# MEASUREMENT CONSEQUENCES OF SELECTED STANDARD-SETTING MODELS*

Richard M. Jaeger/University of South Florida

When Jim Impara first designed this symposium, he was far more thorough than other organizers I've encountered. He not only recruited paper presenters and discussants, he developed some ideas on the topic of standard-setting and was kind enough to give us some direction. It's taken me about six months to overcome Jim's kindness, and in the process, I fear I've done my colleagues here a grievous wrong.

Impara suggested that we consider four models for standard-setting: a "regression model," a "professional judgment model," an "externally imposed standards model" and an "externally validated model." By title, all of these sound plausible. In the rush to complete our symposium proposal, Howard Stoker valiantly attempted to define each of these models and to place them in some historical perspective; Lorrie Shepard wrote a volume of lucid words on how each model should be applied; and I tried to invent some psychometric problems that might be associated with each model.

Once the symposium proposal was out of our hands and under consideration by National Council on Measurement in Education, Lorrie, being the diligent scholar among us, went to the literature to see what others had said about each of Jim's standard-setting models. She found evidence that Jim would score well on the Torrance Test of Creativity—for each model was Jim's invention, and the literature provided no clues to definition or application. We immediately exchanged a number of letters and engaged in frantic phone calls. Now comes my contribution to the damage. It seemed to me that there really weren't four distinct classes of standard-setting models, only two: "judgmental models" and "empirical models." That sounded convincing to my colleagues, and we set about to build our respective pieces of the standard-setting pie: definition, application, and analysis of psychometric consequences.

For my part, considerably more reflection and a more strenuous attack on the literature of criterion-referenced testing and competency-based applications of measurement convinced me that my initial categorization of models is without merit. All standard-setting is judgmental. No amount of data collection, data analysis, and model building can replace the ultimate judgmental act of deciding which performances or which levels of performance are meritorious or acceptable and which are unacceptable or inadequate. All that varies is the proximity of the judgment-determining data to the original performance. What I had earlier labeled a "judgmental model" I would now call a "proximal model" or "direct model," and what I earlier labeled an "empirical model," I would now call a "distal model" or a "derived model." In either case, subjective judgment of merit is inescapable.

One can hardly consider the act of forming judgments apart from the use of those judgments. Millman (1973), in a discussion of alternative procedures for setting performance standards, assumes that the ultimate purpose of judgment is to determine whether an individual could, if given the opportunity, successfully complete some prescribed percentage of the tasks in a well-described domain. Assessment of performance on the tasks in a domain is assumed to be an end in itself. Millman also restricts consideration of standard-setting to situations in which attribute performance, rather than variable performance, is of interest. The datum of interest is whether an individual could succeed in completing a given percentage of tasks; there is seemingly no interest in the actual difference between an individual's performance and the standard that defines minimally acceptable performance.

In this paper I will, likewise, restrict my remarks to attribute performance models. However, I will venture beyond the point where assessment of domain performance is an end in itself. The restriction is not dictated by my judgment of the ultimate utility or potential interest of variables models, only by time and energy. Consideration of purposes that go beyond assessment of domain performance appears to be an important and practical step. Suppose that you could administer all of the tasks in some well-defined domain and observe the result. For example, suppose that you knew with certainty that a given pupil could solve 72.9 percent of all two-digit addition problems. Or, to borrow an example from Ebel (1962), suppose that you knew with certainty that a given pupil could correctly recognize the definitions of 45.2 percent of the words in *Webster's New Collegiate Dictionary*. Would certain knowledge of these facts be useful? I strongly doubt it. Neither statistic would provide a prescriptive basis for remedial education. After all, if it's worth learning to solve any two-digit addition problems, it's worth learning to solve them all. More work is needed to raise the 72.9 percent to 100 percent, true. But which problems can the student solve now? Why were 27.1 percent missed? The statistic provides no answers or clues. Is recognition vocabulary of 45.2 percent of the words in Webster's New Collegiate adequate? Is this a valuable finding? I suspect that a judgment on adequacy depends on tacit inferences to other domains: If recognition vocabulary is good, isn't generative vocabulary more likely to be good? If recognition vocabulary is as high as 40 percent of Webster's words, isn't sight reading likely to be satisfactory too? And won't this enhance school performance in the upper-level grades? And isn't reading skill necessary for success in professional occupations? Clearly, there are hundreds of potentially valuable (and possibly necessary) inferences of ultimate interest. Each inference depends on critical assumptions about the relationships between domain performance and behaviors external to the domain. Psychometrical-

*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.

ly, these assumptions are just as vulnerable as any other component of performance assessment.

## Considerations and Purposes

Cronbach (1971) indicates that validity is not an attribute of a measure, but is associated with various inferences that are based on the results of measurement. A given measurement result, obtained under certain conditions and for certain populations, validly supports some inferences and does not validly support others. Certain threats to validity arise because of the measure itself, others because of the way the measure is used, and still others because of the inference to be made. So it is with standard-setting models.

In this paper, I will briefly consider the use of a number of standard-setting procedures in support of two kinds of inferences: inferences to the performance of individuals on a well-prescribed domain of tasks, and inferences to the performance of individuals on some "ultimate" criteria that lie outside a sampled domain. For each procedure in each application, I will attempt to identify the principal threats to the validity of the desired inference. I will describe the nature of the threats and offer some conjectures on their seriousness. Discussion of the seriousness of threats to validity must be speculative, because the needed research is yet to be done. Perhaps this paper provides its best service by identifying a body of needed research.

## Models and Problems

### Proximal Standard-setting Models

*Inferences to Domain Performance.* Where proximal data are used to set a performance standard and the inference of interest is to performance on a domain from which assessment tasks have been sampled, several procedures can be conceived. For each it is assumed that a judgment has been made on the percentage of tasks in the domain that must be completed successfully in order to label an individual "successful." This judgment could be based on an explicit definition of the kinds of tasks that compose the domain, such as an item form (Hively, and others, 1968), or on examination of all of the tasks that compose the domain, as would be possible in Ebel's definition of domain-sampled tests (1962).

Once adequate domain performance has been specified, there remains the practical problem of setting a standard for successful performance on a sample of tasks from that domain. Thus two judgmental acts are required to set standards in this situation—both domain behavior and sampled behavior must be considered.

Lord and Novick (1968), Hambleton and Novick (1973), Nedelsky (1954), Ebel (1974) and others have discussed ways of determining "mastery" scores once adequate domain performance has been defined. Some methods depend on the judged applicability of statistical models and procedures; others depend on direct assessments of the value of individual tasks. Lord and Novick (1968) suggest that the maximum likelihood estimate of a domain score (percent successful) be used as a standard for assessing success on domain of tasks. The binomial distribution is used as a statistical model in this procedure. It is assumed that the set of tasks used to collect behavioral data is randomly sampled from the domain, and that the responses to these tasks are experimentally independent.

Let us now examine threats to the validity of the inference from the results of sampled behavior to the judgment of success on the domain of tasks. First, the standard of domain behavior that defines success was determined judgmentally. Given the same body of information on the nature of the domain of tasks, it is likely that different samples of judges—whether subject-matter experts, taxpayers, parents, or politicians—would set somewhat different standards. Thus the domain standard, given consistent definitional material, will vary randomly, and the judgment of an individual as successful or unsuccessful will likewise vary. The variance of judges' standards of success on various domains of tasks is largely unexplored, but would likely depend on the type of domain and the nature of the domain definition. Thus, there is a second source of judgment variance. How do judgments on standards for domain performance vary as a function of the definition of the domain? Suppose an item form is provided, or sample tasks are shown, or the entire domain of tasks is available for perusal. How do these factors influence standard-setting? Again, research is needed.

Once the domain performance standard is set, the validity of a maximum likelihood inference from a sample standard to the domain standard is threatened in at least two ways. First, the sample of performance might not be representative of domain performance; bias error may be present in the sampling of tasks. Second, the sample of tasks might not be of adequate size; too high a proportion of those judged successful on the sample of tasks might be below the passing standard on the domain, or too high a proportion judged unsuccessful might be above the passing standard. If the sampling of items conforms to the rules of simple random sampling, (or those of a variety of other probability sampling procedures), the maximum likelihood estimator of percent successful on the domain of tasks will be unbiased. Millman (1972, 1973, 1974) has provided tables that aid in the determination of the number of tasks that must be sampled in order to control the magnitude of random errors of estimation of a domain score.

Four sources of error then—random error among judges who set standards for domain performance, error due to the description of tasks in a domain, bias error due to inappropriate sampling of tasks, and random error due to an inadequate sample of tasks—threaten the validity of inferences from sampled behavior to domain performance in this case.

Novick, Lewis and Jackson (1973) propose the use of a Bayesian estimator of domain performance, based on the sampled performance of all individuals tested on a given occasion. All of the threats to the validity of a standard based on maximum likelihood estimation would be present with this procedure as well. There are at least three additional error sources. First, a Bayesian estimate of an individual's domain performance depends not only on the sampled performance of that individual, but on the performance of all of those assessed in the same group. Thus, using a Bayesian estimator, an individual could be judged successful when tested with one group, and be judged unsuccessful when tested with another group. Both bias errors and random errors in the selection of examinees are possible. The first may be due to a faulty sampling procedure or an inadequately defined examinee population, the last due to an examinee sample of insufficient size. Known theory (Novick and Jackson, 1974) can be used to determine the variability of Bayesian estimates of

domain performance, as a function of the composition examinee samples, examinee population definition, or examinee sampling procedures, but has not been explored in the context of competency-based assessment, and is another subject in need of investigation. It is known that Bayesian procedures tend to regress an individual's sampled score toward the mean of the distribution of performances of those in the sample group. The magnitude of the regression effect increases as the difference between the mean and the individual's performance score increases. Thus, an individual might well be judged unsuccessful when examined with a group of poor performers, but judged successful when examined with a group of high performers. Empirical research on the biasing effect of examinee selection on Bayesian success estimates is ripe for investigation.

Hambleton and Novick (1973) suggest that a decision on an examinee's performance, relative to a domain standard, should consider the consequences of the decision, as well as a sample-based estimate of performance. Some cost or loss is incurred if persons whose true performance is above the domain standard are judged to be performing below that standard. Likewise, there is a loss (possibly different in amount) if persons whose true performance is below the domain standard are judged to be successful. The ratio of these losses, as well as the probabilities of incurring each type of error, can be used to make decisions corresponding to each possible result of sampled performance so as to minimize expected loss. Either the binomial model based on individual performance, or the beta-binomial Bayesian model that uses individual and group performance, can be used to estimate the probabilities of scores above or below the domain standard, given a sample of performance. If the assumptions of the beta-binomial model are appropriate, errors of misclassification will be smaller than those resulting from the binomial model. If the Bayesian prior distribution assumptions don't hold, even larger errors can occur. It would be interesting to conduct an experiment using actual performance data, together with performance histories, to check the correspondence between a priori distribution choices and actual distributions. Although the Bayesian approach has been shown to be theoretically robust by Novick and Jackson (1974) and others, the magnitudes of errors likely to be made in specifying a priori distributions are unknown. Threats to the validity of inferences from sampled performance to domain performance identified for the binomial and Bayesian models above, apply with equal force to the decision-theoretic procedure when binomial or Bayesian estimation procedures are used. In addition, judgments on the magnitudes of losses incurred as a result of errors are subject to variability across judges as well as bias errors that might be associated with different types of judges; e.g., teachers might have different perspectives on losses than would guidance couselors.

Several researchers have suggested that sampled tasks be examined individually and that judgments be made on such factors as 1) the probability that an examinee whose performance barely exceeds the domain standard could perform the task (Angoff, 1971), 2) the value or worth of the task in estimating domain performance and the probability that a barely successful examinee could eliminate each incorrect option in a sample of multiple-choice items (Nedelsky, 1954). In each case, it is assumed that separate judgments of adequate domain performance have been made. For each of these methods of standard-setting, decisions on individual task difficulty are combined through an appropriate formula to establish a standard of performance on sampled tasks.

Inferences from sample performance to domain performance, based on methods that require examination of individual tasks, have the same validity threats as do those based on maximum likelihood estimation. In addition, the sample of judges used to evaluate individual tasks may be biased in their judgments. It is also likely that judgments of task difficulty will vary randomly across samples of judges. Thus, an examinee judged successful on the basis of one set of sampled judgments might be judged unsuccessful on the basis of another. The contribution of interjudge variability or systematic basis to the invalidity of such inferences is not known; again, research is needed.

Another class of standard-setting procedures requires the collection and evaluation of performance data before the standard is established. In these procedures the need for judgment is not eliminated; the object of judgment is merely shifted.

In one such procedure, the purpose of evaluation is selection, and the object is to select those whose domain performance is better than that of some percent of a population of examinees. Thus, if the data were available, the domain performance of population of examinees would be used to establish a domain standard. In practice, both examinees and tasks are sampled, and the distribution of performances of sampled examinees on sampled tasks is used to set a performance standard. Threats to the validity of inferences based on this procedure are numerous. The sampling of tasks and the sampling of examinees are subject to bias and random errors. Further, the judges used to set standards based on sample results may be systematically biased, and the variability of standards set by different samples of judges may be unacceptably high. Here too, empirical research is needed to establish the absolute and relative magnitudes of errors from these sources.

*Inferences to Ultimate Criteria.* Throughout the preceding discussion I have assumed that the inference of interest was to the proportion of tasks in a specified domain that an examinee could complete successfully, if given the opportunity. I have already argued that such inferences are often uninteresting or insufficient.

To make inferences that go beyond the domain sampled clearly involves additional threats to validity. First, one must question the appropriateness of the sampled domain: If one could observe performance on the entire domain of two-digit addition items, would an assessment of this performance support an inference to performance on all addition problems? Would it support an inference to performance on problems involving all four basic arithmetic operations, provided only two-digit numbers were used? The questions are endless and the answers are few. Often our interest in current performance only substitutes for our true interest in later performance, perhaps years hence. To support inferences beyond the sampled domain of tasks, we must determine stable relationships between proximal and ultimate performance criteria. And these relationships must be realistic for a wide variety of examinee types assessed under many different conditions. Given our lack of success with seemingly elementary relationships in learning hierarchies, these threats to the va-

lidity of inferences based on standards of sampled performance appear serious indeed.

## Distal Standard-setting Models

*Inferences to Domain Performance.* A third set of standard-setting procedures involves the use of behavior in a related domain of tasks or on some external measure of performance to set the standard of performance for the domain of central interest. As an example of these procedures, suppose the domain of interest was composed of those tasks judged to be essential to successful classroom teaching. To observe performance on a sample of tasks for a large group of examinees would be costly and difficult. The use of actual schools would be required (a procedure that could be questioned on ethical bases) or complex simulations would have to be designed and conducted. As a less-expensive surrogate for the domain of interest, one might be tempted to define a domain composed of pencil-and-paper tasks like those on the National Teacher Examination (NTE).

If a procedure analogous to the one described were used, a two-stage chain would be required to support inferences from observed performance to performance on the domain of interest. Three kinds of judgmental standards would be required. First, a standard of performance on the domain of ultimate interest would be needed. Second, a standard of performance on the sampled domain would be required. Finally, it would be necessary to set a standard of performance for observed behavior. The inferential chain would extend from the observed performance to the sampled domain, and then from the sampled domain to the domain of ultimate interest.

Threats to the validity of inferences based on the three types of standards abound. Inadequate descriptions of either or both domains of tasks could bias standard-setting. The samples of judges used to establish standards could be unrepresentative of an appropriate population of judges, resulting in bias error. The samples of judges used to establish standards could be inadequate in size, resulting in unacceptably large random errors. If relationships between performance on tasks from the sampled domain and those from ultimate domain of interest were determined empirically (using samples of tasks and examinees sufficiently restricted in size to make the assessment feasible and affordable), bias and random sampling errors from three sources would threaten the validity of the findings. We have little information on the probable magnitude or seriousness of these threats.

*Inferences to Ultimate Criteria.* There is really nothing new to be said for these cases; the threats to validity of inferences already suggested for the simpler situations apply here as well. As might be expected, the validity of inferences from performances on a sample of tasks from one domain, with a standard inferred from a second domain, and desired inferences to yet a third domain (the essence of these cases), is subject to a great many risks. Enumeration of these risks would be repetitive. Just consider the threats to validity of inferences based on the maximum likelihood model, add those associated with inferences to ultimate criteria, and include those associated with the distal standard-setting models already described. The list is discouraging, if not mind-boggling, and our knowledge of the magnitude of errors and the severity of such validity threats is extremely limited.

## An Attempt to Summarize and Some Final Judgments

Threats to the validity of inferences, based on various standard-setting models, are summarized in Table 1. The table is in matrix form with standard-setting models forming one dimension, and threats to validity of inferences forming the second. Two classes of models are indicated–proximal and distal–although utility of the distinction is open to question. Likewise, two classes of inferences are used to sort threats to validity. Threats to validity listed under "Inferences to ultimate criterion performance" should be considered supplements to, rather than substitutes for, those listed under "Inferences to domain performance." An "x" in the body of the table indicates that a threat in a given column applies to a standard-setting model in a given row.

The research that exists on standard-setting procedures in competency-based education appears to be largely theoretical. Millman's work on the relationship between random errors and the size of task samples is based on a well-known binomial probability model and several plausible, but untested, procedural assumptions. Hambleton and Novick's results on the application of decision theory and beta-binomial analysis to performance estimation likewise makes use of elegant statistical theory.

Throughout this paper, I have identified questions for which research-based answers are apparently unavailable. Application of statistical models and theoretical formulation are unlikely to provide answers to these questions. There is need instead for empirical investigation involving human standard-setters in real or simulated judgmental situations, using real performance data and real descriptions of task domains. Perhaps a theory of validity will emerge from this research eventually, but tentative guides to practice, sorely needed, will be developed in the interim.

**Table 1.**

**Threats to the Validity of Inferences Based on Selected Standard-setting Models**

| Standard-setting Model and Procedure | Threats to Validity of Inference to Domain Performance | | | | | | | | | | Additional Threats to Validity of Ultimate Criteria | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 |
| 1. Judges set domain passing score directly<br>a. Maximum likelihood estimate of domain passing score used as sample passing score | X | X | X | X | | | | | | | X | X | X |
| b. Bayesian estimate of domain passing score used as sample passing score | X | X | X | X | | | | X | X | | X | X | X |
| c. Decision-theoretic model (with losses) used to set sample passing score | X | X | X | X | | | | | | X | X | X | X |
| d. Judges determine criterion worth of each task directly | X | X | X | X | X | | | | | | X | X | X |
| 2. Estimate of a percentile used to set passing score<br>Tasks are administered to a criterion group; percentile is used to set passing score | X | X | X | X | | X | X | | | | X | X | X |
| 3. Performance in a related domain or on an external measure used to set passing score<br>a. Contingency table analysis based on two domain samples of tasks (e.g., one costly, one inexpensive) | X | X | X | X | | | | X | X | | X | X | X |
| b. Mathematical model relating two domains or performances on domain tasks and an external measure | X | X | X | X | | | | X | X | | X | X | X |

**KEY**

**Table 1. Threats to the Validity of Inferences Based on Selected Standard-Setting Models**

**Threats to Validity of Inferences to Domain Performance:**

1 = Bias in setting domain standard due to inadequate domain definition
2 = Random error among judges who set domain standard
3 = Inappropriateness of item sampling procedure: bias error in sample standard
4 = Inadequate item sample size: random error in sample standard
5 = Judgment bias in the consideration of individual tasks
6 = Lack of representativeness in the criterion group used to determine percentiles
7 = Inadequately large criterion group, resulting in random error
8 = Lack of representativeness in examinee group, leading to bias error
9 = Inadequately large examinee group, leading to random error
10 = Model bias due to inaccuracy in loss judgments: false positives or false negatives overvalued

**Additional Threats to Validity of Inferences to Ultimate Criteria:**

11 = Bias error due to invalidity of domain identification
12 = Error due to inconsistency of domain-criterion relationship
13 = Bias error due to invalidity of model for domain-criterion relationship

## References


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

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

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

Ebel, R. L. *Measuring educational achievement.* Englewood Cliffs, N.J.: Prentice-Hall Publishing Co., 1974.

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

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

Lord, F. M., and Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley Publishing Co., 1968.

Millman, J. Tables for determining the number of items needed on domain-referenced tests and the number of students to be tested (Technical Paper No. 5). Los Angeles: Instructional Objectives Exchange, April, 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.* Berkeley, Calif.: McCutchan Publishing Co., 1974.

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

Novick, M. R., Lewis, C. , and Jackson, P. H. The estimation of proportions in m groups. *Psychometrika,* 1973, *38,* 19-46.

Novick, M. R., and Jackson, P. H. *Statistical methods for educational and psychological research.* New York: McGraw-Hill, 1974.