to the incompetent; other advocates share the optimism of the behaviorists, believing that the existence of the standards will in themselves increase learning. The complexities of the sociopolitical origins of minimum competency testing are discussed by Resnick in another chapter of this volume. Clearly, the technical issues of reliability, validity, and standard setting are influenced by the external accountability purposes of minimum competency testing programs. Largely because of the seriousness of the consequences, technical problems which could be ignored in an instructional setting will be magnified. The further removed the testing is from the classroom (i.e., state rather than locally administered, or high school graduation, only, rather than year-to-year promotion), the more strained is the comparison of minimum competency testing with the use of tests in objectives-based instructional programs. The similarities are not enhanced much even when minimum competency testing is combined with mandatory remediation courses. In the individualized instructional setting, the testing is frequent and bad decisions are easily corrected. In minimum competency

testing, the consequences are more serious than studying the wrong lesson for a week and are not as easily redressed. In addition, the more gross or general the measure of achievement, as is the case when one or two tests cover 12 years of public school education, the poorer the match is likely to be between tests and instruction. Minimum competency tests may be developed to be criterion-referenced but may not be used as an integral part of instruction as originally envisioned. These issues have special bearing on the validity of minimum competency tests.

TEST CONSTRUCTION AND VALIDITY

The usual organization of a chapter on test theory is to consider chronologically the steps of test construction and the various methods appropriate for collecting evidence of validity. This strategy is not useful, however, because it requires repeating all of the assumptions made during test development when considering validity and leaves an incorrect impression that validation is a single step which follows the construction of the test. Only in strictly actuarial circumstances, where conceptual validity is not at issue, can validity be established by post hoc statistical analysis. In most instances, the validity of a test depends on both the logic of the test development and the empirical evidence gathered at each stage of development. It is useful, therefore, to consider the validity issues raised by inferences made at each step of the test development process and to identify the corresponding methods appropriate for each step.

Measurement validation is a crucial part of any investigative effort in the social sciences, because tests are only approximations of the underlying traits or behaviors one wishes to observe. Inference is, therefore, always required. Validity has to do with how faithfully test performance reflects the intended attribute being assessed. It may be thought of as the accuracy (Millman, 1974) or appropriateness (American Psychological Association, 1974) of the inferences made from test scores. In his comprehensive essay on test validation, Cronbach (1971) emphasized that validity is not an inherent characteristic of a test but rather depends on its use. Measures will have different degrees of validity for different purposes because different interpretations are implied. It is the leap from the test score to an assumed characteristic which must be validated. The soundness of these inferences can be examined by recognizing the incremental inferences made at each step in the test development process. As will be discussed in the next sections, the validity of a minimum competency test depends on validity at several stages: selection of the right domain, how well the domain is defined and explicated, and whether test items are selected to adequately represent the domain.

Domain Selection

The purpose of most minimum competency testing programs is to ensure that high school graduates have the skills necessary for success in adult life (Pipho, 1977, 1978). In North Carolina, for example, legislators have instructed the state board to give tests which ensure that graduating students "possess those skills and that knowledge necessary to function independently and successfully in assuming the responsibilities of citizenship" (Pipho, 1977, p. 40). The highest level of inference for this purpose corresponds to the first and last steps in the test development process; the skills included in the test must be a contributory cause in "life-role success." If the domain of the test is not selected with this purpose in mind and the connections verified, it is *unlikely* that the test can be considered valid.

The literature on criterion-referenced testing is not helpful for this phase of the development-validation process. Most authors (Fremer, 1974; Hambleton et al., 1978; Millman, 1974; Popham, 1975) assume that the intended domain has already been fairly well circumscribed by someone else and begin with directions for further elaborating the domain to facilitate accurate representation. Hambleton and Eignor (1979) acknowledged that "competencies must be prepared or selected before the test development process can begin" (p. 8), but the advice they offer for test construction is more applicable to domain specification, considered in the next section. The necessity for predictive validity is addressed only briefly by Subkoviak and Baker (1977) in their review of criterion-referenced measurement:

> When a test is designed for the purpose of classifying individuals as masters or nonmasters, there is generally the belief that test results are related to outcomes on other variables such as success or failure in subsequent endeavors. Thus, predictive validity, or the strength of relation between test results and other outcomes, is a primary consideration. (p. 294)

Both Harris (1974b) and Millman (1974) are cited as sources on the topic of validating dichotomous classifications. However, both of these authorities are interested in the problem in a circumscribed instructional setting. They do not have to be concerned with the higher level of inference required by minimum competency testing. Rather, as Harris wrote, the test construction process "begins with a careful specification of what is to be learned and how it is to be learned" (p. 109). Nonetheless, Harris also acknowledged that appropriate criteria for the validity of mastery tests are performance on a transfer task and degree of subsequent success (in an instructional sequence). The misfit of criterion-referenced methodology for minimum competency testing is most obvious in the proximal-distal criterion problem. With proximal academic goals it is reasonable to assume that the only effort required is careful explication of the domain; but for distant goals there is the more difficult question of what the goals should be. No one is certain what

skills prepare students for a good life. Furthermore, the kinds of long-range field studies that would have to be undertaken to answer such questions would be on a scale greater than present educational research efforts on any topic including reading.

Madaus (1978) is one of the few authors to consider more explicitly the problem of predictive validity which exists if minimum competency tests are linked to adult functioning. He envisioned an enormous validation task, since a criterion measure of successful adult performance would first have to be agreed upon. Madaus drew a parallel to the difficulties encountered in employment testing; but actually the problems would be much larger because the prediction would have to be accurate not for one job but for all possible jobs, and for success off the job as well. Moreover, if one had an acceptable definition of success there would still be the difficulty of establishing a relationship between performance on the test and the subsequent criterion. An extensive treatment of validity issues in competency-based measurement is available in Linn (1979b). He also likened the problem of what to include in the test, given the purpose of the test, to the problem of demonstrating job relevance in employment testing (see Linn, 1976). For employment purposes, tests such as traditional IQ measures may not be justified as selection devices solely because they correlate with job success; the direct relationship of test content to job performance must be demonstrated by means of job analysis (*Griggs et al.*;[1] Civil Service Commission, 1977). Anastasi (1976) gave a useful summary of the kinds of data-gathering activity that would enlighten a job analysis. To be effective, the job analysis must focus on "those aspects of performance that differentiate most sharply between the better and poor workers" (p. 437). In the context of minimum competency testing, this would mean, for example, that modest correlations of test results with parental socio-economic status (as offered by Hills, 1979) would not be adequate evidence of validity. Rather, skills such as the ability to fill out an income tax form would have to discriminate well between successful and unsuccessful adults. If the counter examples are very numerous (i.e., too many businessmen hire accountants instead of completing their tax returns), then some other content must be sought that is more accurately a prerequisite for success.

Novick (1979) commented on the hopelessness of using even a test-anchored diploma as an adequate screening device for a large number of employers. It would perhaps be more reasonable, he suggested, to construct many different job-specific tests to be administered by employers rather than schools. Hambleton (1979) may reflect the recent sentiments of many educators who are tackling the validity problems by narrowing the claims associated with minimum competency testing. Whereas it might be nearly impossible to demonstrate the validity of measures of life skills, validation of tests of basic skills is much more straightforward. Therefore, one could

argue that schools are on surer ground if they insist on basic academic skills as exit requirements. A change in emphasis to more school-relevant basic skills may especially be prompted by validity issues identified by legal analysts. McClung (1977, 1978), a lawyer, introduced the terms curricular validity and instructional validity (cf. Getz & Glass, 1979). He suggested that minimum competency testing programs would violate due process of law if the tests measured objectives that students had not been taught. Not only must all test elements be apparent in the curriculum, there must also be evidence that the intended instructional goals were actually taught. The choice between life skills and basic skills was one of the issues identified in *Minimum Competency Testing: A Report of Four Regional Conferences* (Miller, 1978).

Two observations can be made about the switch to basic skills as the focus of minimum competency testing:

(1) The change is a contradiction of the earlier popular definition of competency-based education. Spady (1977) specifically states that competencies were life skills not basic academic skills.

(2) The original validity conundrum is not solved if the purpose of giving the basic skills test is still to ensure successful adult life.

In fact, even greater inferences are thereby required about transfer of training to life beyond school, since these applications would no longer even be simulated in the test.

Whether minimum competency tests are designed to measure life skills or basic skills, all of the traditional types of validity evidence described in the *Standards for Educational and Psychological Tests* (American Psychological Association, 1974) are required. Content validity, which is demonstrated if test items are a representative sample of the behavioral domain, is addressed in the section on domain definition. Both content validity and predictive validity, discussed above, are necessary but not sufficient to establish the construct validity of minimum competency tests. Although proponents of criterion-referenced tests initially argued for exclusive adherence to logical rather than empirical tests of validity (Harris, 1974a; Millman, 1974), the need for construct validation or at least predictive validity is now more widely accepted (Hambleton et al., 1978; Messick, 1975; Popham, 1978). What is generally true for criterion-referenced tests is especially applicable to minimum competency tests. Linn, for example, noted that "the very word 'competency' implies a construct" (1979b, p. 119).

Construct validity is the more inclusive and demanding type of validity. It is required whenever test scores are used to draw inferences about an underlying capability, that is, whenever behaviors represented on the test are not directly of interest but are only proxies for the intended criterion behaviors. The theoretical relationship of test performance to other performances must be confirmed by reality. These connections should be

demonstrated logically as the test is constructed and should be confirmed, after the fact, by both correlational and experimental studies (Cronbach, 1971). The predictive validity studies for high school graduation tests are one kind of correlational investigation. When competency tests are used for grade-to-grade advancement decisions it is necessary to show that students retained benefit from additional instruction and could not have functioned with higher level curricula (Linn, 1979b). Nitko (1974) offered an extensive discussion of the kind of study appropriate for validating hierarchies and curriculum sequences. The more removed testing is from day-to-day instruction the more important it is to determine its validity for placement decisions. Finally, testing and required remediation are an intervention whose effectiveness can be tested experimentally. Ultimately, the validity of competency testing depends not only on the predictive relationship between the test and success in life, but also on an *increase* in "adult functioning" for those who fail the test at first but eventually meet the standard.

The picture is bleak for solving these validity problems. Inventing new technology is not likely to help. In other situations, better testing methodology has improved the validity of test use by achieving a better match between the test behaviors and the predicted criterion. There is no prospect for a better match in minimum competency testing programs, however, so long as the test is meant to anticipate the requirements for all possible life roles. The only entirely defensible conclusion is that the technology is not up to the job of certifying competency for high school graduation. A possible intermediate position is to recognize that tests are more likely to have validity when testing is proximal to the instructional use: then, the link to the curriculum is direct, remediation is feasible, and inferences are not made about the transfer of skills to nonacademic performance. Members of the National Academy of Education (1978) have concluded, for example, that a series of competency tests in the early grades, used for diagnostic purposes, could be workable, while high school graduation standards are not. These issues are discussed at greater length in the final section of this chapter where recommendations are distinguished by test use.

Domain Definition

Detailed specifications of the domain to be assessed ensures that (1) the important behaviors are clearly identified so that the logical contribution to construct validity can be judged, and that (2) the match of test items to the domain will be enhanced by unambiguous definition. Although traditional norm-referenced test development rules call for specification of the content universe, recent developments in criterion-referenced testing offer guidelines for accomplishing much greater precision in this endeavor. Minimum

competency programs could conceivably adopt existing standardized tests and set minimum passing scores (Madaus, 1978; Pipho, 1977); but most experts offering technical advice assume that since those tests were not designed for such a purpose and could hardly meet even content validity requirements (and surely not construct validity demands), that competency tests will be developed following the criterion-referenced paradigm (Fremer, 1978; Hambleton & Eignor, 1979). Traditional norm-referenced tests are survey tests covering a wide array of objectives within a particular subject. Therefore, they would not provide in-depth assessment of specific competencies and would often include material even in basic skills tests that would not be considered a minimal requirement.

Ebel (1962) was among the first to propound "content-standard" tests. The crucial characteristic of such tests is that "the processes by which the scores are obtained—the test construction, administration, and scoring—are explicit and objective enough so that independent investigators would obtain substantially the same scores for the same persons" (p. 16). His purpose was to structure test development well enough so that percent scores for a fixed content domain would have a standard meaning. Behavioral objectives (Mager, 1962) were the first strategem adopted in an effort to specify test content systematically. Objective-referencing has subsequently been rejected, however, because it offers too vague a definition of the content (Baker, 1972; Hively, Patterson & Page, 1968; Millman, 1974; Popham, 1975). As Millman noted, even with behaviorally stated objectives there is still sufficient latitude so that "the nature of the final test depends largely on the idiosyncracies of the item writers" (p. 325).

Domain-referencing is the term used synonymously with criterion-referencing (Hambleton et al., 1978; Popham, 1978b) to signify greater clarity in the delineation of behaviors assessed by a test. The theory of domain-referenced testing represents a wedding of the detailed content analysis suggested by Skinner (1954) and generalizability theory (Cronbach, Rajaratnam, & Gleser, 1963; Cronbach, Gleser, Nanda, & Rajaratnam, 1972), whereby all the relevant conditions of observation are identified. Adequate definition of the domain requires specification of both stimulus and response dimensions (Cronbach, 1971; Hively, et al., 1968; Millman, 1974; Osburn, 1968). Popham proposed that test specifications have the following elements: general description, sample item, stimulus attributes, response attributes, and a specification supplement. The response component could either be statements that "delimit the classes of response options from which the student makes *selected responses* or explicate the standards by which an examinee's *constructed responses* will be judged" (p. 122).

Item-generating rules are alternatives to domain specifications. Various approaches have been tried to provide a more detailed blueprint than

Popham's specifications. Millman (1974) gave excellent summaries and examples of several of these including Popham's earlier amplified objectives (Popham, 1974, 1975), item transformations based on linguistic analysis (Anderson, 1972, Bormuth, 1970), item forms that provide algorithms for replacing item parts (Hively et al., 1973; Osburn, 1968), and mapping sentences based on Guttman's facet analysis (Guttman, 1969; Jordan, 1971; Tunks, 1973). Berk (1979) evaluated six strategies as to how well they provided unambiguous domain definition and satisfied eight practicality criteria. He concluded that item transformations, item forms, and algorithms offered the most rigorous and precise specifications, but also noted that each had been effectively applied in only one content domain (reading, mathematics, and attitude assessment, respectively). Amplified objectives, Popham's more recent domain specifications (1978b), and mapping sentences were judged to be the more practical methods.

The more elegant item-generating rules are not likely to be so useful for minimum competency tests as they are in well-defined academic subjects. Of course, once sample items have been written, item forms might serve the lesser purpose of producing parallel items to be used on alternate forms of the test. Elaborate item-writing rules are an improvement on sloppy "content categories"; but unfortunately they are not proof against the idiosyncracies of item writers. In the examples of domain specifications given by Hambleton and Eignor (1979), it is easy to see that the wisdom of item writing is now required to write domain specifications. Moreover, unless there is a larger fabric, so that many small homogeneous domains can be seen to fit together to exhaust the intended content area, it is possible that certain small domains will be assessed uniformly while other domains are missed. Because even the most sophisticated methods of domain definition do not guarantee success at the item-writing stage, review procedures are necessary to ensure that items are appropriate and that the larger domain is represented completely. These safeguards are considered in the next section.

Item Selection and Item Analysis

The content validity of a test depends on the correspondence between the intended content domain and the actual content of the test items. In the literature on domain-referenced testing, the traditional concept of content validity (APA, 1974) has been expanded to include the specification of response conditions as well as the substance of behavioral tests. Adequate domain specifications guarantee content validity by leaving no room for deviations at the item-writing stage. Ebel (1962), for example, stipulated that the requirements for items should be so clearly delineated that it would make no difference whether he or his secretary constructed the test. If the

domain definition provides an adequate blueprint, then items can be selected for a specific test by following a simple random or stratified-random sampling plan. These principles of test construction have also been advocated by Harris (1974a, p. 87), Harris, Pearlman, and Wilcox (1977), and Shoemaker (1975).

Although it is desirable to strive for these ideals in test construction, it is not possible to achieve them fully. Except for the domain of arithmetic computation, it is usually impossible to specify every detail of test content. In fact, the more abstract an instructional goal, the more difficult it is to create an algorithm for generating all relevant test items. Reading comprehension might be demonstrated, for example, by selecting one of several single-sentence summaries of a brief story; but it is difficult to convey the proper synthesis to be achieved in the correct sentence. It would even be hard to be consistent about how closely the correct answer should resemble a sentence in the story, or how to use vocabulary from the story. Linn (1979b) and Subkoviak and Baker (1977) noted that domain definitions are rarely explicit enough to make it clear what constitutes a representative sample from that domain. Linn was especially pessimistic about the likelihood of adequate definition and content validation for competency tests.

Popham (1978b) acknowledged that because of the complexity of the behaviors being assessed and the practical constraints that exist, it will be impossible to achieve complete clarity in test specifications. Recent efforts to provide guidelines for criterion-referenced measurement have been aimed at reducing the ambiguity in intended content. Recognizing that typical domain specifications will still leave much to the wisdom of the item writer, both Popham (1978b) and Hambleton and Eignor (1979) recommended using the additional insight provided by norm-referenced item rules developed from years of experience. For example, question stems should be free of grammatical cues as to the correct choice, and distractors should be constructed from plausible wrong answers (e.g., see Ebel 1972; Stanley & Hopkins, 1972).

After items have been written, it is advisable to obtain independent confirmation of the match between each item and the domain specifications. Hambleton and Eignor (1979, p. 47) and Hambleton and Fitzpatrick (1979) offered many practical suggestions for collecting such judgments. Not only should the substance and format be judged, but also content specialists should evaluate the comprehensiveness of the *set* of items. Cronbach (1971) proposed a duplication experiment as a more rigorous test of the adequacy of domain specifications and of the content validity of resulting tests. Independent teams would use the same rules and would construct parallel tests. Ideally, when empirical results are compared for the two tests they should differ only by sampling error.

The use of empirical analysis to evaluate the quality of criterion-

referenced test items is controversial. Probably the most emphatic opposition was stated by Harris et al. (1977): "the study of item responses—as in reliability estimation and item analysis—plays no role in the test development process" (p. 4). The concern, of course, is that selection of items or even revision by statistical criteria will distort representatives of the item pool as a sample from the intended domain. Such tamperings are believed to destroy the logical links between measurement and instruction. This idea is not new. In 1935, Buros directed the following criticism at item validation procedures:

> It is regrettable that the subject-matter specialists acquiesced so readily to the so-called "dictates of objective measurements." It seems inescapable that such methods of statistically validating achievement tests insidiously tend to strengthen the status quo, to impede curricular progress, to perpetuate our present grade classification, to differentiate rather than to measure, to conceal unlearning, and to give an illusory sense of continuous learning from grade to grade. (See Buros, 1978, p. 1,975).

It has been argued that because traditional item-analysis techniques depend on score variability they will systematically reject items with high proportions correct—a result to be expected from successful instruction—(Millman & Popham, 1974). Millman (1974) allowed only that item analysis could be used to detect flawed items: Once they have been scrutinized for logical flaws, however, they would have to be revised or replaced to maintain test scores as accurate estimates of domain scores. In their reviews of criterion-referenced testing technology, Hambleton et al. (1978) and Subkoviak and Baker (1977) advocated the use of empirical methods as long as the user is mindful of the potential for distorting the domain definition. Since a complete and explicit domain is usually not achieved in practice, item analysis can be used to identify instances where item writers have erred in their conceptualization. Hambleton and his colleagues pointed out that although variability should not be the guiding principle in test construction, it can be important in demonstrating validity. The test constructor can intentionally select samples that vary in competence (i.e., masters and nonmasters) and then make sure that item responses are associated with the known differences in groups. For example, Haladyna (1974) provided empirical evidence that classical analysis techniques are more interpretable and improve the construction of criterion-referenced tests if variability is increased by combining pre- and postinstructional groups.

The debate over using item analysis with criterion-referenced tests is easily won if the specific application is competency testing. In this case, one no longer has to choose between two purposes of criterion-referenced testing to estimate domain scores or make mastery classifications; the purpose of competency testing is to distinguish the competent from the incompetent. The test must have construct validity for this discrimination.

(In fact, as will be discussed in the next section on reliability, it must also be a dependable and efficient measure of this dichotomy.) When the meaning of a test hinges on the cutting-score, domain scores for individuals are of secondary importance. Rather than seeking to report an examinee's proportion-correct score with the smallest possible interval of error around each score, the goal is to minimize error at the cutting score. Obviously, test efficiency in making dichotomous classifications is gained at the expense of accurate instructional information along the full performance continuum. Subkoviak and Baker (1977) provided a readable discussion and simulated example of differences in item selection that will occur depending on whether the test purpose is to estimate the true percent correct or discriminate between better and poorer students (in this case along the full continuum instead of only two mastery states). Their example also illustrates how the estimates of domain performance will be much less accurate when a nonrandom sample of items is used. When items are selected to discriminate at a cutting-score, estimates of domain scores will be the most inaccurate for individuals the furthest from the cutoff.

Unquestionably, items must be selected for competency tests which discriminate most accurately between masters and nonmasters; these will be items that have a reasonably high correlation with criterion groups (presuming we can identify them) and that have maximal variance (difficulty of .5) for examinees whose performance level is exactly at the cutoff point. Different indices of discrimination appropriate for dichotomous classifications and comparisons among indices are found in Brennan (1972), Cox and Vargas (1972), Crehan (1974), and Haladyna and Roid (1976). Hambleton and Eignor (1979) suggested that in the future it may not always be necessary to sacrifice accurate estimation of domain scores to achieve valid discriminations. Latent trait methodology might eventually be used to estimate descriptive domain scores more accurately, despite the administration of only the more discriminating items.

Although the sample invariant properties of this methodology are an improvement over other statistical techniques, it is only applicable within reasonable limits and cannot be used to extrapolate to performance that is completely untested.

RELIABILITY

The literature on reliability for criterion-referenced tests is substantial and has been thoroughly reviewed by Hambleton et al. (1978), Linn (1979a), and Subkoviak and Baker (1977). Given the excellence of these references, the purpose of this section is to recapitulate briefly the issues that distinguish various approaches to reliability estimation. The discussion is shorter than that for other major methodological topics in the chapter because there is

less that is unique about this problem as it applies to minimum competency testing. Because minimum competency testing requires mastery classification rather than domain score estimation, the literature regarding statistics for the latter purpose can be ignored. To some extent, minimum competency testing creates unusual validity and standard setting problems which interact with reliability estimation. Implications of these features are given special note. Beyond those, minimum competency tests are similar to other mastery tests for reliability purposes.

In 1969, Popham and Husek argued that methods for estimating reliability based on classical test theory are inappropriate for criterion-referenced tests because they depend on score variability. Such methods were well suited for norm-referenced tests which are designed to provide a full range of scores and determine an individual's relative standing on a trait dimension. The correlational procedures used for estimating reliability directly reflect the kind of correspondence desired in the measures. For example, in test-retest reliability the correlation coefficient reflects the extent to which individuals maintain their same rank order (and interval distances) on the two test administrations. It does not indicate whether they each achieved the same score as before, since everyone could increase their score by 10 points and the correlation would still be perfect, if rank orders did not change. Woodson (1974a, 1974b) provided an early rebuttal to Popham and Husek's argument, pointing out that in order to measure, a test must discriminate. Although a criterion-referenced test may not produce variability if administered to a group of masters only, it must surely discriminate if both masters and nonmasters are tested, or it could hardly be valid.

Currently there is some concensus that Popham and Husek (1969) and Millman and Popham (1974) might have overstated the case against the use of correlational analysis to assess the reliability and validity of criterion-referenced tests. Their motive for doing so was probably to forestall the distortion of content coverage that would occur if these indices were mindlessly applied without notice of the effects of group variability and with increasing variance as a guiding rule in test construction. As has already been discussed, Hambleton et al. (1978) offered a solution to the lack of variance that Popham and Husek might have considered:

> In hindsight, they might have suggested that test developers "create" test score variance by "pooling" the test performance of two groups of examinees—those expected to be "masters" of the material included in a test (perhaps a group of examinees after instruction) and those who would be expected to be "nonmasters," perhaps a group of examinees prior to receiving instruction. It would then be possible to apply any of the classical reliability approaches and interpret the results in the usual way. (p. 15)

Linn (1979a) gave a more extended review of the score variability debate and also concluded that the problem can be avoided practically. He

emphasized the conceptually important point, however, "that variability should not be the guiding principle [in test construction] nor be allowed to distort [content] representatives" (p. 94).

Mastery Classifications Versus Domain Scores

Several approaches have been taken to solve the problem of judging test consistency in the presence of restricted score variability. Harris (1972), for example, noted that the standard error of measurement is an index of dependability unaffected by score range (or selection of a cut-score). Or, when the reliability question can be conceptualized as a sampling problem—how closely test results resemble those for the universe of interest—Cronbach's generalizability theory (Cronbach et al., 1972) is most appropriate.

However, when the test purpose is to make mastery classifications rather than to estimate domain scores for individuals, a different paradigm is appropriate. Hambleton and Novick (1973) introduced the idea of decision consistency rather than score consistency. If individuals are to be classified into mastery states, the important question is, how consistent are the classifications between two administrations of the same test or between parallel forms? They suggested the observed proportion of agreement as an index of reliability:

$$p_o = \sum_{k=1}^{m} p_{kk}$$

where p_{kk} is the proportion of examinees classified in corresponding mastery states on the two administrations. For minimum competency testing the number of mastery states, denoted m, will be two (or at most three).

Correction for Chance Agreement

Swaminathan, Hambleton, and Algina (1974) agreed with the consistency of mastery classifications as a definition of reliability, but criticized the simple proportion of agreement because it did not take into account the amount of agreement that would occur by chance. For example, if 100 percent of all examinees were declared masters on two separate testing occasions then the results would be perfectly consistent ($p_o = 1.0$) but little would be known about the dependability of the test. What is desired is an index that denotes the amount of agreement over and above the amount of chance consistency attributable to the marginal proportions. However, because chance agreement changes with the marginal values it is difficult to make these adjustments intuitively. Therefore, Swaminathan et al. (1974)

adopted Cohen's (1960) coefficient as an index of reliability adjusted for chance.

$$\kappa = \frac{P_o - P_c}{1 - P_c},$$

where P_o is again the observed proportion of agreement and P_c is the amount of agreement that would occur just by chance. The value for P_c is given by:

$$P_c = \sum_{k=1}^{m} P_{k\bullet} P_{\bullet k}.$$

The terms $P_{k\bullet}$ and $P_{\bullet k}$ are the proportions of examinees assigned to mastery state k on the first and second administrations.

Coefficient κ can be interpreted as the proportion of consistent classifications contributed by the test, beyond the chance agreement attributable to the particular proportions of masters and nonmasters on the respective testings. The properties of κ have been well described (Cohen, 1960, 1968; Fleiss, Cohen, & Everitt, 1969; Hubert, 1977). An excellent discussion of considerations in choosing between p_o and κ is given by Subkoviak (1980).

As has always been our experience in dealing with measures of association, it depends on the particular application which index is more appropriate. Just as the various measures of agreement proposed by Goodman and Kruskal (1954) allow one to adjust for various factors which may influence interpretation, κ provides a more direct measure of test quality that is not confounded by the selection of a particular cut-off score. It might be especially useful during test development when estimates of reliability are desired but selection of the final cut-off score is also being deliberated. The uncorrected coefficient p_o, however, is a better indicator of the consistency of decisions as they actually occur, given the chosen cut-off score and the heterogeneity of the population. It is a composite index which summarizes decision consistency attributable not only to the set of items and test length but also to other conditions of the decision situation including the cut-off score.

Hambleton et al. (1978) noted that the magnitude of κ is dependent on the cut-off score and the heterogeneity of the group of examinees and attributed to Millman the admonition that all of these be reported along with κ to ensure accurate interpretation. It might also be well to repeat the old psychometric advice that the characteristics of any field test sample must be like that of the population for whom the test is intended, especially both the mean and variance.

Single Test Administration

As originally proposed, computation of κ or p_o requires two test administrations. Recently several methods have been developed for estimating these indices from a single test administration. Huynh (1976a) derived an estimate of κ; Marshall and Haertel (1975) and Subkoviak (1976) provided an estimate of the unadjusted proportion of agreement (p_o). Of course the latter can also be adjusted for chance agreement (see Subkoviak, 1980).

Single testing estimation procedures are more desirable for classroom applications of criterion-referenced tests, where it may not be feasible to conduct test-retest studies. Minimum competency testing programs, however, are likely to be implemented on a large scale at either the state or district level where sufficient resources are available for substantial field trials. Therefore, the more convenient procedures may not always be the method of choice. Subkoviak (1978) empirically compared the three single testing procedures and the Swaminathan approach (coefficient κ). The Swaminathan procedure produced unbiased estimates but had relatively large standard errors for classroom size samples ($n = 30$). Again, it should be noted that minimum competency tests are likely to be developed in situations where larger reliability studies would be warranted. Subkoviak concluded by recommending the Huynh approach, from among the single test procedures, because it "produces reasonably accurate estimates, which appear to be slightly conservative for short tests" (p. 115).

Choice of Loss Function

Hambleton and Novick's (1973) decision-consistency approach was not the only solution proposed for tailoring reliability indices to fit the uses of criterion-referenced tests. Other methods were introduced which quantify the amount of mastery or nonmastery rather than the number of misclassifications. Livingston (1972) adapted classical test theory for criterion-referenced testing by reconceptualizing variance not as deviation of scores about the mean but as deviations about the criterion score. His criterion-referenced reliability coefficient is defined as

$$K^2_{XT} = \frac{\sigma_T^2 + (\mu_X - C_X)^2}{\sigma_X^2 + (\mu_X - C_X)^2}.$$

where T and X, respectively, are used to denote true and observed scores and C is the cut-off score. Obviously, the Livingston coefficient will increase as a function of norm-referenced reliability and the distance of the criterion score from the mean. Harris (1972) criticized the latter feature because the increase in reliability caused by selecting a cutting score far from the mean

was not a reflection of the dependability of the test per se. Of course, the question as to whether the location of the cut-off score should be allowed to influence the interpretation of reliability is much like the debate over whether p_o or κ is the better index.

Hambleton and Novick (1973) criticized Livingston's approach on much more fundamental grounds, arguing that when the purpose of testing is to determine mastery, one is not interested in the degree of mastery or nonmastery status but only in whether misclassification occurs. They emphasized that the inappropriate choice of loss function (square-error loss rather than threshold loss) was a much more serious deficiency of the classical approach than its reliance on variance. Actually there are differences of opinion as to whether it is undesirable to reflect *degree* of misclassification. It can be argued that in many criterion-referenced applications, especially competency testing, selection of a cut-score, however well reasoned, results in an artificial dichotomizing of a continuous trait. Therefore, small deviations near the cut-score reflect much more ambiguity as to actual competence; and one would, indeed, like to count (i.e., weight) more extreme deviations as more serious instances of unreliability. The dilemma one faces in selecting the appropriate conceptualization of error (and corresponding reliability index) is summarized by Brennan and Kane (1977), "a squared-error loss function has the advantage of being sensitive to the magnitude of errors, but the disadvantage of being sensitive to all errors of measurement including those that do not lead to misclassification" (p. 287). A more dramatic way of stating the drawback to squared-error approaches is found in a quotation attributable to Kenneth Boulding: "The trouble with least-squares estimation techniques is that in the treatment of error there is no distinction between whether one has stopped five feet short of the cliff or gone five feet beyond."

Unfortunately, an ideal index, which would reflect only misclassification errors but would weight these by the degree of error, is not available. In the interim it is advisable to use a combination of approaches to summarize both decision-consistency and magnitude of errors. For the latter purpose, it is preferable to use the standard error of measurement (as discussed in Hambleton et al., 1978 or Harris, 1972) or classical coefficients with adequate variance assured. If a squared-error index were desired, the Brennan and Kane's (1977) approach discussed in the next section is an improvement on Livingston's (1972) coefficient.

Generalizability Theory

Brennan and Kane (1977) concurred with Livingston's conceptualization of error as squared deviations in person scores across testing occasions;

however, they also improved on the classical approach by applying Cronbach's et al. (1972) generalizability theory. The benefit is that error due to random sampling of items from the domain is also taken into account. Their index of dependability, defined using expected squared deviations from C is:

$$M(C) = \frac{\epsilon_p\,(\mu_p - C)}{\epsilon_i\epsilon_p\,(X_{pi} - C)},$$

where ϵ_p denotes the expected value over all persons in the population of persons and ϵ_i indicates the expected value over all samples of items of the size used in the test. Just as in Livingston's formula, dependability is a function of the distance between the mean domain score and the criterion score ($\mu_p - C$). In fact, the authors advisedly used the symbol $M(C)$ to show its dependency on C. Observed differences are indicated by $X_{pi} - C$, where X_{pi} is the observed score for person p averaged over the sample of items.

Linn (1979a) noted that one of the unfortunate consequences of eschewing approaches based on score variability is that the usefulness of generalizability for criterion-referenced measures has been missed. It would also be unfortunate if generalizability theory were ignored because of the disadvantages associated with measuring error as deviations from the cut-score. Not only does the sampling paradigm of generalizability theory match the domain specification requirements of criterion-referenced testing, but as Cronbach stated, "Estimates of (variance) components are particularly illuminating when an instrument is used for absolute measurement" (1975, p. 602). Brennan's (1979) more recent paper provides an extensive and comprehensive discussion of applications of generalizability theory to the dependability of domain-referenced tests. The use of variance components to identify the sources of error and lack of domain representativeness is more important than the attempt to quantify reliability using deviations from the cut-score.

STANDARD SETTING

The spirit of the minimum competency testing movement is to reaffirm standards. Both political and practical requirements imply that standards must be set to decide who has passed the test. But the problem of how to set justifiable cutoff scores does not have a ready solution. This together with the question of validating the substance of the test are the two largest obstacles to testing for minimal competence. Interestingly, they are also the two methodological areas where minimum competency testing most diverges from classroom applications of criterion-referenced testing. For tests of academic subject matter, content validation is more straightforward and may be satisfied by detailed domain specifications; testing for

competency, especially as a prerequisite for life success relies on much greater inferences which are difficult to validate. Selecting cutoff scores is a related but distinct problem. One must determine how much knowledge is sufficient. In the classroom, one may rely on traditional views of what constitutes mastery (Bloom, 1968), or select by trial and error the performance level that seems to ensure success in later instruction. In any case, the most important feature of classroom testing for mastery is that errors in test decisions can be easily redressed. For large-scale minimum competency testing programs, it is both more difficult to select a particular cutting-score and more serious to make a mistake.

The reviews on criterion-referenced testing methodology previously cited (Hambleton et al., 1978; Subkoviak & Baker, 1977) contain sections on setting passing standards because many applications require mastery classifications as well as domain score estimation. In addition, major reviews which focused entirely on the standard-setting issue, have been written by Millman (1973), Meskauskas (1976), and Glass (1978b).

To organize the literature on standard setting, Meskauskas (1976) identified a fundamental distinction between state and continuum models. These approaches differ by the way in which authors conceptualize the underlying trait being assessed. In state models it is assumed that mastery is all-or-none. In these circumstances it is not difficult to set a performance standard, since it must be 100 percent by definition, with perhaps some allowance for measurement error. Standard-setting methods based on this view of mastery include those of Emrick (1971), Macready and Dayton (1977), and Roudabush's dichotomous true-score model (1974). State models are credible for certain physical performance tasks and for very discrete cognitive tasks (such as adding and subtracting, overlooking occasional errors in balancing a checkbook). However, for most cognitive skills taught in school or influenced by school learning, state models are inappropriate because the skill is continuous, acquired in undiscernible bits rather than by crossing a threshold. Even in reading, only a state of complete illiteracy is identifiable (i.e., not even letter recognition). The minimum acceptable level of reading is on a continuum somewhere between reading only "stop" signs and comprehending the *Wall Street Journal*.

For the skills assessed in minimum competency testing, only continuum models are applicable. This means that a continuously distributed trait must be artificially dichotomized to allow classification decisions. It is this problem—of drawing a pass-fail line when no clear distinction exists between the competent and incompetent—which leads to the arbitrariness of standard-setting methodology.

Every author agrees that all standard-setting techniques are judgmental (Shepard, 1976, 1979b; Glass, 1978b; Hambleton & Eignor, 1979; Jaeger, 1976, 1979; Popham, 1978a, 1978c). They all involve subjective choices

about what constitutes mastery. Glass (1978b) evaluated six different types of standard-setting methods and concluded that none was adequate, essentially because the distinction they seek to approximate does not exist. Glass would not say that very good performance cannot be distinguished from very bad performance, only that there is no point where nonmastery changes to mastery. When trying to measure competence it is reasonable to conceive of the trait, but it is not logical to theorize that absolute distinctions exist between masters and nonmasters when we have already acknowledged that the trait is continuous. Therefore, it is not possible that either present methods or improved methods, developed after redoubled research efforts, will be able to uncover "true" standards. Because Glass' review was so pessimistic, and because the alternatives he proposed always required a basis of comparison (e.g., change scores) or some marketplace determination of quotas, a great deal of energy has been devoted to refuting his arguments (Block, 1978; Hambleton, 1978; Hambleton & Eignor, 1979; Popham, 1978a; Scriven, 1978). The chief rebuttal has been that standard setting may be arbitrary but it is not capricious (Popham, 1978a).

The APA test standards (1974) require that a test user have a "rationale, justification, or explanation of the cutting scores adopted" (p. 66). Popham argued that it is possible for school boards to satisfy this requirement by deliberating and selecting reasonable proficiency levels "which they consider acceptable" (p. 298). The picture painted is one of much more considered judgment and care than the mindlessness and willy-nilly standard setting that Glass criticized. Although it is true that standard setting can be done more wisely than has been typical, the point should not be lost that the arbitrary cutting points still result in artificial dichotomies. Not only will there inevitably be misclassifications due to unreliability, but even for those "correctly" classified there will be very little difference between individuals adjacent on either side to the cutoff score.

Given that standard setting involves an artificial demarcation on a performance continuum, it will be difficult to argue that the line should be drawn one place instead of another. Human judgments are fallible. In the case of minimum competency testing programs, the consequences of wrong judgments can be serious as evidenced by the recent court case in Florida, *Debra P.* vs. *Turlington.*[2] Popham (1978a, 1978c) suggested that such programs are no more arbitrary than conventional grading practices; moreover, they can be based on more careful measurement and collective judgments. Although it is true that centralized competency programs can bring the best technology to bear on the problem, it is also true that much more will hinge on a single score than hinges on a single grade. Previously, a student graduated on the basis of passing grades in numerous courses; an unfair F or a generous D was likely to be balanced out over the course of a student's career. If instead, graduation depends on a single test, errors will

be greater and more difficult to balance for the individual. Of course, multiple opportunities to take the test protect against measurement errors; and remedial course work should actually change a student's proficiency level, thereby offering some benefit if standards are accidentally set too high. It is presumed, however, that cutoffs cannot be set too easy or the entire exercise will be meaningless; and if instead the error is made in the other direction, there will be enormous costs to both the individual and the system, even if most students eventually pass the test. This discussion merely serves to emphasize that there are reasonably high stakes attached to the selection of cutoff scores.

The following sections contain short summaries of different standard-setting methods. Each is considered for its possible contribution to the selection of reasonable cutting scores. Deficiencies in each method are also especially stressed because of the harm that can result when apparently scientific methods are used to make crucial decisions without recognizing the subjectivity involved. In the end, a combination of approaches is recommended as the best protection against the shortcomings of individual methods.

Standards Based on Judgments of Test Content

Standard-setting methods that require judges to review test content before choosing a passing score are an obvious improvement on traditional standards which remain fixed at 70 percent (or 85 percent in many objectives-based programs). Historically, when 70 percent was universally accepted as a passing grade, one can imagine that teachers adjusted the difficulty of test content so that the prespecified standard would correspond to an acceptable performance level (see Airasian, Kellaghan, Madaus, & Ryan, 1972). One can also surmise that this juggling may not have been very conscious, and that sometimes serious misjudgments would be made so that too many would pass or fail the test. Later, grading practices became more variable and relied more on normative information, so that 50 percent might be an A on a Calculus test if it were the best paper in the class. Norms, of course, have their own problems such as failing to recognize when all students in a given class do very well or very poorly compared to other classes. The return to absolute standards brings the history of grading practices full circle. However, the introduction of careful domain specifications precludes the "juggling" of test content necessary to maintain 70 percent as a reasonable standard. Public acceptance of passing scores around 70 or 75 percent because of precedents in the schools 20 years ago is misguided; these cutoffs could be exceeded often or only rarely depending on the content of the test.

Because standards do not exist in nature waiting to be discovered by our

methods and because an absolute rather than normative cut off is desired, experts will have to decide on standards by reviewing test content. When criterion-referenced tests are used to decide placements in classrooms, the judges should be teachers and subject-matter specialists. However, the political nature of minimum competency testing programs may require that additional relevant stake holders (i.e., parents, employers, and members of minority groups) be involved in the standard setting process. The Nedelsky (1954) method for judging test content has the longest history and has been used extensively on certification exams in medicine (Bobula & Standish, 1974; Meskauskas, 1976; Meskauskas & Webster, 1975; Paiva & Vu, 1979). Judges are instructed to consider each multiple-choice test item and decide how many options are so obviously wrong that a minimally competent student will be able to reject them. (In other words, the student is not expected to be able to make the more esoteric distinctions between the correct answer and the next best answer.) The minimum passing level for each item is determined by computing the chance score for the remaining alternatives. For example, on a 5-option question where two answers are clearly wrong, the minimally competent student will guess from among the remaining choices and has a one-third probability of being correct. Each judge's minimum passing score is obtained by summing these item values; the results from all the judges are then averaged.

Apparently Nedelsky also wished to take into account that not every minimally competent student would earn exactly the average chance score. Therefore, to apply the formula one must also decide what percentage of exactly borderline students should pass; the percent is converted to a constant on the normal distribution axis and multiplied times the standard deviation of the individual judge's passing scores; the average standard is then adjusted upwards or downwards by this amount. One has to question why he did not use the standard deviation due to chance. As Meskauskas (1976) noted, the use of a term reflecting variability in judges' standards seems unjustified, "although the basic concept of fine-tuning the setting of a minimum-passing point on the basis of probability value may well have utility" (p. 136).

The Nedelsky approach has the advantage of recognizing the effects of chance success, but it presents judges with an artificial and often puzzling task. For example, Paiva and Vu (1979) noticed that judges had trouble disassociating their judgments from their own difficulty in answering the questions. This is disturbing, since their scoring of the items should not depend on the difficulty in choosing between the correct answer and the next-best answer; rather the Nedelsky method focuses judges' attentions only on eliminating the easy distractors rather than making the tough discriminations. Also, Meskauskas and Webster (1975) reported huge

differences in individual judge's standards which are not resolved by the Nedelsky method.

The Angoff (1971) method offers judges a slightly easier task than the Nedelsky procedure. They have to consider each item and estimate the probability that a minimally competent individual will get it right. These probabilities are then summed and averaged across judges. Step-by-step directions for this and other standard-setting methods are given in a manual by Zieky and Livingston (1977).

When instructing judges for the *National Teacher Examination*, Educational Testing Service (1976) improved on the Angoff method by presenting the judges with a probability scale, including a "do not know" position. Although it is hardly different from asking them to pull a probability from thin air, it does call to mind that one might improve judgments by using several techniques developed for the subjective scoring of essay tests (Coffman, 1971; Stanley & Hopkins, 1972). For example, before assigning probability values a judge should attempt to sort items into ranked categories from hardest to easiest. Then each set of items can be double-checked to make sure the items are homogeneous in difficulty (should not be sorted into additional stacks) and are distinguishable from the next higher and lower categories. After the judge is satisfied with the sorting of items, he or she can assign uniform probabilities to each stack.

Ebel (1972) proposed a slightly more complex procedure. Judges are asked to categorize items by both difficulty and relevance. Relevance refers to how central an item is to the competence being assessed. Probabilities of getting the item correct are then assigned to each category and used to weight the items in computing the minimum passing score. By this method some allowance is made for missing relatively easy but less relevant items. However, judges may find it difficult to keep the dimensions distinct, since familiarity (and hence relevance) is usually a factor in item difficulty.

One of the few empirical studies comparing standard setting methods was done by Andrew and Hecht (1976) using the Nedelsky and Ebel methods. A panel of eight judges met on two separate occasions to set standards using those two methods; care was taken to counterbalance for order effects. The percentage of items expected to be answered correctly by the minimally competent student was 68 percent for the Ebel method and 49 percent by the Nedelsky method. Glass (1978b) found the greater than 20 percentage-point difference disconcertingly large. By making some assumptions about the test distribution, he further showed that the discrepancy would result in an even greater difference in passing rates: 95 percent of the students would pass the test if the Nedelsky cutoff were used and only 50 percent would pass using the Ebel criterion. Glass took these results as evidence of the inadequacies of the methods.

Hambleton (1978), however, averred that the Andrew and Hecht findings are not damning to the standard-setting methods because the "directions to the judges were different, and procedures differed, [therefore] no one should expect the results from these two methods to be similar" (p. 283). Andrew and Hecht also attributed the differences to "different philosophical assumptions and varying conceptualizations" (p. 49). This defense is inadequate, however, because it presumes that one knows well enough from the model descriptions what the implicit differences will be in the definition of the minimally competent individual. In measurement one is always content when measures of different things yield different results; but if two instruments are intended to measure the same thing and disagree widely, the conclusion is that one or both are seriously in error. Little work has been done to state a priori how differences in procedures affect the definition of minimal competence. These distinctions would have to be drawn before one could select different methods for different purposes.

It is plausible that among the Nedelsky, Ebel, and Angoff methods the Nedelsky criterion will be lowest, because the task of eliminating the clearly wrong answers is easier than actually choosing the correct answer. Brennan and Lockwood (1980) used generalizability theory (in a study involving five judges and 126 items) to quantify the variability in cutting scores attributable to differences in the Angoff and Nedelsky methods. The Nedelsky procedure resulted in lower cutoffs and greater variability in judges ratings. The variance components estimated from a mixed-effects analysis of variance were four times greater for differences in procedures than for differences in rater means (over procedures). Brennan and Lockwood did more than other authors to analyze how differences in procedures lead to differences in results. The most important conceptual difference, of course, is whether probabilities are estimated by eliminating distractors. There are also some artifactual differences caused because the Nedelsky procedure restricts the rater to a small discrete number of unequally spaced probabilities (because of the number of answer choices). When trying to select among methods, the Nedelsky approach should not be used unless elimination of wrong answers is clearly consistent with how a minimally competent individual would be expected to answer the test. In medicine, for example, imagine a test question that asks which drug should be prescribed for a particular disease. Two of the choices are obscure drugs which would have no effect, one choice is a drug that will kill a patient with that condition, one answer is the correct drug, and the last answer is the traditionally correct drug which has undesirable side-effects. If judges believe that the minimally competent examinee should only be able to eliminate the fatal drug, the Nedelsky procedure will lead to a different and lower standard than if they think it requires minimal knowledge to distinguish between the two effective drugs, one with bad side-effects. In most minimum competency testing

situations test specifications and items have been written so that the hardest discrimination (choosing between the right answer and the next best answer) reflects the minimum competence to be assessed. If judges believe that for many questions the minimally competent individual should be able to eliminate all of the distractors, the Nedelsky approach will be unwieldy and will not result in any better quantification than if the judges simply guess about the probability of correctly distinguishing between the two hardest choices.

Using a reasonably large number of judges is the only way to ensure the political credibility of the standards (see Jaeger, 1978, regarding representative sampling procedures). For various judgmental methods, however, the question remains as to how to combine individually set standards. Simple averages create a false concensus and do not acknowledge differences in the wisdom of various judges. Brennan and Lockwood (1980) suggested a "reconciliation" procedure, where judges are asked to meet and discuss their differences to arrive at a final standard. It might even be wise to consider instances of extreme differences of opinion as evidence for questioning the validity of a particular item. Certainly it is important that agreement on averaged standards across groups of judges not be interpreted as compelling evidence of having uncovered the "true" standard. If panels of judges have been constituted randomly and standards arrived at by averaging, then of course the means will differ only by sampling error; but profound differences of opinion within the panels will not have been resolved (cf. Bernknopf, Curry, & Bashaw, 1979).

A judgmental method proposed by Jaeger (1978) will be referred to again in a later section because it involves the use of normative data as well as inspection of test items. Jaeger acknowledged the dilemma posed by different values. He recommended that median standards be derived for each type of judge. (Medians will be less influenced by extreme individuals than will means.) The lowest standard set by any group will then be adopted as the final standard.

Standards Based on Judgments about Groups

Judging test content may prove to be a contrived task, since judges will have to strive constantly to keep in mind what they expect a minimally competent student to be like. Although judges can reasonably conclude that a minimally competent student cannot answer questions that they cannot answer, they will usually have to make allowances for large differences between themselves and the marginal student.[3] Zieky and Livingston (1977) suggested that it may be easier to judge the performance of real students rather than to judge test items. Hambleton, Powell, and Eignor (1979) called methods based on this approach "combination models" because they

involve judgments about groups followed by empirical work to select the cut-score.

The Contrasting-Groups method (Zieky & Livingston, 1977) requires that judges know in advance a great deal about the proficiency of the students they will judge. This will usually mean that classroom teachers must be the judges: If parents or other interested parties were to become involved they would have to devote considerable time to learning each student's capabilities either by interviewing or individual testing. The judges then identify students they know to be masters and those they know as nonmasters (most applications call for simply excluding cases that are not easily classified). The cutoff score on the competence exam is determined by giving the test to both the masters and the nonmasters and choosing the score that best separates the two groups. This will usually be the point of intersection between the two distributions (assuming equal samples or standardization to constant area); however, the point can be adjusted up or down if the two kinds of error (false masters and false nonmasters) are not considered equally serious. Because of legal ramifications in minimum competency testing, masters who are misclassified will probably be considered the more serious, especially if the error is due to lack of validity rather than instability in the scores. Figure 1 is an illustration of overlapping mastery and nonmastery distributions and two different passing scores set either to balance the two types of error or to reduce false negative errors.

The Contrasting-Groups method is deceptively simple. Although no empirical studies exist, it is obvious that differences in judges' conceptualizations of mastery will affect the cut-score. For example, if judges have a tendency to classify borderline students as masters, the cutoff score will be lower than if a preponderance of borderline cases are classified as nonmasters. This shifting will occur even if the instructions call for excluding borderline cases, since there is no accurate way to define "how borderline is borderline." These procedures will nevertheless be essential as part of the test validation effort. It will be useful to see how much the two a priori groups overlap. Although the point of intersection could hardly be defended as the automatic standard, this information will be an important ingredient in the rationale for the final standard; this is essentially the "classification" problem of discriminant function analysis (Kendall & Stuart, 1966).

The Borderline-Group method (Zieky & Livingston, 1977) is the complement of the Contrasting-Groups method. The same knowledgeable judges are asked to identify students they consider to be borderline masters of the subject matter being assessed. The test is then administered to these students and their median score is used as the standard. Hambleton et al. (1979) considered this to be a conceptually more difficult task than identifying sure masters and nonmasters and therefore prefer the Contrasting-Groups method. Also, as Zieky and Livingston noted, it is

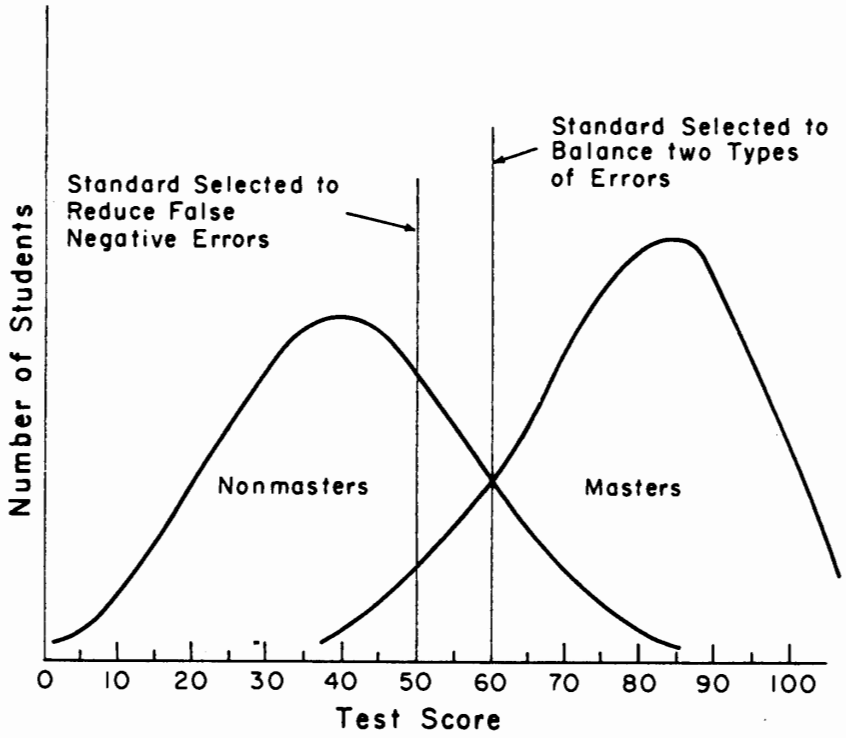

Figure 1. Mastery and nonmastery distributions with hypothetical standards.

difficult to obtain the recommended minimum 100 students for the empirical study, since the borderline cases are such a small percentage of the total group to be tested. Because the Borderline-Group method does not have any obvious advantages over the Contrasting-Groups method, its use is not recommended.

Empirical Methods for Discovering Standards

Every reviewer of standard-setting methods has used slightly different categorizations to describe the various approaches. There is the least consistency in the labels and subdivision for methods using empirical data. Perhaps this is because most approaches have more than one salient characteristic; for example, they may follow both a decision-theoretic conceptualization and use Bayesian statistical methods or they may require both an external criterion and a utility function. The methods discussed in the previous section involved both judgments and empirical data, but the judgmental step was considered the more crucial aspect. Two major

categories have been adopted here to typify the remaining empirical methods. The first set includes those methods which assume that a standard exists in the performance of the population. Statistical methods are applied to uncover the inherent standard. Methods assigned to the second category are actually not for setting a standard; rather a standard has already been chosen by some other method and the purpose is to fine tune or adjust the test standard.

Berk (1976) proposed an empirical method nearly identical to the Contrasting-Groups method. However, he sought to eliminate the problem of defining (and judging) mastery and nonmastery by selecting two criterion groups: one instructed and the other uninstructed. Some subjectivity is required, since the instructed group must have received "effective" instruction before one can presume that they are masters. The cutting score is selected that maximizes the agreement between the test classifications and criterion groups.

Berk's original intent was to use this approach with short criterion-referenced tests in instructional settings. Hambleton and Eignor (1979) concluded that the method is promising for this purpose. However, even in situations for which the method is most appropriate, there will not be a "true" standard to be discovered. The optimal cutting score identified will depend on the degree of nonmastery in the uninstructed group and both the duration and effectiveness of the instruction received by the group expected to be masters. For minimum competency testing the method is not applicable at all, because it is impossible to identify instructed and uninstructed groups for the competencies tested, since the skills are presumably acquired during 12 years of schooling. Moreover, the presumption that an instructed group will be predominantly masters is hardly valid; obviously if instruction guaranteed mastery the need for minimum competency testing programs would not have arisen in the first place. Hambleton and Eignor were equally pessimistic: "Other extreme groups might be formed (for example, "successful" adults and "unsuccessful" adults) and their performances compared on the test for the purpose of setting an optimum standard. Clearly though, results from such comparisons can be explained in numerous ways and, therefore, results of this sort have limited practical value" (p. 385).

Block (1972) developed a standard-setting model called "educational consequences," named for its attempt to maximize subsequent learning or some other valued outcome. This scheme is slightly different from those discussed in the next section which assume the existence of an external criterion of success (i.e., a point that separates successful from unsuccessful); the educational consequences approach only assumes a functional relationship between the present test and the later behavior it is expected to influence. (Of course, unless the test content is an end in itself, this

relationship is necessary for validity.) It does not assume that cutoff or dichotomy exists on the criterion dimension. The cutting score on the test, selected after experimental studies, is the one that maximizes performance on the outcome dimension. (Multiple desired outcomes can be "maximized" by forming a composite or by adjusting the standards set from separate analyses.)

Glass (1978b) severely criticized this method, which was intended to avoid arbitrary judgments by discovery of a more scientific standard. Unless the relationship between the test and the valued outcome is nonmonotonic (i.e., increases and then decreases), a 100 percent test standard will be optimal. This is obviously unrealistic and hardly worth the extensive investment in field trials to determine. And as Glass suspects, monotonic curves (with no intermediate maximum point) will almost always occur whenever both the test and outcome are cognitive variables. If perchance an attitudinal outcome could be found with a monotonic decreasing relationship to the test, it could be combined with the cognitive outcome to produce the desired nonmonotonic function. However, there is no science or methodology for deciding on the appropriate weights to use in forming such a composite. Such an approach may help in quantifying the effects of different judgments, but it will not avoid the problem of making judgments.

Block and Berk originally thought of their methods in the context of more circumscribed instructional settings. Block's method may still have application in situations where instructional units are in a demonstrated hierarchy and where plateaus (if not decrements) can be identified in the developmental graphs. The greater obstacle to solving the standard-setting problem in minimum competency testing with this approach is stated by Hambleton and Eignor (1979):

> There is yet another problem, perhaps even more serious than the non-monotonicity problem. One can't maximize a valued outcome if the outcome can't be defined in any reasonable manner. In sum, to utilize Block's method, there would have to be concensual agreement on what a valued outcome of being competent is. This would seem to be as difficult a task as trying to get people to define behaviors associated with minimum competency. (p. 385)

Their comments also bode ill for the solution of validity issues raised earlier. Finding a valued outcome or external criterion of success for standard-setting purposes is the same problem as trying to define success so as to validate the test's predictive validity.

The Use of Norms in Setting Standards

Hambleton et al. (1979) identified three different ways that norms can be used to establish cutoff scores. The first method is the same as the use of successful criterion performance to distinguish masters from nonmasters on

the test in the Contrasting-Groups method, Block's (1972) Educational Consequences Model, and Huynh's (1976b) external criterion method. These methods fail for lack of a valid and defensible criterion (see especially, Burton, 1978; Jaeger, 1978).

The second normative approach that Hambleton et al. find objectionable is simply to take a percentile rank in the norm group as the standard. This occurs, for example, when the minimal competence standard is set at the eighth-grade level without regard for the test content. They disapproved especially of the passing score for the California High School Proficiency Exam set at the 50th percentile of graduating seniors. In this case, however, such a decision may not be as unjustifiable as it sounds. Having made some effort to construct a test of real-life skills for which there was no previous data, both the fairness and the validity of the test would depend on how average high school students would score. Given the political purpose of the test, to allow "qualified" students to leave high school early, it would have been hard to defend a substantively determined standard that was very much above or below the median. Judges can decide about the credibility of a norm value as reasonably as they can choose an acceptable passing score. (Of course, normatively set standards must be based on a representative sample of the population of interest so that the standard will not fluctuate with the particular group of individuals taking the test.)

A third use of norms commended by Hambleton et al. is to supplement the judgmental process of determining the cutoff point. This use of typical performance data is recommended by Conaway (1976, 1979), Jaeger (1978), and Shepard (1976, 1979b). The use of normative data does not have to lead to such nonsensical goals as "everyone will be above the national average." Data about the actual performance of a representative sample can simply make judges better informed. Judges' opinions about desirable performance levels are based on "informal" norms from their own experience. Often extreme differences in judges' opinions about criterion levels are caused not by intentional differences in stringency but by unrealistic expectations gained from experience with unusual populations.

The argument for the importance of normative data can be pressed more strongly by examining the original taboos against norms. Norms were eschewed in criterion-referenced testing and competency testing because the decisions to be made were quota-free. That is, the testing situation did not require that a prespecified number of examinees pass or fail (Cronbach & Gleser, 1965). When quotas exist, it is straightforward and nonarbitrary to set a cutoff score. Candidates are ranked by test scores and then, counting down from the top, as many are taken as are needed to fill the quota. If there is no quota, however, an absolute judgment must be made.

The distinction between quota-free and fixed-quota testing is not as clear as one might suppose. Prestigious private colleges, always operating with a

quota, might occasionally admit more freshmen because judgmental discriminations at the cutting point were too difficult to make; or they might admit a smaller number one year because the quality of applicants fell below some absolute standard of excellence. Similarly, so-called quota-free situations are not really free of limits and relative judgments. Millman (1973), for example, suggested that the financial cost of providing remedial instruction could be taken into account in adjusting the standard. In the case of minimum competency testing, implicit quotas exist. Given the fallibility of standard-setting methods, one would immediately call a standard invalid if it failed 90 percent of the senior class in an average or good high school. The test would be invalid as a measure of minimums if it failed 50 percent or even 25 percent. At the other extreme, since there are known incompetents in the schools (*Peter Doe* vs. *San Francisco Unified School District*[4]), a standard would be automatically invalid if it failed none of the seniors taking the test (assuming a large representative sample). Judges convened to set standards would be advised to confront directly the question of where, in the range from 5 to 20 percent, the failure rate should be. The legitimacy of the standards will be ensured just as much by examining the failure rates as by inspecting the test content. Both types of judgments are necessary.

Empirical Methods for Adjusting Standards

The following methods conform to a decision-theory approach, even when the authors do not explicitly so name it. The nature of the decisions to be made, in this case dichotomous classifications of masters and nonmasters, is taken into account in formulation of the measurement problem. The best reference which sets criterion-referenced testing in this context is by Hambleton and Novick (1973). Due credit is given by Hambleton and Novick for adaptations and terminology taken from earlier work in personnel selection theory by Cronbach and Gleser (1965). The methods are not, strictly speaking, aimed at creating standards. If a standard already exists, either for the test or indirectly for a criterion measure, then these methods can be used to adjust the given standard to minimize error.

The decision situation is represented in Table I. The problem of correctly locating the passing score on the test, given a standard on an underlying performance continuum, is nearly identical to the issue of decision validity for competency tests. Does the discrimination made by the test accurately reflect the latent dichotomy? Some models treat all errors as equally serious; others seek to reduce either false-positive or false-negative decisions on the basis of assigned utilities.

Millman (1973) summarized techniques for moving a cutoff score up or down depending on the financial and psychological costs associated with the two kinds of misclassification. A great deal of work has been done since that

TABLE 1
Latent Performance Continuum

		Incompetent $\pi < \pi_o$	Competent $\pi \geq \pi_o$
Test Decision	Fail $X < C$	Correct Nonmasters	False Negatives
	Pass $X \geq C$	False Positives	Correct Masters

Note. π_o = cutting score on the latent variable;
C = cutting score on the test;
π = the individual's domain score;
X = the individual's test score.

time to provide models for incorporating benefit and cost data. It is highly advisable that already subjectively determined standards still be made more generous or more conservative to protect against the more serious kind of error. In minimum competency testing programs false negatives, students who fail the test unfairly, are more serious than lucky incompetents. Although it is likely that judges will wish to lower standards slightly to prevent these mistakes, they cannot be set so low that everyone passes or the entire exercise would be meaningless. Little advice has been offered to help the decision maker decide what weights to assign to the different costs. In a particular situation there may be general agreement about which loss is greater, but the selection of a 2-to-1 or 3-to-1 loss ratio will be arbitrary and will make a difference in passing rates. Perhaps the best advice for the administrator who is called upon to provide these insights, is to stimulate the effects (on passing rates) of different utility values before agreeing to the final loss ratio.

Huynh's (1976b) empirical Bayes approach is one of the better known procedures for setting cutoff scores given the existence of an external criterion. Huynh was initially interested in situations where criterion-referenced tests would be used to determine when a student's mastery of a topic was sufficient to allow him or her to progress to the next topic. Success on the next unit of instruction, called the referral task, is used as the criterion. Given π_o, the cutoff score for success on the criterion task, C for the test is chosen so that in the usual fourfold table, Table I, the average loss, $P(\pi < \pi_o, X \geq C) + P(\pi \geq \pi_o, X < C)$, is the smallest. Glass (1978b) called this approach "bootstrapping on the other criterion scores." Its most serious

deficiency is that it begs the question of how to justify the standard on the criterion measure. And for minimum competency testing purposes neither a criterion nor a standard of success exists. A procedure with similar rationale was suggested by Livingston (1976): He used stochastic approximation techniques to arrive at the value of *C*.

Two other procedures, those of Livingston (1975) and Van der Linden and Mellenbergh (1977), also rely on the existence of an external criterion. In addition, they use linear functions to quantify the expected loss from the two kinds of error. Because these methods only displace the standard-setting problem to the criterion variable, they will not be so useful, even in adjusting a standard, as some of the Bayesian methods which also incorporate utility functions.

Kriewall (1969, 1972) was one of the first to cast the standard-setting problem in decision theoretic terms and to propose a model specifically accounting for classification errors. To use Kriewall's method, one must already have some idea of the range of test scores where the cutoff score will be located. Then by selecting different values for upper and lower bounds to the mastery range and using a model of independent Bernoulli trials, one can study the consistency of classifications for an individual precisely or exactly at the cutting score. To apply this model, however, one has to assume item homogeneity that is not likely to exist with minimum competency tests. Moreover, use of this procedure is analogous to selecting a cutoff score to maximize reliability rather than choosing a passing point that is substantively more defensible. Perhaps it is for this reason that Hambleton et al. (1979) found the model more applicable for determining test length. For this purpose it is identical to the procedure suggested by Millman (1972, 1973).

Various Bayesian approaches exist for adjusting cutoff scores. They all presuppose an available standard, which may be thought of as the desired level of performance on the entire domain. The simplest approach was illustrated by Davis and Diamond (1974). They assumed that the test administrator would have no other information about a student's true competence than the test score. Using Bayes theorem, they demonstrated how high the cut-score on the test would have to be to ensure that an examinee's true score was above the domain standard with a specified degree of confidence.

A more complicated and probably more useful approach was introduced by Novick and Lewis (1974). It involves specifying loss ratios and prior information on the distribution of examinee competence. Of course, a minimum criterion level has already been set so these manipulations will only supplement the standard-setting effort. The use of prior information on examinees in minimum competency testing is likely to be impractical and subject to the same validity questions as the tests themselves. The

specification of loss ratios, however, may be helpful to decision makers in considering the consequences of a range of cutting scores. Although the decision maker is not given any advice about how to determine the relative costs of false positives and false negatives, tables and computer simulations developed from this model do help visualize how much the cutting score and passing rates will change if different loss ratios are adopted. Novick and Lindley (1978) suggested a model for selecting a utility function to describe the degree of risk characterized by different levels of uncertainty. The normal distribution and other families of distributions appear to be more realistic for characterizing gains and losses than a threshold utility function which treats errors as equally serious regardless of how far they are from the cutting point.

A Composite Method for Setting Standards

All standard-setting methods are arbitrary and fallible (Jaeger, 1976; Glass, 1978b). Some methods are better than others for selecting a rational and defensible cutting point; but none of the models provides a scientific means for discovering the "true" standard. This is not only a deficiency of current methods but is a permanent and insolvable problem because the underlying competencies being measured are continuous and not dichotomous. There cannot be a clear and unambiguous distinction between masters and nonmasters.

Given the inadequacies of each approach to standard setting, the best solution is to avoid making absolute judgments whenever possible. In the last section of the chapter, alternatives are suggested for instances when minimum competency testing is intended for school accountability rather than individual certification. When cutoff scores are believed to be essential, then a combination of methods should be used so that the insights gained from each approach can all contribute to the final standard. This is similar to the "triangulation" of measurement processes recommended by Webb, Campbell, Schwartz, and Sechrest (1966). If each method of measurement is fallible, the use of multiple measures will ensure more interpretable results, since we have reason to believe that the same errors will not be repeated in each measurement technique. (The situations are not perfectly analogous, of course, since there is not a "truth" for the standard-setting methods to converge upon.)

The best way to construct a composite model is to select the best method from each of the types reviewed. Judgments about test content are essential to avoid complete reliance on normative standards. Judges should be selected to represent important audiences (Jaeger, 1978; Shepard, 1976) and should have the opportunity to work independently before deliberating,

so that records can be kept of persistent differences of opinion for judges of a particular type in different panels. The Angoff (1971) and Jaeger (1978) methods are both practical and consistent with the conceptualization of minimum competence. Regardless of which is selected, the judges should be given normative data to consider in making their ratings (Conaway, 1979; Hambleton et al., 1979; Jaeger, 1978; Shepard, 1976; Zieky & Livingston, 1977).

The Contrasting-Groups method (Zieky & Livingston, 1977) is the preferred method based on judgments about groups. Although it may seem expensive to use both an empirical study and judgments about test content, confirmation of the test's relationship to the distinction between master and nonmaster criterion groups is essential for validation in any case. The extra computations necessary to identify the optimal cutoff score are trivial. Beyond this, however, for minimum-competency testing applications, the empirical methods for discovering standards (Berk, 1976; Block, 1972) add nothing to the method based on judgments about masters and nonmasters. Also, the various statistical techniques (Livingston, 1975, 1976; Huynh, 1976b; Van der Linden & Mellenbergh, 1977) which presume the existence of a standard on an external criterion dimension are inappropriate, since an external criterion has neither been defined nor operationalized.

Normative data are essential for setting realistic passing rates. Given the definition of minimum there is only a small range of values for the percentage failing that is plausible. A different representative group of judges from those who implement the Angoff or Jaeger method should decide on acceptable failure rates. (The same group could be used if the normative data provided are not in a form to "give away" the failure rate expected from a logically determined standard.) The Bayesian methods for adjusting standards proposed by Novick and Lewis (1974), and more recent work by Novick and Lindley (1978) on realistic utility functions, should be helpful in conceptualizing how much a cutoff must be adjusted to reflect beliefs about the seriousness of different misclassification errors.

If these steps are followed one can imagine a result where an administrative body has the following information: a range of cutoff scores and an average cutoff agreed upon by several panels of judges, an optimal cutoff score identified empirically by the Contrasting-Groups method, a range of failure percentages agreed upon by a different group of judges, and a set of adjustments to be applied to the substantively chosen standard reflecting different plausible loss functions (also agreed upon by the judges). These pieces of information will undoubtedly suggest different cutoff scores and must be reconciled. If the proposed cutoffs cover more than half the range of the test, the evidence is compelling that a defensible standard cannot be set at all, and stronger legal steps might be taken to prevent implementing the testing program. If, however, the proposed standards

differ less, covering perhaps only a range of 20 percent in proportion correct score, then some intermediate value can be selected and rationalized by considering the relative importance of the different sources of information.

USES OF MINIMUM COMPETENCY TESTS

It is now conventional wisdom to consider the relative costs of the two types of errors when setting standards. But a prior question one might ask is, "What is the purpose of certification?" Given the extreme difficulty of validating a minimum competency test and setting a cutting score, there may be alternative ways to accomplish the same purpose that do not depend on standards-based testing. Three different uses for minimum competency tests can be envisioned as: (1) pupil classification for instructional purposes, (2) pupil certification, and (3) program evaluation. Conclusions about the adequacy of testing technology will depend on which of these purposes is addressed.

Pupil Classification for Instructional Purposes

Both pupil classification and pupil certification involve decisions about individuals. These two purposes for testing differ, however, with respect to frequency of testing and proximity of the testing hurdle to classroom level decisions. Throughout this chapter it has been assumed that minimum competency tests for high school graduation are of the second type. Their primary purpose is certification, and they are distant from the teaching-learning process. This is true even if remediation courses are required for students who fail the test; the relatively short and comprehensive competency tests are not likely to be thorough diagnostic instruments.

The first use of competency tests, for pupil classification within the classroom, is identical to the original purpose of criterion-referenced testing. It requires extensive pools of domain-referenced test items, linked to a specific curriculum, that can be administered at the discretion of the teacher to make short-term instructional decisions about needed review, workbook assignments, reading-group placement, advancement to new material, or remedial tutoring. It is implausible that such tests could be administered statewide, for example. To provide useful diagnostic information the tests have to be tailored to a specific curriculum and administered just at the time when the teacher is uncertain about what to do next with a particular child.

Minimum competency tests for the high school diploma and classroom criterion-referenced tests are at opposite ends of a continuum. Somewhere in between (but still closer to the certification use) are annual tests of competency for grade-to-grade promotion. This continuum of proximity to

the classroom was recognized implicitly by the National Academy of Education Committee on Testing and Basic Skills[5] and strongly influenced their conclusions about the adequacy of technology at the different levels.

> The NAEd Panel believes that any setting of statewide minimum competency standards for awarding the high school diploma—however understandable the public clamor which has produced the current movement and expectation—is basically unworkable, exceeds the present measurement arts of the teaching profession, and will create more social problems than it can conceivably solve.
>
> Setting minimum competency standards for high school graduation is the final step in a completely unsuccessful effort to reduce to precise behavioral terms what education is all about. Attempts to impose competency-based teacher education and competency-based preparation of school administrators have surely failed. The effort to determine and assure minimum competency standards for high school graduation will also fall of its own weight, for the scaffolding of existing test designs is too weak to carry such an emotionally-laden and ambiguous burden. Continuing any extensive efforts and funds in this direction is wasteful and takes attention away from the major tasks of improving our schools.
>
> However, the Panel is in agreement that a series of standardized tests at the lower grade levels used for diagnosing individual student weaknesses, pinpointing remediation needs, and building public pressures if school-wide performances in basic skills continue over time to be consistently low, could be positive influences on student learning. Such tests can have negative effects if they become the sole magnet of educational energies. But used with care, and with increased public (particularly teacher and parental) understanding of the diagnostic and therapeutic use (and limitations) of tests, minimum competency testing can be a positive educational development. (1978, p. 9)

The only disadvantage to classroom uses of criterion-referenced tests is cost: in dollars for test development and in both students and teacher time. Often teacher judgments about placement decisions will be just as accurate and less disruptive. Used judiciously, however, when the teacher is unsure of the student's level of comprehension, testing can greatly improve the focus and effectiveness of instruction.

The methodological problems reviewed in this chapter do not pertain so strongly to classroom level uses. Test validation is essential but not so difficult because such great inferences are not made regarding transfer of training to skills outside of school. The standard-setting problem is not so grave because decisions are mediated by normative or relative comparisons (e.g., informal groupings of similar students are made) and errors in placement can be easily corrected.

Pupil Certification

The technical issues in minimum competency testing reviewed in this chapter are colored by the use of the tests to certify that students have learned "what they should have learned in school" or "what they need to know to succeed in life." The validity problems are insurmountable. They are similar in kind but very different in magnitude to the problems of

validating professional licensure examinations. In the case of law or medicine, there is better consensus about the content domain because there is only one career, not all possible careers, to represent. Also cutoff scores are influenced (though of course not automatically determined) by marketplace considerations such as the number of candidates and the number of doctors (or lawyers) needed.

Methodological problems in minimum competency testing are more serious than classroom level testing because the decision made with the test is so permanent. Popham (1978a, 1978c) argued that setting standards in minimum competency tests is no more arbitrary than traditional grading practices. However, it is reasonable to assume that any unfairness in these procedures will balance out over the years for each student; whereas, this is not possible in a single test.

Alternatives to minimum competency testing have been proposed. How reasonable they might be depends on the motive for initiating minimum competency testing. Assuming that the purpose was to make high school diplomas more meaningful to prospective employers, Novick (1979) suggested that measurement experts instead construct a whole series of tests better tailored to different occupations. Employers could select relevant tests and administer them to applicants. Page (1978) proposed "scaled certification" rather than "irrational dichotomies." Students could be given several tests of different desired competencies. Standardized scores reflecting their level of accomplishment, like those used on the SAT and ACT, could then be reported on their high school transcripts. A student's qualifications might then be reflected in a profile of scores such as 10th percentile in reading, 15th percentile in mathematics and 15th percentile in language skills.

When the case of *Peter Doe* vs. *San Francisco Unified School District* is cited to explain the purpose of minimum competency testing, it is assumed that the schools could have intervened to make sure all 12th-graders could read if there had only been a test administered to identify the deficiency. If the purpose of testing is to make sure that the school makes every possible effort and does not pass on failing students unwittingly, then perhaps some sort of staffing process should be required in early grades similar to that required for Individual Educational Programs (IEPs) in special education (by P.L. 94-142). Then, at least social promotion to the next grade would not occur unless parents, teachers, and other experts agreed that (1) everything possible was being done to teach the child and (2) promotion was desirable despite inadequate mastery of basic skills. Competency tests in the early grades could be one source of information for such a staffing process; evidence of regular functioning in the classroom would also be considered. Other alternatives are discussed in the next section, when the purpose is not

to certify the individual but to hold schools accountable for acceptable achievement levels.

Program Evaluation

Whenever the percentage of students meeting a standard is used to report on the quality of an instructional program, minimum competency testing is being used for program evaluation or accountability purposes. The concern is no longer with making go-nogo decisions about individuals: To graduate or not from high school? To repeat second grade or not? Instead, a second layer of standards must be imposed to determine if an acceptable number of students are answering correctly an acceptable number of items. The technology of standard setting, which is hardly adequate for pupil certification, is absolutely insupportable for program evaluation purposes. First, the double layer of standards creates two sources of error. Second, when comparisons are made among subject areas, errors in setting the standards will be indistinguishable from program strengths and weaknesses; that is, if a standard is set just a little too high (multiplied over all the students) the instructional program will look weak in this area. Similarly, programs (or program components) will look better than they are if the standard is just a little too lenient.

The results of the Florida Functional Literacy Test are the most striking example of how fallible standards can lead to wrong conclusions about differences in program effectiveness. On the basis of competency tests educators, legislators, and reporters concluded that students were worse off in mathematics than in communication skills. (See Glass, 1978a, for an extensive discussion.) The passing score for both the mathematics and communications (reading and writing) tests was set at 70 percent. This was done by committees of curriculum specialists from local districts without ever seeing the test items. (The committees did have the test objectives, but they were of a general nature and were not written as domain specifications.) When the tests were administered, 35 percent of the seniors failed the mathematics test and only 10 percent failed in communications. The response in Florida has been to spend more resources on remedial mathematics programs. The apparent weakness in mathematics, however, could have been caused entirely by having set tougher standards for that subject.

In order to make sound judgments about relative strengths and weaknesses in an instructional program, a firm basis of comparison is needed. A nonartifactual benchmark can be found in normative or longitudinal data. Although traditional standardized tests may not measure the desired content, there are other sources of normative comparisons (Shepard, 1979a). For example, the objective-referenced items of the

National Assessment of Educational Progress are reported with national percent correct; and the State of California developed a second- and third-grade Reading Test, including in it items from several publishers all with national norms. If tests are selected with content validity for a particular curriculum, then differences in average percentiles can be interpreted as real strengths and weaknesses in a school program. This occurs, for example, if a district finds that its scores are at the 80th percentile in mathematics computation but only at the 50th percentile in mathematics concepts. School officials in Florida could have been much more certain of the need to allocate more resources to mathematics than reading if their average scores had been respectively 25 and 5 percentile ranks below the national median.

Glass (1978b) offered another alternative to arbitrary standards for judging educational programs. Improvement or decline compared to previous performance is readily interpretable:

> Perhaps the only criterion that is safe and convincing in education is change. Increases in cognitive performance are generally regarded as good, decreases as bad. Although one cannot make satisfactory absolute judgments of performance (Is this level of reading performance good or masterful?), one can readily judge an improvement in performance as good and a decline as bad (p. 259).

An advocate of minimum competency testing might insist that preventing incompetents from receiving a diploma is so important, and will have so salutary an effect on the efforts of individuals, that it is worth the risk to use tests with weak validity and equivocal cutoff scores. So long as the very best professional judgments have been sought, and precautions against errors have been taken, a reasonable line can finally be drawn; and education will be better off in the long run for having made these tough decisions. The "lesser of two evils argument"—better to have fallible standards than no standards (Scriven, 1978)—applies to instances when dichotomous decisions are required for individuals. It has absolutely no bearing when the data are to be used to judge groups or programs. Dichotomies at the minimum competency end of the scale do not tell us what is happening to other students in a school or school system, because the rest of the score distribution is obscured. In addition, interpretation of results requires a second set of standards to decide what percent of students should be passing the test. To determine the effectiveness of an educational program it is much more informative to study the entire performance continuum. The value attached to performance levels can be supplied by relevant comparative data. (See Popham, 1976, regarding the use of norms with criterion-referenced tests.) In this way one can conclude whether students at the mean are achieving as much as they should, as well as whether students in the bottom quartile are learning enough.

CONCLUSION

Technical issues associated with criterion-referenced testing have been reviewed with special attention to their meaning in the context of minimum competency testing. The nature of the test decision, to classify individuals as competent or incompetent, determines the way in which reliability and validity questions are addressed. The methodological areas of greatest concern are validity and standard setting. Minimum competency testing is very different from classroom uses of criterion-referenced testing because of the greater inferences required regarding the construct validity of the test substance and because of the serious consequences attached to the selection of arbitrary cutoff points. The testing technology is good for classroom uses (for which it was originally designed). The same technology, however, is inadequate for certification of high school graduates. Moreover, the percent of students passing a competency exam is a poor statistic for judging the effectiveness of a school program because it tells nothing about the educational attainments of students who are distant from the cutoff score. Alternative ways to address these ends were considered, such as competency testing and staffing procedures, in the early grades to ensure individual competence and conjunctive use of norm-referenced and criterion-referenced tests to evaluate educational programs.

FOOTNOTES

[1]*Griggs et al.* v. *Duke Power Company* 401 U.S. 424, 1971

[2]*Debra P.* v. *Turlington*, 78-892-Civ-T-C, 1979.

[3]Some authors have suggested that judges should always be told the correct response to each question so that trying to figure out the answer will not hinder their judgments (Bernknopf, Curry, & Bashaw, 1979). However, the judges' own struggles with an item may influence their judgments of difficulty. Further, unless derelicts or other "unsuccessful" adults are sought as judges, it can be argued that there should not be questions on the test that the judges cannot answer.

[4]*Peter Doe* v. *San Francisco Unified School District* 60 C.A. 3d 814, 1976.

[5]The NAEd Panel was composed of Stephen K. Bailey, Graduate School of Education, Harvard University (Chairman); John B. Carroll, University of North Carolina, Chapel Hill; Jeanne Chall, Graduate School of Education, Harvard University; Robert Glaser, University of Pittsburgh; John I. Goodlad, University of California, Los Angeles; Diane Ravitch, Teachers College, Columbia University; Lauren Resnick, University of Pittsburgh; Ralph W. Tyler, Science Research Associates, Chicago; and Robert L. Thorndike, Teachers College, Columbia University.

REFERENCES

Airasian, P. W., Kellaghan, T., Madaus, G., & Ryan, J. P. *Payment by results: The analysis of a 19th century performance contracting program.* Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, April 1972.

American Psychological Association, American Educational Research Association & National Council on Measurement in Education. *Standards for educational and psychological tests.* Washington, D.C.: American Psychological Association, 1974.

Anastasi, A. *Psychological testing.* (4th ed.) New York: Macmillan, 1976.

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

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

Angoff, W. H. Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational Measurement.* Washington, D.C.: American Council on Education, 1971.

Atkinson, R.C. Computer-based instruction in initial reading. In *Proceedings of the 1967 Invitational Conference on Testing Problems.* Princeton, N.J.: Educational Testing Service, 1968.

Baker, E. L. *Using measurement to improve instruction.* Paper presented at the annual meeting of the American Psychological Association, 1972.

Barr, R., & Dreeben, R. Instruction in classrooms. In L. S. Shulman (Ed.), *Review of Research in Education*, Volume 5. Itasca, Ill.: F. E. Peacock, 1977.

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

Berk, R. A. *A critical review of content domain specification/item generation strategies for criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Bernknopf, S., Curry, A., & Bashaw, W. L. *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, 1979.

Block, J. H. (Ed.). Introduction. *Mastery learning: Theory and practice.* New York: Holt, Rinehart, & Winston, 1971. (a)

Block, J. H. (Ed.). Operating procedures for mastery learning. *Mastery learning: Theory and practice.* New York: Holt, Rinehart, & Winston, 1971. (b)

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

Block, J. H. Schools, society and mastery learning. New York: Holt, Rinehart, & Winston, 1974.

Block, J. H. Standards and criteria: A response. *Journal of Educational Measurement*, 1978, *15*, 291-295.

Bloom, B. S. Learning for mastery. *Evaluation Comment* Vol. 1. UCLA-CSEIP, 1968, *1*, n.p.

Bloom, B. S. Mastery learning. In J. H. Block (Ed.), *Mastery learning: Theory and practice.* New York: Holt, Rinehart, & Winston, 1971.

Bloom, B. S. *Human characteristics and school learning.* New York: McGraw-Hill, 1976.

Bobula, J. A., & Standish, M. Minimum pass level study. *Report to the Faculty.* Center for Educational Development, University of Illinois College of Medicine, 1974, 25-31.

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

Brennan, R. L. A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 1972, *32*, 289-303.

Brennan, R. L. *Some applications of generalizability theory to the dependability of domain-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association and National Council on Measurement in Education, San Francisco, April 1979.

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

Brennan, R. L., & Lockwood, R. E. A comparison of the Nedelsky and Angoff cutting score procedures using generalizability theory. *Applied Psychological Measurement*, in press.

Brickell, H. M. Seven key notes on minimum competency testing. In B. S. Miller (Ed.), *Minimum competency testing: A report of four regional conferences*. St. Louis, Mo.: CEMREL, 1978.

Buros, O.K. Fifty years in testing: Part I. Some reminiscences, criticisms and suggestions. In O.K. Buros (Ed.), *The eighth mental measurements yearbook*, Vol. II. Highland Park, N.J.: Gryphon Press, 1978.

Burton, N. Societal standards. *Journal of Educational Measurement*, 1978, *15*, 263-271.

Carroll, J. B. A model of school learning. *Teachers College Record*, 1963, *64*, 723-733.

Carroll, J. B. Problems of measurement related to the concept of learning for mastery. *Educational Horizons*, 1970, *48*, 71-80.

Civil Service Commission, Equal Employment Opportunity Commission, Department of Justice, Department of Labor. Uniform guidelines on employee selection procedures. *Federal Register*, 1977, *42*(251), 65,542-65,552.

Coffman, W. E. Essay examinations. In R. L. Thorndike (Ed.), *Educational measurement*, (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cohen, J. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 1960, *20*, 37-46.

Cohen, J. Weighted Kappa: Nominal scale agreement with provision for scaled disagreement of partial credit. *Psychological Bulletin*, 1968, *70*, 213-220.

Conaway, L. E. Discussant comments: Setting performance standards based on limited research. *Florida Journal of Educational Research*, 1976, *18*, 35-36.

Conaway, L. E. Setting standards in competency-based education: Some current practices and concerns. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency based education*. Washington, D.C.: National Council on Measurement in Education, 1979.

Cox, R. C., & Vargas, J. C. *A comparison of item selection techniques for norm-referenced and criterion-referenced tests.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, February 1972.

Crehan, K. D. Item analysis for teacher-made mastery tests. *Journal of Educational Measurement*, 1974, *11*, 255-262.

Cronbach, L. J. How can instruction be adapted to individual differences? In R. M. Gagné (Ed.), *Learning and individual differences*. Columbus, Ohio: Charles E. Merrill, 1967.

Cronbach, L. J. Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Cronbach, L. J. Dissent from Carver. *American Psychologist*, 1975, *30*, 602-603.

Cronbach, L. J., & Gleser, G. C. *Psychological tests and personnel decisions*. Urbana, Ill.: University of Illinois Press, 1965.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. New York: Wiley, 1972.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, 1963, *16*, 137-163.

Davis, F. B., & Diamond, J. J. The preparation of criterion-referenced tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement*. (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Dickson, G. E. (Ed.). Research and evaluation in operational competency-based teacher education programs. *Educational comment*, Vol. 1. Toledo, Ohio: University of Toledo, 1975.

Ebel, R. L. Content standard test scores. *Educational and Psychological Measurement*, 1962, *22*, 15-25.

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.

Ebel, R. L. The case for minimum competency testing. *Phi Delta Kappan*, 1978, *60*, 546–548. (a)

Ebel, R. L. The case for norm-referenced measurements. *Educational Researcher*, 1978, *7*, 3-5. (b)

Educational Testing Service. *Report on a study of the use of the National Teachers Examination by the State of South Carolina*. Princeton, N.J.: Educational Testing Service, 1976.

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

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

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of Kappa and weighted Kappa. *Psychological Bulletin*, 1969, *72*, 323-327.

Fremer, J. *Using and interpreting the results of criterion-referenced tests: Promise, progress, and unresolved problems.* Paper presented at the first annual Johns Hopkins University National Symposium on Educational Research, "Criterion-Referenced Measurement: The State of the Art." Washington, D. C., October 27, 1978.

Getz, J. E., & Glass, G. V Lawyers and courts as architects of educational policy: The case of minimal competence testing. *The High School Journal*, 1979, *62*, 181-186.

Glaser, R. Instructional technology and the measurement of learning outcomes. *American Psychologist*, 1963, *18*, 519–521.

Glaser, R. Adapting the elementary school curriculum to individual performance. In *Proceedings of the 1967 invitational conference on testing problems*. Princeton, N. J.: Educational Testing Service, 1968.

Glaser, R. *Adaptive education: Individual diversity and learning*. New York: Holt, Rinehart, & Winston, 1977.

Glass, G. V Minimum competence and incompetence in Florida. *Phi Delta Kappan*, 1978, *59*(9), 602-605. (a)

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

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. *Journal of the American Statistical Association*, 1954, *49*, 732-764.

Guttman, L. Integration of test design and analysis. In *Proceedings of the 1969 invitational conference on testing problems*. Princeton, N.J.: Educational Testing Service, 1969.

Haladyna, T. M. Effects on different samples on item and test characteristics of criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 93-99.

Haladyna, T. M., & Roid, G. H. *The quality of domain-referenced test items*. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1976.

Hambleton, R. K. On the use of cut-off scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 1978, *15*, 277-290.

Hambleton, R. K. *Determining the validity of competency tests*. Paper presented at the 19th Annual Conference on Large-Scale Assessment. Sponsored by the National Assessment of Educational Progress, Denver, Colo., June 1979.

Hambleton, R. K., & Eignor, D. R. Competency test development, validation, and standard setting. In R. Jaeger & C. Tittle (Eds.), *Minimum competency testing*. Berkeley, Calif.: McCutchan, 1979.

Hambleton, R. K., & Fitzpatrick, A. *Review techniques for criterion-referenced test items*. Manuscript in preparation.

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

Hambleton, R. K., Powell, S., & Eignor, D. R. Issues and methods for standard setting. In *A practitioner's guide to criterion-referenced test development, validation, and test score usage. Laboratory of Psychometric and Evaluative Research Report No. 70* (2nd ed.) Amherst, Mass.: School of Education, University of Massachusetts, 1979.

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

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. *Journal of Educational Measurement*, 1972, *9*, 27-29.

Harris, C. W. Problems of objectives-based measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974. (a)

Harris, C. W. Some technical characteristics of mastery tests. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974. (b)

Harris, C. W., Pearlman, A. P., & Wilcox, R. R. (Eds.). *Achievement test items—Methods of study* (CSE Monograph Series in Evaluation, No. 6). Los Angeles: University of California, Center for the Study of Evaluation, 1977.

Hills, J. R. *Construct validation of the Florida Functional Literacy Test*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Hively, W. (Ed.). *Domain-referenced testing*. Englewood Cliffs, N. J.: Educational Technology Publications, 1974.

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

Hively, W., Patterson, H. L., & Page, S. A. A "universe-defined" system of arithmetic achievement tests. Journal of Educational Measurement, 1968, *5*, 275-290.

Hubert, L. J. Kappa revisited. *Psychological Bulletin*, 1977, *84*, 289-297.

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

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

Jaeger, R. *Measurement consequences of selected standard-setting models*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1976.

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

Jaeger, R. M. Measurement consequences of selected standard-setting models. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement*. Washington, D.C.: National Council on Measurement in Education, 1979.

Jordan, J. E. Attitude-behavior research on physical-mental-social disability and racial-ethnic differences. *Psychological Aspects of Disability*, 1971, *18*, 5-26.

Keller, F. S. Goodbye, teacher . . . *Journal of Applied Behavior Analysis*, 1968, *1*, 79-89.

Kendall, M. G., & Stuart, M. A. *The advanced theory of statistics, Vol. III: Design and analysis and time-series*. New York: Hafner, 1966.

Kriewall, T. E. *Application of information theory and acceptance sampling principles to the management of mathematics instruction*. Unpublished doctoral dissertation, University of Wisconsin, 1969.

Kriewall, T. E. *Aspects and applications of criterion-referenced tests*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Kulik, C. L., & Kulik, J. A. PSI and the mastery model. In B. A. Green (Ed.), *Proceedings of the second national behavioral instruction conference*. Washington, D.C.: Center for Personalized Instruction, 1976.

Lindvall, C. M., & Bolvin, J. O. Programmed instruction in the schools: An application of programming principles in "individually prescribed instruction." In P. C. Lange (Ed.), *Programmed instruction* (Sixty-sixth Yearbook of the National Society for the Study of Education, pt. 2). Chicago: NSSE, 1967.

Linn, R.L. *Construct validity and measurement of competency-based education*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, 1976.

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

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

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

Livingston, S. A. *A utility-based approach to the evaluation of pass/fail testing decision procedures* (Report No. COPA-75-01). Princeton, N. J.: Educational Testing Service, 1975.

Livingston, S. A. *Choosing minimum passing scores by stochastic approximation techniques* (Report No. COPA-76-02). Princeton, N. J.: Educational Testing Service, 1976.

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

Madaus, G. F. *Measurement issues and consequences associated with minimal competency testing.* Paper presented at the Spring Membership Conference of the National Consortium on Testing, Arlington, Virginia, May, 1978.

Mager, R. F. *Preparing instructional objectives.* Palo Alto, Calif.: Feardon, 1962.

Marshall, J. L., & Haertel, E. H. *A single-administration reliability index for criterion-referenced tests: The mean split-half coefficient of agreement.* Paper presented at the annual meeting of the American Educational Research Association, Washington, D.C., 1975.

McClung, M. S. Competency testing: Potential for discrimination. *Clearinghouse Review*, 1977, 439-448.

McClung, M. S. Are competency testing programs fair? legal? *Phi Delta Kappan*, 1978, *60*, 397-400.

Meskauskas, J. A. Evaluation models for criterion-referenced testing: Views regarding mastery and standard-setting. *Review of Educational Research*, 1976, *45*, 133-158.

Meskauskas, J. A., & Webster, G. D. The American Board of Internal Medicine Recertification Examination: Process and results. *Annals of Internal Medicine*, 1975, *82*, 577-581.

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

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

Millman, J. Tables for determining number of items needed on domain-referenced tests and number of students to be tested. *Technical Paper No. 5.* Los Angeles: Instructional Objectives Exchange, 1972.

Millman, J. Passing scores and test lengths for domain-referenced measures. *Review of Educational Research*, 1973, *43*, 205-216.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), *Evaluation in education: Current applications.* Berkeley, Calif.: McCutchan, 1974.

Millman, J., & Popham, W. J. The issues of item and test variance for criterion-referenced tests: A clarification. *Journal of Educational Measurement*, 1974, *11*, 137-138.

National Academy of Education. *Improving educational achievement.* Washington, D.C.: author, 1978.

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

Nitko, A. J. Problems in the development of criterion-referenced tests: The IPI Pittsburgh experience. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Novick, M. R. *The great cut-score debate.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

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

Novick, M. R., & Lindley, D. V. The use of more realistic utility functions in educational applications. *Journal of Educational Measurement*, 1978, *15*, 181-191.

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

Page, E. B. Escaping the minimum competency dilemma: Scaled certification. In *Proceedings of the National Conference on Minimum Competency Testing.* Portland, Ore.: Clearinghouse for Applied Performance Testing, 1978.

Paiva, R. E. A., & Vu, N. V. *Standards for acceptable level of performance in an objectives-based medical curriculum: A case study.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Pipho, C. *Update VII: Minimal competency testing* (Report No. 105). Denver, Colo.: Education Commission of the States, 1977.

Pipho, C. Minimum competency testing in 1978: A look at state standards. *Phi Delta Kappan*, 1978, *59*, 585-587.

Popham, W. J. 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). Los Angeles: University of California, Center for the Study of Evaluation, 1974.

Popham, W. J. *Educational evaluation.* Englewood Cliffs, N. J.: Prentice-Hall, 1975.

Popham, W. J. Normative data for criterion-referenced tests? *Phi Delta Kappan*, 1976, *58*, 593-594.

Popham, W. J. As always provocative. *Journal of Educational Measurement*, 1978, *15*, 297-300. (a)

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

Popham, W. J. *Key standard-setting considerations for minimum competency testing programs.* A presentation at the SCE Invitational Winter Conference in Measurement and Methodology, UCLA, Center for the Study of Evaluation, Los Angeles, January 1978. (c)

Popham, W. J., & Husek, T. R. Implications of criterion-referenced measurement. *Journal of Educational Measurement*, 1969, *6*, 1-9.

Rickover, H. G. Do we really need a national competency test? *The National Elementary Principal*, 1978, *57*, 48-56.

Robin, A. L. Behavioral instructions in the college classroom. *Review of Educational Research*, 1976, *46*, 313-354.

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

Scriven, M. How to anchor standards. *Journal of Educational Measurement*, 1978, *15*, 273-275.

Shepard, L. A. Setting standards and living with them. *Florida Journal of Educational Research*, 1976, *18*, 23-32.

Shepard, L. Norm-referenced vs. criterion-referenced tests. *Educational Horizons*, 1979, 26-32. (a)

Shepard, L. A. Setting standards. In M. A. Bunda & J. R. Sanders (Eds.), *Practices and problems in competency-based measurement*. Washington, D. C.: National Council on Measurement in Education, 1979. (b)

Shoemaker, D. M. Toward a framework for achievement testing. *Review of Educational Research*, 1975, *45*, 127-147.

Skinner, B. F. The science of learning and the art of teaching. *Harvard Educational Review*, 1954, *24*, 86-97.

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

Stanley, J. C., & Hopkins, K. D. *Educational and psychological measurement and evaluation*. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Subkoviak, M. J. Estimating reliability from a single administration of a criterion-referenced test. *Journal of Educational Measurement*, 1976, *13*, 265-276.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. *Journal of Educational Measurement*, 1978, *15*, 111-116.

Subkoviak, M. J. Decision consistency approaches. In A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore: Johns Hopkins University Press, 1980.

Subkoviak, M. J., & Baker, F. B. Test theory. In L. S. Shulman (Ed.), *Review of research in education*, Vol. 5. Itasca, Ill.: F. E. Peacock, 1977.

Suppes, P. The uses of computers in education. *Scientific American*, 1966, *215*, 206-221.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. *Journal of Educational Measurement*, 1974, *11*, 262-267.

Talmage, H. (Ed.). *Systems of individualized education*. Berkeley, Calif.: McCutchan, 1975.

Trivett, D. A. *Competency programs in higher education*. ERIC/Higher Education Research Report No. 7, 1975.

Tunks, T. W. An application of Guttman facet theory to attitude scale construction in music. *Council for Research in Music Bulletin*, 1973, *33*, 47-53.

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

Webb, E. J., Campbell, D. T., Schwartz, R. D., & Sechrest, L. *Unobtrusive measures: Nonreactive research in the social sciences*. Chicago: Rand McNally, 1966.

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. *Journal of Educational Measurement*, 1974, *11*, 63-64. (a)

Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests: A reply. *Journal of Educational Measurement*, 1974, *11*, 139-140. (b)

Zieky, M. J., & Livingston, S. A. *Manual for setting standards on the Basic Skills Assessment Tests*. Princeton, N. J.: Educational Testing Service, 1977.