DISCUSSANT COMMENTS: SETTING PERFORMANCE STANDARDS BASED ON LIMITED RESEARCH*

Larry E. Conaway/Research Triangle Institute

As discussant I want to state my appreciation to the chairperson and participants who were extremely thorough and prompt in preparing for the session. The timely delivery of intermediate materials and final papers was very valuable in preparing discussant comments.

My comments stem from two sources. First, in working with educational assessment projects in various states and local districts, I have been involved in actually establishing desired and minimal performance standards for groups of students across cognitive objectives, using consensus procedures. These performance standards have been used to identify program strengths and weaknesses. Second, having received the papers prior to the session and having been both intellectually puzzled and intrigued by many of the ideas, I backtraced the authors through many of their references and read some additional literature in the areas of criterion-referenced testing and performance-based teacher education.

I generally agree with the basic points the authors made: standard-setting is not new; standard-setting always involves human judgment; there are a number of threats to the validity of standards, and there has not been adequate research in this area; standard-setters cannot and will not wait for definitive empirical research so interim guidelines for standard-setting will emerge, based upon available research and practical constraints; and in light of an incomplete research base, those who set and use standards should be careful that they do not do more harm than good through overuse or abuse of standards. Therefore, my comments are not generally critical. Instead, my comments elaborate upon some points that I believe should be emphasized and discuss Shepard's recommendations for setting standards from the perspective of my experience in establishing standards.

Establishing the Standard Versus Estimating the Score

Jaeger provided threats to validity from two separate and distinct sources: (1) establishing the standard and (2) estimating the achieved score. Study of the related literature shows that most of the rigorous literature relates to estimating the achieved score and its associated variance rather than to establishing the standard.

Airasian and Madaus (1972) stated that the area of setting standards was the area of criterion-referenced measurement in most need of research, and Quirk (1974) discussed a similar lack of research in the context of performance-based teacher education. Research in establishing the standard is as important as research in estimating the achieved score; certainly nothing of definite utility will be provided from the research until there is progress in both areas. An additional benefit from research in establishing the standard is that work in this area involves educators and others in tying together testing and instruction.

In summary, there has been very little research in the area of establishing standards through human judgment, and there is a lack of methodological literature in establishing standards and their associated variances. Jaeger was correct in emphasizing the need "for empirical investigation involving human standard-setters in real or simulated judgmental situations, using real performance data and real descriptions of task domains."

Judgmental Standards Must Take into Account Item Content and Difficulty Level

Millman (1973) stated that it is difficult to defend the frequent practice of employing a particular passing score only on the grounds of tradition. Quirk (1974, p. 317) was very specific in his discussion of fixed cutoff scores in performance-based teacher education:

> While they (fixed cutoff scores) sound semiscientific, they do not possess much substantive value. The percentage of items related to an objective which a candidate answers correctly is a function not only of the content of the items but also of the difficulty of the items. An estimate of the difficulty of the items can be obtained either from a logical judgment based on a study of the specific items or from empirical item-analysis data.

After discussing the state of research in standard-setting, the state-of-the-art in domain-referenced testing, and the problems associated with learning hierarchies, Airasian and Madaus (1972) concluded that teachers will have to establish their own standards using expert opinion, experience, face validity of items, and group consensus.

Personal experience in standard-setting proves this literature to be relevant and true. Valid and credible standards depend upon the use of human standard-setters who take their roles seriously and who base their standards upon their experiences with specific performance tasks. Shepard's warning should be taken seriously—"The validity of the standards will depend on the wisdom of the standard-setters." Unless standards are established by some defensible methodology which involves careful human judgment, they will not serve their intended purposes nor will they stand up against the careful scrutiny of those who doubt their validity.

---

*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.

Recommendations for Setting Standards

My comments about Shepard's recommendations for setting standards are based mainly upon personal experience in establishing desired and minimal standards for cognitive objectives by setting desired and minimal performance levels for each test item, using consensus procedures. Before discussing individual recommendations, there is a point to be made. Shepard's model is very general. If this general model proves to have utility in the area of standard-setting, Shepard and others should expand the model by discussing each relevant recommendation specifically for such standard-setting categories as individual student standards, group standards (e.g., classroom, district, state), standards established without external political pressure, and standards established because of external political pressure. Setting standards within categories such as the above requires differing specific recommendations, and the more specific recommendations would have greater value to those involved in the various situations.

Setting Standards Ought To Be an Interactive Process. This is certainly correct since standards involve human judgment rather than being inherently true. After we have teachers establish minimal and desired performance levels for groups of students, we have the group look at their standards in comparison to actual student results and any normative data that are available. Recommendations for change are based upon results from all sources. Teachers generally rely heavily upon the standards they set, but they also reconsider some standards after seeing other results.

The Normative or Experiential Basis of Judgment Ought To Be a Formal Part of the Standard-Setting Process. As long as human judgment is involved, the experiential basis will be a factor; however, there are probably times when the normative basis should be controlled in terms of the time for formal consideration. One reason we began to establish a priori levels of desired and minimal performance was because teachers could not meaningfully determine "what ought to be" when actual student results and normative results were available. Teachers have enough of an experiential base to arrive at meaningful standards using consensus procedures, and these standards then become a useful set of results to be used in conjunction with other results in recommending educational changes. It seems very likely that normative results ought to be a formal part of the standard-setting process for lay people who do not have a broad experiential basis for judgment.

The Most Reasonable Standard Is Improvement. Shepard makes an excellent point in stating that an important first step is designating areas where improvement is needed. In our work, attention has been focused upon instructional areas designated as needing improvement, based upon minimal and desired performance standards in conjunction with other results. Other instructional areas were judged to be satisfactory using the same procedures for establishing standards and analyzing results. The fact that many of these areas in need of improvement were identified using standards set by teachers rather than normative results has been very helpful in gaining the cooperation of the teachers in recommending and implementing changes.

The need for improvement should not be based solely upon normative results. In the National Assessment example it is difficult to tell if Shepard is reporting that others are saying a decline is bad or if she is advocating that position. It is important that some formal judgmental consideration be given to the standard of science performance, in conjunction with the change results, before the conclusion is made that a decline in science performance is bad.

Allow for Differences of Opinion by Involving Various Audiences in Standard-Setting. The political situation can have great impact in this area. When there is no external pressure, those who determine the composition of standard-setting groups can attempt to construct groups which can and will establish valid standards. Even in these situations, they should be aware that active involvement by parents and politicians may provide more credible standards and prevent future political problems.

When there is external pressure the standard-setting procedures may be dictated by non-educators or the standards may be established completely outside the normal educational decision-making arena. Assuming control of standard-setting procedures by educators, Shepard made some valid observations. All audiences should be represented, and their conflicting viewpoints should be aired and considered in reaching consensus. Shepard's suggestion that more than one group be convened (each group representing all audiences) is very practical. When the groups disagree it would probably be beneficial both to report more than one criterion level and to convene another group, composed of representatives from each original group, to attempt to reach a final consensus.

One further idea might be worthy of consideration. It may be possible to involve audiences in establishing standards at different levels. For example, legislators, parents, and other lay representatives could be principally responsible for establishing terminal standards and descriptions for real life performances. Educators could be principally responsible for establishing intermediate standards which would insure that the terminal standards were met. Communication between the lay and educational groups would be necessary, including representation in each other's standard-setting groups. This type of procedure could probably not be instituted where distrust was already present; however, it might be educationally beneficial in stable situations, and it might prevent distrust from developing.

References

Airasian, P. W., & Madaus, G. F. Criterion-referenced testing in the classroom. *NCME Measurement in Education*, May, 1972, *3*(4), 1-8.

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

Quirk, T. J. Some measurement issues in competency-based teacher education. *Phi Delta Kappan*, 1974, *LV*, 316-319.