COMMENTS: MEASUREMENT ISSUES RELATED TO PERFORMANCE STANDARDS*

Tom Haladyna/Oregon State System of Higher Education

My first reaction to the problem of standard-setting in competency-based education is that it really has nothing to do with measurement. In most systematic instruction settings, a test score is obtained from student performance which represents the student's true level of performance with respect to a well-defined achievement domain (in some cases this is an objective). A passing standard is imposed on the scale for the purpose of determining who has passed and who has failed. Where to set this passing standard and noting the consequences of the standards are direct concerns in this symposium. While many measurement problems pertain to these concerns, the actual setting of a standard and the decision making that occurs appears more rightfully in the area of evaluation of test results.

The setting of standards in competency-based education is, indeed, a venture into murky waters. While there is a long history of standard-setting in education, there is not a well-known science of standard-setting (Stoker, 1976). Thus, the relevancy of this symposium is well-established, and the papers presented here represent a beginning concern. Lorrie Shepard's paper is a thorough and excellent analysis of the conditions (definitions and distinctions) surrounding the use of standards. The recommendations she offers make good sense. To reiterate some of her points:

1. Standard-setting is essentially judgmental—albeit eventually empirically confirmed.
2. All judgment is essentially normative.
3. Instructional improvement will be one criterion we will need to consider in competency-based education and standard settings.

Each of these points reveals the need for empirical evidence to augment our rational judgment. Shepard's paper does not tell how we actually set standards, and in this vein, there is still much work to do. However, she has provided us with some of the pitfalls which should be avoided in setting these standards.

Jaeger's paper reveals a number of important substantive problems in the area of measurement, specifically relating to errors of decision making and validity. It is important to note that these problems have an added significance when used in the context of competency-based education, particularly in the area of reliability where decision making occurs and errors of measurement play an important role.

The balance of these comments will be addressed to (1) the context for which standard-setting occurs (2) distal and proximal standard-setting models (3) decision making and Type 1 and Type 2 errors (4) the ultimate dilemma facing anyone who uses criterion-referenced or domain-referenced achievement tests in competency-based instruction.

1. For a number of years, there has been a plethora of approaches to instruction where standards are used to assign all students to one of two categories, pass or fail. These approaches include competency-based education, mastery learning, individually-prescribed instruction, computer-based instruction, computer-assisted instruction, the Keller plan (Personalized Student Instruction), Carroll's model for school learning, and Program for Learning According to Needs (PLAN). While this list is not exhaustive, it does represent the wide range of innovative types of instruction that requires the classroom management of students in a nontraditional way. A host of problems exist with the setting of standards and the determination of who should pass and who should fail in this instructional context, and some of these problems are addressed in this symposium. We presently have no known system for setting standards, and our ability to ascertain the accuracy with which students are classified in pass or fail categories is only beginning to be studied and understood.

2. The proximal and distal standard-setting models described by Jaeger are suspiciously like traditional concepts of content and predictive validities. In the case of the former, the concern is for clear specification of the content domain and for random sampling of items representing that content domain. This is the essence of domain-referenced testing as described by Hively (1974) and Millman (1974a). While some debate exists over how different this domain-sampling approach is (Haladyna, 1975), random sampling of items from a well-defined domain is a recommended procedure in traditional test theory (Nunnally, 1967) as well as the new domain-referenced testing approach. It is the manner in which items are created, via an item-writing, generating algorithm, that distinguishes the old domain-sampling approach from the new. Both approaches appear to satisfy the criteria for good content validity. The proximal standard-setting model and threats to validity described by Jaeger have much to do with the problems encountered with the use of content validity in testing.

In the distal standard-setting model, the concern is some ultimate criterion. As in the case with content validity, there is much with the distal model that is reflected in the traditional concept of predictive validity. The threats to validity described by Jaeger appear salient for a traditional application of predictive validity as well as in the context of competency-based instruction. Additionally, there appears much work ahead in determining where to set this standard and what the consequences of the standard will be. There is a strong need for empirical studies where sequential instruction occurs and where passing standards are manipulated to ascertain the long-range effects of passing standards in terms of both the proximal and distal standard-setting models.

3. In both Jaeger's and Shepard's papers, mention is made of the problem of correctly classifying students into pass or

*Presented in a symposium on Measurement Issues Related to Performance Standards in Competency-Based Education, National Council on Measurement in Education, San Francisco, April, 1976.

fail categories once a standard is set. The problem appears to revolve around error of measurement for a test score of a student. Since true scores cannot be known, only estimated, we employ an observed score and some statistical theory where errors of measurement are used to compute a standard error to assist us in decision making. A confidence interval is established around the passing standard and students can be assigned to these conditions: (a) pass (b) uncertain (c) fail. This problem is well-described by Hambleton and Novick (1973) in their work, and it has been studied by Millman (1974b) and Haladyna. While comments in Jaeger's paper are addressed to the binomial and beta-binomial models, as well as the decision-theoretic work of Hambleton and Novick (1973), other models which have utility in decision making were not discussed. One of these is the Rasch model (Wright and Panchepakesan, 1968), one of a class of latent trait models. Another model is the classical one, which was empirically compared to the binomial model in a study by Haladyna. Neither model was found to be particularly useful for decision making in the instructional context. Each of these approaches offers potential solutions to the problems of decision making, and the Type 1 and Type 2 errors misclassify true passing students as failing and true failing students as passing. Shepard states that these two types of errors can be controlled by manipulating the passing standard. Setting it higher minimized one error at the expense of the other. Emrick (1971) offered a statistical solution to this but it remains virtually untried and untested. It should be apparent that considerable efforts are needed in this area in the future.

Despite our lack of methodology in minimizing errors or misclassification, there are a great many examples of minimizing losses due to errors of measurement. Most professional schools employ criteria to decrease false positives at the expense of false negatives. That is, only the most potentially capable are admitted even though a great many who would have been successful were never given a chance to attend. Empirical data which attests to the superiority of any statistical model is lacking, and this state of affairs represents one of the most urgent areas of research in educational testing.

4. The proximal standard-setting model and the implicit approach to measurement represent what is currently called "domain-referenced testing." Herein, lies an issue of major importance that neither paper has addressed. Operationally defining a content domain, through the use of instructional objectives of item-writing rules, represents a form of educational behaviorism where the resulting item pool forms the basis for defining the concerns of concern. An operational definition of vocabulary might call for the recognition of a correct definition when given four alternative definitions. The percentage of words any student can define on a random sample of tasks represents the behavior which occurs. How we interpret that behavior depends on our approach: the operational definition of vocabulary versus the "construct-referenced." His eloquent plea reveals the schism which appears to differentiate the domain-referenced test movement in achievement testing from a more traditional approach. The operational definition of a trait through item generation rules leads to a strong case for Jaeger's proximal standard-setting model (or content validity), while the construct approach advocated by Messick, among others, appears to be an alternative to the content validity proximal standard-setting model. In the construct approach, descriptions of the trait of concern and empirical verification through testing occurs. Thus, the initial concern with standard-setting may be how we intend to describe the trait we are measuring: as an operational definition or as a hypothetical construct. The choice will lead to (a) a different set of assumptions in item and test construction and (b) a choice of standard-setting models as well as (c) the accompanying threats to validity that Jaeger describes.

References

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

Haladyna, T. *An analysis of two procedures for decision making when using domain-referenced tests* (Paper presented at the Annual meeting of the National Council for Measurement in Education.) Washington, D.C.: 1975.

Haladyna, T. The paradox of criterion-referenced measurement. (Paper presented at the Annual meeting of the National Council on Measurement in Education.) San Francisco: 1976.

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.

Hively, W. Introduction to domain-referenced testing. *Educational Technology*, 1974, *14*, 5-10.

Millman, J. Sampling plans for domain-referenced tests. *Educational Technology*, 1974, *14*, 17-21(A).

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

Nunnally, J. C. *Psychometric theory.* New York: McGraw-Hill, 1967.

Stoker, H. *Models for performance standards: Background and identification.* (Paper presented at the National Council on Measurement in Education.) April, 1976.

Wright, B. D., & Panchepakesan, N. A procedure for sample-free item analysis. *Educational and Psychological Measurement*, 1969, *29*, 23-49.