SETTING STANDARDS AND LIVING WITH THEM*

Loretta A. Shepard/University of Colorado

The avowed virtue of competence-based education is that instructional efforts are accompanied by unambiguous criteria for learning proficiency. Competence-based assessment requires standards to distinguish acceptable and unacceptable performances. The purpose of this paper is to consider practical constraints on standard-setting and to make recommendations. This paper is written to complement Dick Jaeger's paper which treats the psychometric issues involved. It begins where Jaeger's leaves off with interim suggestions in the absence of empirically validated procedures.

In the original proposal for this symposium, it was implied that there are different models for setting standards and that each has certain attributes that makes it the method of choice in particular situations. Instead, it seems that the distinctions are fuzzy and that a single composite plan for setting standards is called for.

Definitions and Distinctions

Before proceeding with practical suggestions for setting standards, some definition difficulties that cloud the issue of setting performance criteria ought to be cleared up. There are some terms that are appropriately synonyms for "standards" and others that are not. There are some concepts that seem perpetually linked with competence criteria that are really unrelated notions.

For example, performance standards are sometimes thought to be the same as normative cut-off scores. That is, criterion scores are set so that a prespecified number or percent of candidates pass. Millman (1973) pointed out that whether an individual succeeds by this type of criterion depends in part on the competence of others taking the test. While this procedure is appropriate when "the number of people to be certified is fixed (Millman, 1973, p. 206)," such a criterion lacks any intrinsic standard. It ensures that the most able candidates will be selected but does not guarantee the skill level of any of them. Correct use of the term "standard" in the context of this paper requires an absolute rather than a relative judgment of performance. With this stipulation understood, "criterion score" may be used interchangeably with "performance standard."

The purpose of standards is to identify mastery of whatever content is being assessed. Performance criteria or standards are a necessary feature of mastery learning of the types advocated by Carroll (1971) and Bloom (1971). The standard is the level of proficiency that each student is expected to attain. One of the difficulties with mastery learning is its basic tenet that given enough time nearly any student can learn what we wish to teach him. Depending on the difficulty of the material, this fundamental assumption may or may not be met. This issue should not, however, be confused with the equally perplexing problem which we are facing of how to determine the level of performance that constitutes mastery. This distinction is very important for subsequent discussion, since when these issues are mixed, standards must be set as minimums so that all students can attain them.

Another question which is often confused with the problem of setting standards is the method of test construction. The terms "objective-referenced testing," "criterion-referenced testing," and "domain-referenced testing" are often used as if they meant the same thing. The confusion of these terms in common parlance will probably never be eliminated, but the distinctions that have been obscured ought to be redrawn. Criterion-referenced testing was meant to describe what students know or have accomplished rather than their relative standing in a group (Glaser, 1963). The name implies that a passing score or criterion is established to determine which students have succeeded. Criterion-referenced testing was accompanied by a more careful specification of the content universe to be assessed. This systematic development of test content is a separate feature and is better characterized by the term objective-referenced testing. Many so-called CRTs are objective-referenced or content-referenced but do not have standards of performance. The name implies that standards or criteria for judging acceptable performance are a part of the testing process. Instead, it may be that the test items have been developed to correspond to a detailed schema of the content, but that no standard has been agreed upon. Certainly setting criteria is partially dependent on properly referencing items to objectives. How meaningful would criteria be if the content universe were not carefully identified and appropriately represented by the items? However, it is important to realize that the methodological insights about how to develop objectives and subobjectives do not contain the solution to the problem of how to set criteria.

Domain-referenced testing, described by Millman (1972) and Hively (Maxwell, 1971), is an even more precise method of test content specifications. The item-forms which characterize domain-referenced testing are intended to be explicit enough to determine the difficulty level of all items generated from a particular form; in this sense, criterion levels are a part of each item form. Nevertheless, when performance is aggregated across items, there is still the problem of what combination of success and failure denotes acceptable performance.

Finally, the concept of competence-based assessment seems to be tangled with notions of learning hierarchies, prerequisite skills, minimum competence, basic skills, and the more popular survival skills. Gagné(1965) is the principal exponent of a seductively simple theory of education. Every learning task is comprised of elements. Learning is facilitated

*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.

when each task is analyzed into its requisite subtasks so that the learner may proceed sequentially through the implicit hierarchy. The theory is intriguing and has some empirical examples (Gagné and Paradise, 1961). It seems to work best in math where, just as in the case of domain-referenced testing, we seem better able to analyze goals into specific subskills. Nevertheless, there are counter-examples where the theory does not work so well. Learning hierarchies are only absolute when the order of skill acquisition is immutable for all individuals. For most learning tasks, the hierarchies are not fixed or absolute. There are always anomalous individuals who have mastered an apparently complex skill without mastering supposed prerequisites: like playing the piano without learning to read music. In math, it may be that knowing how to add two-digit numbers is an absolute prerequisite for adding three-digit numbers; but it may also be that addition facts need not be mastered before learning to multiply. Observations like these pose a conundrum for standard-setters in competence-based assessment. If the purpose of setting criterion levels on the earlier tasks is to ensure success on later endeavors, the counter-examples are very troubling indeed. What if a child has not mastered a prerequisite skill and is assigned additional practice? If the prerequisite skill has no merit in itself but is only being studied as the means to an end, then the extra practice may be a waste of time. If some children can master the final objective without the prerequisites, then we are less certain of our standards for mastery.

The learning hierarchies model is not very practical for setting standards. It is useful in only very circumscribe curricular areas. When the ultimate tasks are broad such as "being a good citizen" or "being successful in life," it is very unlikely that valid hierarchies exist. Standard-setters would be better able to fulfill their role if they could only develop the methodologies to uncover them. "Standards are choices, not essences."

Setting Standards Requires Judgment

If in all the instances that we care about there is no external truth, no set of minimum competencies that are necessary and sufficient for life success, then all standard-setting is judgmental. Our empirical methods may facilitate judgment making, but they cannot be used to ferret out standards as if they existed independently of human opinions and values.

So-called empirical methods for standard-setting usually involve statistical means for maximizing the prediction of some ultimate task from the competency assessment. For example, a cut-off score may be set on a medical certification exam to maximize the agreement between success and failures on the exam and success and failures in the profession. Nevertheless, judgment is integrally involved in the definition of the criterion performance. Although the cut-off score is apparently determined statistically, judges actually decide it depending on whether professional success is defined as annual income or patient longevity.

Standards Should Not Be the Lowest Common Denominator

Those who believe that standards exist external to the judgement of experts are likely to engage in searches for the essential skills. They might, for example, interview plumbers and shop clerks to try and locate which competencies are minimal, and which are held by all employed adults. Such searches will be informative and may be helpful to judges in the same way that normative data are helpful, but the searchers will not turn up any universals. We might as well consider right now what we will do with some very successful plumbers who cannot read or street sweepers in San Francisco who make more money than university professors but cannot handle simple fractions. The search for absolutes leads to absurd reductionism. Perhaps we should study the mentally retarded. What skills are absolutely essential to be able to ride the bus to and from a sheltered workshop? Suppose we thus identify some basic skills that are completely rudimentary. How would these lowest-common-denominator standards serve as meaningful criteria for prospective carpenters and TV repairmen?

Standard-setters ought to begin their task recognizing that counter examples will exist. If reading comprehension is deemed important, the standard ought to be set despite the existence of successful businessmen who cannot pass the test. Lowering the standards until everyone can pass them completely defeats their purpose.

As I remarked earlier, Gagné's learning hierarchies have had a pervasive effect. Belief in his tenets has sent assessors and standard-setters in search of criterion scores as if they were external truths. In addition, an overly simplified view of learning has been adopted. Although any student may demonstrate mastery with lots of coaching and prompting, what disciples of mastery learning don't consider is how unequal students may look on retention and transfer. Instead of recognizing that students may reach different levels of mastery, adherents are likely to seek more and more fundamental tasks so that evidence of success and failure will be black and white instead of gray. Those who believe that standards ought to be the lowest common denominator have failed to recognize that the behaviors we wish to predict, and the life skills we wish to ensure are not singular but are arrayed along a continuum. The competencies that mentally retarded individuals ought to have to lead a full life are not the same as those necessary for other subgroups to lead rich and productive lives.

Judgmental Standard-Setting is Subjective but not Capricious

The theme of the preceding sections is that setting standards is judgmental. The validity of the standards will depend on the wisdom of the standard-setters. No standard inheres in nature for them to discover. If nothing else, standards will be established more sensibly and with less argument and grief if experts understand their limitations and their role. There are no magic ratios, 80 percent, that can be arbitrarily assigned regardless of skill level or content domain.

This recommendation seems so obvious, but it is the single most frequent error in criterion-referenced assessment. In the evaluation of the Michigan assessment results, House, Rivers, and Stufflebeam (1974) criticized the use of the term "minimal" for objectives in math and reading that were not attained by any of the districts in the state. Since some of these districts routinely scored very high on traditional achievement tests, the fault seemed to be with the objectives rather than with the educational system. In retrospect, it appears that not much thought was given to the use of the word "minimal"; either the assessment should have been conducted with the existing objectives but without the minimal designation or some judgment should have been made about which objectives were truly minimal. Obviously, identifying minimums is not easy but experts should ask themselves which objectives are so essential that districts can be considered negligent when the necessary percentage of students does not attain the objective. If experts had asked themselves this question in the Michigan case, they might have forestalled the embarrassing results.

The Harshness of Standards Ought To Be Modified Depending Upon the Seriousness of Two Types of Errors

If standards were set wisely, it is assumed that most decisions that result from the assessment process would be correct. Individuals who were incompetent would fail the test and those who were competent would pass. Mistakes will occur, however, because of measurement error or improper inferences to the real-life skills for which the test is a proxy. These errors are of two types, viz, false positives and false negatives. False positives are those who pass but do not have the necessary mastery; false negatives are those who fail but actually possess the requisite skills.

The seriousness of these two kinds of mistakes will vary ·with the situation. When individuals are certified to practice in various professions as doctors, lawyers, teachers, the cost to society is much greater for the false positives than for the false negatives. In these cases, relatively stringent standards ought to be set to protect the public against unqualified practitioners. The cost to individuals who are thereby unfairly failed is outweighed by the public good. In many instructional settings, however, the reverse is true. In instances when learning hierarchies are valid, strict criteria prevent an individual from encountering material that is too difficult until the prerequisite skills have been mastered. But when learning hierarchies are not accurate, it may be that the most serious costs occur with the false negatives. Individuals who are forced to drill on material they have already mastered may become frustrated or bored. Such a circumstance is particularly untenable when incomplete mastery of early skills does not actually prevent mastery of more complex skills.

Expert judges will have to consider the relative costs of the two types of error and adjust the standards to protect against the most serious mistakes. The current hue and cry for competence-based high school diplomas is motivated entirely by one type of error—those who graduate without seemingly essential competencies. Before these various plans and laws are enacted, someone ought to raise the issue of the other type of error. Are there societal or personal benefits that would warrant keeping some nineteen-year-olds in school? Will we be better off if diplomas are withheld from those who do not meet certain standards? This question has not been raised because in the current rush for minimum standards, most believe that the existence of standards will increase the skills of high school graduates. They probably will. But if standards are to be meaningful, there are bound to be individuals who cannot meet them. Some will take a longer time, others will never make it. Only trivially low standards will be passed by everyone.

Recommendations for Setting Standards

1. Setting standards ought to be an iterative process.

Subsequent recommendations suggest ways to enlighten the judgment-rendering process; but because it relies fundamentally on human wisdom, there are bound to be errors. Thus, criteria should not be fixed for all time. Panels of experts ought to be reconvened when the results are in. They should examine the results, looking especially for examples of false positives and false negatives. The purpose of their review should not be to grant exceptional certifications of mastery, but to find systematic errors that suggest a change in criteria is needed. They will be conducting an informal validation, looking for agreement between the results of criterion assessment and additional data and judgments. In the Michigan example, experts might decide to adjust the standards if all school districts failed unless they believed that all districts were indeed failing.

2. The normative or experiential basis of judgments ought to be a formal part of the standard-setting process.

Expert judges ought to be provided with normative data in their deliberations. Instead of relying on their experience, which may have been with unusual students or professionals, experts ought to have access to representative norms. If new measures are being developed along with the standards, then normative information could be obtained from similar existing measures.

Of course, the norms are not automatically the standards. Experts still have to decide what "ought" to be; but they can establish more reasonable expectations if they know what current performance is than if they deliberate in a vacuum.

3. The most reasonable standard is "Improvement."

In some areas, it may be possible to establish minimal competencies that are absolute and consensual, where everyone agrees that a skill is essential and mastery is clearly distinguishable from non-mastery. For example, all physicians ought to be able to recognize and treat shock. However, in most areas of education and training, these absolutes do not exist. In these instances, "improvement" may be the only defensible standard. Having looked at data on current performance for a state or a district, experts ought to identify those objectives for which performance has been satisfactory and those which need improvement. There is still the difficulty of deciding how much improvement is needed, of actually setting the standard. But, designating areas where improvement is needed is an important first step. If the standard-setting process is really iterative, the standard will evolve over repeated assessments. It will be "set" when the experts decide for a particular objective that the performance level is acceptable and ought to be "maintained" and that attention for improvement be focused elsewhere.

National Assessment was intended to provide information about what students and adults in the United States know and can do. It was the creation of those who advocated criterion-referenced testing, but no criteria were specified. There was very little excitement over the result of the first assessment rounds because most ?udiences were unable to judge for themselves whether the reported percentages were good or bad. The standards were missing. Now there is a great deal of excitement because the science and writing results have shown decline from the first to the second assessment. After verifying that the science decline could not be explained by a change in objectives or a curricular shift from the physical to the biological sciences, it was not difficult to assign meaning to the results. A decline is bad.

4. Allow for differences of opinion by involving various audiences in standard-setting.

Jaeger (1976) points out in his paper that the validity of standards depends on the sampling of judges. This is true whether the judges establish the standards directly by assigning criterion scores or indirectly by defining success in a criterion group.

It is a fairly obvious and hackneyed recommendation that all relevant audiences ought to be represented in the setting of standards. It is worth mentioning, however, since in some instances where legislators or militant parents have insisted on the setting of standards, they have left the actual judgments to professional educators. This is perplexing since there is no reason to believe that the educators will reflect the values held by parents and taxpayers. The impetus for competence-based assessment in the first place may have been differences in values and priorities. If an organization of employers has lobbied for competence-based diplomas because high school graduates cannot make change, their values ought to be represented in the setting of standards. It is only postponing the clash if educators substitute standards in the affective domain.

Representation from groups who disagree may be the most straightforward way of dealing with differences in values. Brickell (1974) wrote a very amusing paper about the external factors that influence evaluation. The evaluator was buffeted by enormous political pressures in all settings but one. The fairest and most politically neutral situation was that in which all sides were represented.

Once experts have been identified from relevant audiences, they should not be tossed together to reach consensus on standards. Obviously, consensual standards would be easier to implement. But, personality dynamics in a particular group of judges could create phoney consensus. The best protection against artificial consensus is to convene more than one group of expert judges and have them meet separately. If they arrive at the same standards or nearly the same standards independently, their agreement will be observable and will be much more dependable than if they had met jointly.

If groups of experts do not agree on what standards should be, alternatives still exist for arriving at standards. If the purpose of the standards is to allow evaluation of school systems, it will be possible to report assessment results in light of more than one criterion. This is similar to the apples and oranges problem encountered in comparative evaluation. One program may look better by one set of criteria; the other may shine when judged by alternative criteria. The only fair and comprehensive way to judge them is to apply all of the criteria to each.

If the standards are to be used to make decisions about individuals, then one set of criteria is needed. The criterion should either be the most stringent or the most lenient of those proposed depending on which type of error is more serious in that situation. If false negatives would be more costly to individuals and society, then the standards should be lower than in the instances when false positives must be screened out.

5. Final caution: Focusing on minimums may limit the height of educational attainment.

Competence-based education certainly does not imply that learning will stop after the basics have been mastered. It only implies that learning is sequential and that mastery of early tasks will be accomplished before passing on to the next. However, the current popularity of this concept with legislators and the public carries with it the term minimum. Certainly for individuals, there cannot be much quarrel with learning the basics before attempting more complex tasks. But for entire school systems, there may be some unforeseen consequences if the exclusive focus becomes the attainment of minimums. Ultimately, we will have to face the choice of whether to teach the last student in a school to work with decimals instead of teaching a classmate the beginnings of biochemistry. Once competence-based assessments are fully installed, parents and taxpayers ought to ask for assurances that minimums are not being attained at the expense of excellence.

Anti-testers are fond of the argument that testing minimums will limit educational growth. Rather than being a plea for less testing, however, this caution warns of the need for evidence at both ends of the performance continuum.

For individuals as well, there is the concern that required minimums will become the maximums, that students will stop trying when minimums are attained. Part of the solution to this problem is the comment provided earlier that the life skills that are ensured by competence-based assessment are arrayed along a continuum. A full and productive life for most individuals requires more than the minimums. Perhaps assessment and rewards for accomplishment beyond the minimums are means for increasing growth towards these goals.

References

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

Bloom, B. Mastery learning and its implications for curriculum development. In E. W. Eisner (Ed.), *Confronting curriculum reform.* Boston: Little, Brown, 1971.

Brickell, H. The influence of external political factors on the role and methodology of evaluation. (Paper presented at the meeting of the American Educational Research Association.) Chicago: 1974.

Carroll, J. Problems of measurement related to the concept of learning for mastery. In J. H. Block (Ed.), *Mastery learning: Theory and practice.* New York: Holt, Rinehart and Winston, 1971.

Gagné, R. *The conditions of learning.* New York: Holt, Rinehart and Winston, 1965.

Gagné, R., and Paradise, N. Abilities and learning sets in knowledge acquisition. *Psychological Monographs,* 1961, *75* (14).

Glaser, R. Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 1963, *18,* 519-521.

House, E., Rivers, W., and Stufflebeam, D. *An assessment of the Michigan accountability system.* The Michigan Education Association and the National Education Association, March, 1974.

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

Maxwell, G., Hively, W., Lundin, S., Rabehl, G., and Sension, D. Curriculum evaluation in the MINNEMAST project: A case study in domain-referenced testing. Minneapolis: University of Minnesota, 1971.

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, April, 1972.

Millman, J. Domain-referenced measures. *Review of Education Research,* 1973, *43,* 205-216.