A Framework for Methodological Advances in Criterion-Referenced Testing

Ronald A. Berk
The Johns Hopkins University

A vast body of methodological research on criterion-referenced testing has been amassed over the past decade. Much of that research is synthesized in the articles contained in this issue. The fact that this issue is devoted exclusively to criterion-referenced testing sets it apart as a quintessential journal publication on the topic.

This paper is intended to provide a broad framework for understanding and evaluating the individual contributions in the context of the literature. The six articles appear to fall into four major categories: (1) test length; (2) validity; (3) standard setting; and (4) reliability. These categories correspond to most of the technical topics in the test development process (see, e.g., Berk, 1980b; Hambleton, 1980).

Test Length

If a teacher or curriculum specialist asked a psychometrician, 'How many items should be written for each objective?' or 'How many items should be sampled from the domain?' what answer *could* be given? Whether the test is norm referenced or criterion referenced, there is no available source that recommends a magical number of items. The guidelines offered in most measurement texts and technical papers on the topic are rather nonspecific; however, this perplexing issue cannot be dismissed or ignored, as are the properties of validity and reliability analyzed in subsequent sections of this paper. Every test maker, teacher through test publisher, must answer the test length question.

The problem of determining test length can be approached from a practical perspective based on research evidence and/or from a purely technical perspective. The former has been explicated previously in terms of a multiplicity of factors, including importance and type of decision making, importance and emphases assigned to objectives, number of objectives, and practical constraints (Berk, 1980c); the latter constitutes the orientation of Wilcox's article.

Wilcox has organized his review according to three achievement test conceptualizations based on three different types of true score: (1) the number of items a student would answer correctly if every item in the item domain was answered; (2) the proportion of skills among a domain of skills that a

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 4, No. 4 Fall 1980 pp. 563–573

student has acquired; and (3) latent trait models. For the first conceptualization where the cutoff score is known, Wilcox evaluates the binomial probability function, binomial error model, Bayesian decision-theoretic model, and beta-binomial model. Other Bayesian approaches are examined for situations where the cutoff score is unknown. In the second section two approaches are considered: One deals with a specific skill viewed in terms of a population of students and the second focuses on an individual student in terms of a domain of skills. The final section is devoted to solutions using latent trait models. Wilcox raises questions about the appropriateness of these models for the test length problem and the condition where there are items that do not fit a particular model.

After this rather exhaustive survey of potential solutions to the question posed at the beginning of this discussion, Wilcox's findings seem inconclusive. He offers no specific recommendations on how to proceed or what "best" solutions merit further attention. However, what is most troublesome is that there is no designation as to who can and should use any of those models. Certainly, they are prohibitive for classroom teachers and, probably, for a large number of district evaluators, given the required statistical sophistication and the computer resources essential to calculate the various estimates. There are at least two exceptions: the binomial models described for the first conceptualization and the latent trait models, which have been quite popular in several school districts and state departments of education.

Validity

Among the topics addressed in this issue, validity has been the most neglected in the context of the criterion-referenced measurement literature. Despite its importance in test construction and score interpretation and use, only three treatments of validity have been reported, and they were published only within the past two years (Hambleton, 1980; Linn, 1979b; Millman, 1979). Linn's article reviews some of the key issues in validity. His contribution will be assessed in light of these recent developments.

One theme permeating much of the research in criterion-referenced testing is that content validity via explicit content domain specifications is necessary to assure clarity and meaning in test score interpretation. Numerous strategies that extend or replace existing objectives-based specifications have been proposed (see reviews by Berk, 1980a; Millman, 1980; Popham, 1980; Roid & Haladyna, 1980). Linn points out that this emphasis on procedures to establish a domain definition and content representativeness is, in fact, necessary but not sufficient to support the diverse interpretations and uses of criterion-referenced test scores. Other kinds of evidence are needed.

These "other kinds of evidence" should be gathered from traditional criterion-related and construct validation studies. Several methods have been described by Hambleton (1980) and some concrete examples are provided by Linn. What is particularly significant about Linn's discussion of this issue, however, is his persuasive argument for a unified conceptualization of validity in contrast to the current compartmentalized conceptualization of validity. The latter tends to perpetuate the notion that there are alternative approaches and that a test maker need only choose one of them.

The impetus for a unified conceptualization stems largely from the ideas advanced by Cronbach (1980), Guion (1977, 1978, in press), Messick (1979), and Tenopyr (1977). Linn strongly urges test makers to view the different types of validity as approaches to accumulating certain kinds of evidence rather than as alternative approaches. Toward this end, he suggests that test makers initially concentrate on the score interpretations, uses, and inferences instead of on the type of validity that is required.

The message conveyed by Linn is lucid, constructive, and meaningful. Its realization will constitute a shift in thinking and, ultimately, in practice. The stress on obtaining evidence to substantiate

the various decisions that could be made with test scores should produce a more complete documentation of validity for all achievement tests. Given the present and anticipated uses of criterion-referenced and minimum-competency test scores, this new direction is especially important.

Standard Setting

More than 20 different cutoff-score methods for criterion-referenced tests have been recommended in the literature. Despite several extensive reviews of these methods by Hambleton (1980), Hambleton, Powell, and Eignor (1979), Meskauskas (1976), and now Shepard (in this issue), standard setting is still the most complicated technical topic in criterion-referenced measurement. The articles by Shepard, Macready and Dayton, and van der Linden address various aspects of the topic. This section will propose a framework for understanding their contributions to the research.

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

State Models

State models assume that mastery or true-score performance is an all-or-nothing state, and the standard is therefore set at 100%. Deviations from this true state are presumed attributable to "intrusion" (false mastery) and/or "omission" (false nonmastery) errors. After a consideration of these errors, the standard is adjusted to values less than 100%. Glass (1978) referred to these models as "counting backwards from 100%" (p. 244).

Given the amount of research that has accumulated on standard setting, state models have received relatively little attention. The Macready and Dayton article provides the most comprehensive survey of state models to date. Although the authors claim that the models are nonjudgmental in nature, they possess many of the same judgmental and empirical characteristics of the decision-theoretic approaches for continuum mastery models. In fact, their article might be retitled "Decision-Theoretic Approaches for State or Latent Class Models." A further discussion of this point follows.

Judgmental-empirical models.[1] The various models employ decision rules to identify the cutoff score that minimizes expected loss due to classification errors (see, e.g., Bergan, Cancelli, & Luiten, 1980; Emrick, 1971; Macready & Dayton, 1977). These rules require judgment in designating the loss ratio. The subjectivity involved in this process is described at length by Shepard. Macready and Dayton indicate that all decision-making must incorporate, implicitly or explicitly, a weighting of losses.

[1]Although the emphasis here is on state models for setting standards, some of the models reviewed by Macready and Dayton deal with item reliability (e.g., Knapp, 1977; Werts, Linn, & Jöreskog, 1973; Wilcox, 1977a, 1977b).

Yet they also note that this judgmental component can be eliminated by setting the loss ratio equal to 1.0. In addition, the authors recommend a judgmental assessment of parameter estimates in conjunction with the absolute and relative statistical assessments of model fit. Clearly, judgment is an integral part of the decision-theoretic state models.

There are a few limitations of the models that render them less compatible with current practices in criterion-referenced testing than the continuum models. One limitation is that some of the models (e.g., Knapp, 1977; Roudabush, 1974; Wilcox, 1977a, 1977b) are based on mastery of only one or two items. Decisions at the item level would be appropriate, for example, in the context of algorithmic testing, as in Scandura's (1977) structural learning theoretic approach. Unfortunately, the use of a single item to measure attainment of a skill is extremely restrictive in view of the structure and imprecision of most content domain specifications. Coupled with this limitation is the requisite homogeneity of the domain. Only discrete pieces of information (such as facts or terminology) or skills where perfection is essential would produce an adequate model fit. This contrains the application of the models to low-level cognitive skills and ultra-specific objectives. The third limitation pertains to the requisite homogeneity of the student population that is tested. The models assume that masters answer all items correctly and that they have an equal chance of incurring an inappropriate response (omission error) to an item. The converse assumptions exist for nonmasters. Most intact classes are more heterogeneous than these assumptions would permit. Probably the composition of certain specially formed classes would provide the necessary homogeneity.

Given this rather negative appraisal of the present state of state models, one final comment regarding their potential seems warranted. Macready and Dayton offer a three-factor scheme for defining all classes of latent state models based on the attributes of level of item response, model type, and presence of covariate. They also propose several model extensions (e.g., covariate state models) in order to overcome some of the aforementioned limitations. These contributions furnish a meaningful structure and direction for future research on state models. The potential of those models can be improved through deliberate efforts to establish greater congruence with actual testing practices.

Continuum Models

Continuum models assume that mastery is a continuously distributed ability that can be viewed as an interval on a continuum; i.e., an area at the upper end of the continuum circumscribes the boundaries for mastery. This conceptualization appears to fit the design and intent of most criterion-referenced tests. It is therefore not surprising that the bulk of the research has concentrated on continuum models. The majority of the cutoff score methods developed within the past five years fall into this category. Consequently, the reviews cited at the beginning of this section, including Shepard's, have focused primarily on these methods.

Although it was inevitable that the experts on criterion-referenced testing would not agree on a "best method" for setting standards, there is consensus on one issue—all of the methods involve some form of human judgment. A completely objective, scientifically precise method does not exist. Regardless of how complex and technically sophisticated a method might be, judgment plays a role in the determination of the cutoff score and/or in the estimation of classification error rates.

Judgmental methods. These methods are based on judgments of the probability that competent persons would select particular distractors in each item (Nedelsky, 1954) or would answer each item correctly (Angoff, 1971; Ebel, 1979; Jaeger, 1978). The subjectivity of these item content decisions used to arrive at an overall cutoff score was expressed succinctly by Shepard: Judges have the sense that they are "pulling the probabilities from thin air." This problem is reflected in the variability among judgments within a single method and also across methods (see Shepard's review of recent

studies of this problem). For further descriptions of these methods and related issues, interested readers should consult the reviews by Popham (1978) and Zieky and Livingston (1977) in addition to those previously mentioned.

Judgmental-empirical methods. This category consolidates all other cutoff score methods that Shepard has labeled "standards based on judgments about groups," "empirical methods for discovering standards," and "empirical methods for adjusting standards." The reason for combining these into one category is for convenience and simplicity. The methods are based on some type of judgment and actual or simulated data, judgmental data, and/or distribution assumptions. To clarify this point and to justify this classification, the specific judgmental and empirical components in ten continuum methods that have been given wide visibility in the literature are defined in Table 1. The role of judgment should not be underestimated. That is, the judgmental component usually supplies the foundation for much of the statistical estimation of probabilities of correct classification decisions and false mastery/false nonmastery decision errors.

The judgmental-empirical methods differ according to other characteristics as well: (1) their overall purpose; (2) definition of the criterion variable; (3) consideration of utilities; (4) statistical sophistication; and (5) practicability. Perhaps the most important initial distinction that Shepard makes between these methods pertains to their purpose. Only the Berk (1976) and Zieky and Livingston (1977) approaches are intended to select a cutoff score. All of the remaining methods presume that a standard already exists on a criterion or latent variable. Subsequently, this standard is translated into a cutoff score for the test, and decision error rates based on various assumptions are estimated. In some cases, those rates can be used to adjust the cutoff. Van der Linden emphasizes that the decision-theoretic models are not techniques for setting standards or optimizing mastery decisions; they *are* techniques for minimizing the consequences of measurement and sampling errors once the true cutoff has already been chosen.

The decision-theoretic approaches reviewed by van der Linden are associated almost exclusively with the last method in Table 1. They are the most theoretically and statistically complex continuum methods. Van der Linden's presentation of six methods is organized in terms of the type of loss function employed. This seems quite useful, since the type of decision to be made with the scores and the assumption about losses determine the choice of a standard setting method and a reliability index. (See also Shepard's summary of some of these methods.) Three loss functions are reviewed: (1) threshold loss (e.g., Hambleton & Novick, 1973; Huynh, 1976; Mellenbergh, Koppelaar, & van der Linden, 1977); (2) linear loss (e.g., van der Linden & Mellenbergh, 1977); and (3) normal ogive loss (e.g., Novick & Lindley, 1978). Although the threshold loss function has received the most attention to date, the normal ogive loss appears to have considerable potential in criterion-referenced testing. Further research on loss functions is suggested to address not only the standard-setting issues but also the problems related to reliability (see Traub & Rowley's article), test length (see Wilcox's article), and test score equating. Inter alia, one should be cognizant of the distinction indicated previously—that the decision-theoretic models constitute a circuitous solution to the cuttoff score problem by augmenting as opposed to actually determining a standard (cf. Hambleton et al., 1979).

Decision Applications

From the methods reviewed in the Shepard, Macready and Dayton, and van der Linden articles, certain factors emerge that seem crucial to the selection of a cutoff score method, the most important of which is the decision to be made with the test scores. Shepard describes three kinds of decision application: (1) pupil diagnosis; (2) pupil certification; and (3) program evaluation. The reviews of the

Table 1
Judgmental and Empirical Components of Continuum Methods for Setting Cutoff Scores and/or Estimating Error Rates (Listed in Order of Increasing Overall Complexity)

			Empirical Component		
Method	*Source*	*Judgmental Component*	*Actual Data*	*Judgmental Data*	*Distribution Assumptions*
Educational Consequences	Block (1972)	Selection of criterion variable	X		
Criterion Groups	Berk (1976)	Selection of intact criterion groups	X		
Contrasting Groups/ Borderline Groups	Zieky & Livingston (1977)	Selection of individuals to comprise comparison groups	X		
Binomial Model	Kriewall (1972)	Setting boundaries for mastery and nonmastery ranges			X
Utility Based	Livingston (1975)	Selection of criterion variable; assignment of benefits/costs	X	X	
Linear Loss Function	van der Linden & Mellenbergh (1977)	Selection of cutoff for latent variable; assignment of losses	X	X	X
Stochastic Approximation	Livingston (in press)	Selection of performance criterion	X	X	
Control Comparison	Wilcox (1979a)	Selection of control by panel of judges	X		X
Beta-Binomial Model (Empirical Bayesian)	Huynh (1976), Huynh & Saunders (1979), Wilcox (1979b)	Selection of referral task	X		X
Bayesian Decision Model	Novick & Lewis (1974), Schoon, Guillon, & Ferrara (1979), Swaminathan, Hambleton, & Algina (1975)	Setting prior probabilities and loss ratio	X	X	X

decision-theoretic state and continuum models stress the first instructional application, although their practicability in that context is highly questionable. Shepard urges the use of particular judgmental and judgmental-empirical continuum methods for pupil certification decisions. However, she considers their utility for pupil diagnosis and program evaluation somewhat restricted.

Reliability

Similar to the preceding topic, the literature is replete with studies and reviews of criterion-referenced test reliability. More than a dozen indices have been proposed; and critical reviews of those indices have been conducted by Berk (in press), Hambleton, Swaminathan, Algina, and Coulson (1978), Linn (1979a), Millman (1979), and now Traub and Rowley (in this issue). This section will examine the contribution of the latter review.

Traub and Rowley's presentation is organized according to two major dimensions: type of variable (continuous or binary) and intended use of test score (decision making or measurement). The first dimension was adopted from Graham and Bergquist's (1975) distinction between binary and continuous models of criterion-referenced measurement. These are analogous to the state and continuum models defined earlier. The second dimension, adopted from the signal work of Hambleton et al. (1978) on the topic, differentiates between reliability of decisions made with the scores and reliability of measurements or scores themselves. The two dimensions are then used to generate four possible applied situations: (1) continuous variable, binary decisions (e.g., mastery-nonmastery); (2) binary variable, binary decisions; (3) continuous variable, test scores; and (4) binary variable, test scores. The various approaches to reliability are discussed in response to each situation.

Situation 1

Almost the entire review is devoted to Situation 1. It is clearly the most common situation encountered in practice. Traub and Rowley employ a loss function structure to assess the specific indices consonant with that situation: (1) threshold loss; (2) linear loss; and (3) squared-error loss. Of course, the first two loss functions correspond to van der Linden's. As noted in the Standard Setting section, the loss function provides a meaningful structure for selecting statistics consistent with the decision application. Since the choice of a cutoff-score method should precede the choice of a reliability index (see Berk, in press, for rationale), a consideration of the appropriate loss function is one way to assure compatibility between the cutoff score (and evidence of decision accuracy) and reliability (and evidence of decision consistency), i.e., the loss function should be the same for both, where possible. For example, if an optimal cutoff score is selected where the losses associated with the decision errors are assumed to be equal, then a threshold loss function agreement index such as p_o or κ should be chosen as the measure of reliability.

Traub and Rowley present an extensive survey of the one- and two-administration approaches for estimating the threshold loss function indices p_o and κ. Interested readers should compare their conclusions with those of Berk (in press) and Subkoviak (1980).

The evaluation of squared-error loss function approaches focuses on the work of van der Linden and Mellenbergh (1978), Livingston (1972), and Brennan (1980a). Particular emphasis is given to Brennan's (1980a) generalizability theoretic approach. His articles on that topic should be consulted for an in-depth understanding of the theory and another possible framework for studying threshold loss and squared-error loss function indices (Brennan, 1980b).

One distinguishing feature of Traub and Rowley's review is their examination of linear loss function indices. This loss function has not been covered by extant reviews. Two indices are described: Livingston and Wingersky's (1979) index of decision-making efficiency and van der Linden and Mellenbergh's (1978) index incorporating three types of risk.

Situations 2, 3, and 4

The remaining three situations receive relatively limited attention, due in part to the unavailability of appropriate statistics. Situations 2 and 4, which are based on the state model, are especially difficult to address. Situation 3 is more manageable, since methods derived from classical test theory and generalizability theory can be used to provide domain score estimates. Specific statistics for this situation have been proffered recently in Berk (in press).

Implications for Practice

One of the overall conclusions Traub and Rowley draw from their review is that the approaches have little practicality for classroom teachers and program evaluators. Although further research at these levels of decision-making is indisputably needed, there are a few available statistics that are easy to compute, to understand, and to interpret. Hambleton and Novick's (1973) two-administration estimate of p_o based on a test-retest design is one example (see Berk, in press, for details). For program evaluation decisions, errors associated with average domain score estimates derived from matrix sampling designs have potential.

Conclusions

The foregoing assessment of the articles in this issue indicates that each paper contributes to the literature on criterion-referenced measurement by devising a framework for future investigations on the topic and by synthesizing proposed methods and related technical issues. In some instances, new methods or extensions of existing methods are also discussed.

The analyses of the articles in this issue revealed a single common thread running through all of them: The selection and estimation of the technical properties are governed primarily by the interpretation of the test scores and the decisions that are made with them. In other words, the types of decision-making will dictate the length of the test, the types of validity evidence, the location of the cutoff score, and the types of reliability evidence.

As mentioned at the outset of this paper, a considerable amount of research has accumulated over the past decade. It has consistently increased in quantity and improved in quality. Despite the unduly pessimistic state of the art conveyed by some of the authors, the methodological advances contained in this issue, as well as elsewhere, are very encouraging.

References

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

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

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

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

Berk, R. A. *Domain-referenced versus mastery conceptualization of criterion-referenced measurement: A clarification*. Paper presented at the annual meeting of the American Educational Research Association, Boston, April 1980. (b)

Berk, R. A. Practical guidelines for determining the length of objectives-based, criterion-referenced tests. *Educational Technology*, 1980, *20*, 36-41. (c)

Berk, R. A. A consumers' guide to criterion-referenced test reliability. *Journal of Educational Measurement*, in press.

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

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980. (a)

Brennan, R. L. *Parametric expressions for some single-administration indexes of dependability for domain-referenced interpretations*. Discussant comments for the paper session "Reliability in criterion-referenced testing" presented at the annual meeting of the National Council on Measurement in Education, Boston, April 1980. (b)

Cronbach, L. J. Validity on parole: How can we go straight? In W. B. Schrader (Ed.), *New directions for testing and measurement. Measuring achievement: Progress over a decade* (No. 5). San Francisco, CA: Jossey-Bass, 1980.

Ebel, R. L. *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice-Hall, 1979.

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

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

Graham, D., & Bergquist, C. *An examination of criterion-referenced test characteristics in relation to assumptions about the nature of achievement variables*. Paper presented at the annual meeting of the American Educational Research Association, Washington, DC, March 1975.

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

Guion, R. M. "Content validity" in moderation. *Personnel Psychology*, 1978, *31*, 205-213.

Guion, R. M. On trinitarian doctrines of validity. *Professional Psychology*, in press.

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

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*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 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.

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

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

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

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

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

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 measurement*. Washington, DC: 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, DC: 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, NJ: Center

for Occupational and Professional Assessment, Educational Testing Service, 1975.

Livingston, S. A. Choosing minimum passing scores by stochastic approximation techniques. *Educational and Psychological Measurement*, in press.

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement*, 1979, *16*, 247-260.

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

Mellenbergh, G. J., Koppelaar, H., & van der Linden, W. J. Dichotomous decisions based on dichotomously scored items: A case study. *Statistica Neerlandica*, 1977, *31*, 161-169.

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

Messick, S. *Test validity and the ethics of assessment*. Paper presented at the annual meeting of the American Psychological Association, New York, September 1979.

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

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

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

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), *Problems in criterion-referenced measurement* (CSE Monograph Series in Evaluation, No. 3). 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.

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

Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. Baltimore, MD: The Johns Hopkins University Press, 1980.

Roid, G. H., & Haladyna, T. M. The emergence of an item-writing technology. *Review of Educational Research*, 1980, *50*, 293-314.

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

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

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

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

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

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

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

van der Linden, W. J., & Mellenbergh, G. J. Coefficients for tests from a decision theoretic point of view. *Applied Psychological Measurement*, 1978, *2*, 119-134.

Werts, C. E., Linn, R. L., & Jöreskog, K. A congeneric model for Platonic true scores. *Educational and Psychological Measurement*, 1973, *33*, 311-318.

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

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

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

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

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

Author's Address

Ronald A. Berk, Division of Education, The Johns Hopkins University, 105 Whitehead Hall, Baltimore, MD 21218.