A Standard Vision

Gregory E. Stone, Ph.D.

Who passes and who fails? What does it mean to pass? How can a fair and meaningful standard be established? Such questions are routinely asked within many different educational and evaluative settings. The stakes are high, the requirements important – a public at large depends upon these measurement devices to graduate and pass qualified candidates.

There are as many different models, empirical and otherwise, for establishing passing standards as there are examinations themselves. Some reflect complex relationships between statistical technique and judgement making, others a simplicity of qualitative purpose. All attempt to create a reasonable decision, and most are subject to significant criticism on grounds of equity, precision, and meaningfulness. In this article a conceptual and fundamental framework within which all models may be evaluated is discussed.

Regardless of the model, every standard setting method must effectively demonstrate the desired criterion, be reproduceable, and remain genuine. It is important to note that in the efforts of standard setting, golden rods and sacred cows are of little use. Ultimately the process is genuinely evaluative, and it becomes the goal of the standard setter to define a systematic, logical and understandable quantifiable method for conduct of this qualitative exercise.

The first requirement, effective demonstration of the desired criterion, is fundamental. In criterion referenced standard setting, the criterion hopes to represent a specific body of content knowledge. Theoretically, the act of passing a test demonstrates successful mastery of this content. This interpretation of a passing outcome is only reasonable if the standard adequately reflects the content. A demonstration of adherence to content I propose to call criterion validity, in support of the criterion referenced standard. While a departure from common quantitative descriptions of validity of criterion standards, it appears both logical and desireable. Unfortunately such validity is achieved very rarely.

Traditional standard setting systems (like Angoff, for example) gather together groups of experts in a subject area and ask them to predict candidate performance. A typical question posed to these experts is "how many examinees out of 100 will answer each item correctly?" Summations and averages of these predictions of performance ultimately become the standard.

Even a superficial review of such a judgement making process reflects that the desired content-based criterion is being missed. Outcomes are necessarily linked to data input. When predictions of performance are used as 'input' it follows that the products of that predicted performance becomes the 'output'. The criterion emerging from predicted performance must be a performance criterion, not a content criterion.

To establish a content-based standard, judges must define the criterion in a manner that addresses it directly. Meaningful definition is only achievable through an exercise focussing on a qualitative evaluation of the concepts within the subject matter, rather than via unwarranted and impractical predicated quantities. Thus far, only Rasch-based models have been able to demonstrate effective content validity. In particular, the Objective model (Stone, 1994, and Gross and Wright, 1965) collects judgements in terms of essentialness of content presentation, and has successfully demonstrated a singularity between qualitative judgement and quantitative outcome. Objective models allow content experts to be content experts – by selecting content of importance.

The second quality, that of reproduceability, is a concept not foreign to measurement. Generally considered reliability in quantitative circles, it is a question of reproduction of results. Standards must be able to demonstrate that they are applicable on more than a single version of an examination. A criterion 'standard' implies a level of achievement within a criterion. If the standard changes with each unique examination or grouping of items, how can a reasonable level of achievement be considered? A simple test of reproduceability is available to check standards.

Consider the passing rates for two content-similar, but not necessarily item-identical, examinations. If the standard is reliable, should the passing rates not also be the same? Not necessarily. There are three facets in a typical examination setting – the difficulty of the particular examination, the abilities of the examinees, and the standard used for passing. Theoretically the first two vary, whereas the latter (the standard) should not. To test for reproduceability, the examination forms must first be equated (in Rasch methodology most likely through common-item equating). Using a standard linear transformation, differences in examinee ability between the two groups can be controlled. The result will be two different groups of examinees where difficulty and ability are controlled. Testing for reproduceability (consistency) is as simple as visually inspecting the pass rates for each group. If identical (within the defined error), then the standard defined meets this requirement for reproduceability – and is, in short, reliable.

The third quality of a useful standard finds its roots in genuine scientific credibility. In few other aspects of measurement has this been such a pervasive problem. Unfortunately standards and standard setting is such a politically sensitive issue that the methods themselves have tried to adapt to these number games. Is 60% too low a pass rate? Then move the standard up to a level that will pass 70%. Don't call it fudging, call it "adjusting" and try to find a statistic (maybe the SEM or Mean person performance) that can somehow be used to justify the move. Standard setting is notorious for fudging.

In the real world, political and other considerations are important and often impact upon measured, considered decisions, like standards. Apart from politics, the real issue for the measurement professional is one of honest reflection. When standards must be changed, the role of a measurement expert is to express those changes and educate the stakeholders. What sort of content knowledge is being left out of the new standard? How may curricula be informed to raise the level of student performance? Instead of addressing these changes directly, many choose complicated "adjustment" techniques and errantly believe that the standard has somehow remained the same, just adjusted or corrected. Research honesty and integrity in creating a genuine standard that remains true to its defined meaning is imperative for the process.

Ultimately there may be many ways to define performance standards. However, there are at least three fundamental qualities that may be used to judge their merit. The redefined notions of validity, reliability and genuineness should be considered performance benchmarks. While only one model has thus far demonstrated each – the Rasch-based Objective model - the article expresses a desire that other models too will put themselves to these simple, yet fundamentally necessary tests.

To illustrate one way, through which the reliability of passing standards may be assessed, consider Figures 1 and 2. Each presents data concerning the passing rates observed on four national, high-stakes examinations. Each uniquely created exam was constructed using the identical content outline, but each contained a different set of specific items. The diamond pointed line represents actual passing rates on each successive administration using the same (equated) standards. The square pointed line represents what the passing rate would have been had difficulty of the examination and group person ability been controlled. A glance at the figures shows a clear linearity within the Objective standard - evidence of its reliability - while the Angoff standard does not. Instead, the Angoff standard itself or the error associated with it, produces wildly different results from administration to administration. Such results suggest a fairly unreliable process. Which passing rate should one believe? Why the fluctuation when all moveable factors have been controlled?

Figure 1

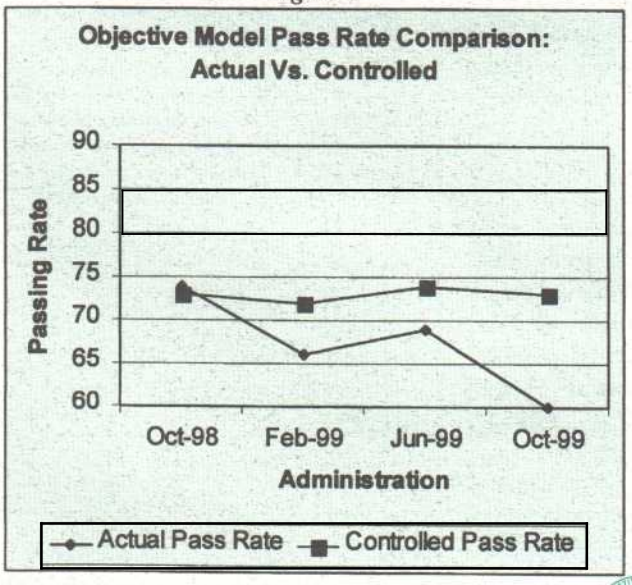

Figure 2