Standard Setting: The Next Generation

(Where Few Psychometricians Have Gone Before!)

Ronald A. Berk

WARNING: NES has determined that this material may contain harmful or dangerous ideas.

CAUTION: DO NOT drive or operate dangerous machinery while reading this material.

If DROWSINESS occurs, it shouldn't be surprising.

DO NOT induce vomiting; the content will take care of that.

Truckloads of articles, pamphlets, and book chapters have been published on the topic of standard setting. When charged with the task of executing a particular standard-setting method, executors throughout the land have exhibited high blood pressure, high anxiety, ulcers, and, in one case, even hairballs.

Due to these possible physiological effects of standard setting, it has been suggested that some professionals who are involved in the standard-setting process seriously consider taking the new medication STANDARDSET® (generic name is CUTSCOR). A description of this medication is shown in Figure 1.
Figure 1

STANDARDSET®
(CUTSCOR)

The most widely prescribed medication for raising standards in America.

Are you doing everything you can to raise your standards, but it just never seems to be enough?

Tests prove STANDARDSET® helps raise standards when adjustments based on consequential-validity evidence aren't enough. Of course, not everyone gets the same results. And not everyone on STANDARDSET® reaches his/her standard-setting goal.

Consider the following case study.

Elmo has low standards, and his father is now out of work because he performed at extremely low standards.

This put Elmo at risk. Low standards are serious, especially if you have two or more risk factors, including an unrepresentative sample of judges, unqualified judges, inadequately trained judges, or high intrajudge or interjudge variability.

Once he realized the risk, Elmo tried really hard to raise his standards.

For six months he said "no" to test-centered methods and "yes" to examinee-centered methods. He kept his fingers crossed. The psychometrician said that the contrasting-groups method alone doesn't always raise standards enough. It could be due to poor examinee selection, a large proportion of false positives, or just bad examinees; no one knows for sure.

The psychometrician prescribed STANDARDSET®.

The psychometrician added STANDARDSET® to Elmo's standard-setting program and his standards really started to go up!

Is STANDARDSET® right for you?

Ask your psychometrician. STANDARDSET® should not be used by people who are afraid of human judgment, performance assessment items, multiple cut-scores, iterations, consensus building, behavioral anchoring, or decision policy analysis, people who are not sure whether to say "polychotomous" or "polytomous," or people who are opposed to the Angoff method and any of its modifications.

There can be side effects.

It is recommended that your psychometrician perform routine statistical tests before and after treatment. Tell your psychometrician if you experience any unexplained pains while taking STANDARDSET®, as this could be a sign of serious side effects, such as adverse impact and costly litigation.

STANDARDSET® is available in 500 mg tablets or timed-release suppositories.

Ask your psychometrician about STANDARDSET®. For a free booklet, call 1-800-CUT-SCOR.

BERK PHARMACEUTICALS
(A name you used to be able to trust!)
After two decades of research and practice, the measurement community is a smidgen closer now than it was then to grappling with the pesky and prickly complexities of standard setting.

At present, there are nearly 50 standard-setting methods documented in the literature. At the epicenter of every method proposed since prehistoric times is "human judgment." Its subjectivity and imprecision wreak havoc in the minds of quantitatively trained people, which might explain some of the grumbling and murmuring among standard-setting researchers at professional conferences. The various methods require judgments about test content (the items or the test itself), examinees' performance, or a combination of both. You may be thinking, "How soon can we expect to see an objective, unbiased, nonjudgmental method?" I think we have a better chance of seeing nutritional information on the side of a Slurpee™ or seeing an American Football Conference team win the Super Bowl! If you're still clinging to this dream of a nonjudgmental method, I suggest that you let it go! In fact, what were previously labeled as judgmental and empirical methods were renamed by Jaeger (1989) as test-centered and examinee-centered methods, respectively, to acknowledge that all methods are judgmental.

Recently, the test-centered methods used by the National Assessment of Educational Progress (NAEP) to set achievement levels for the 1990 and 1992 assessments were the objects of considerable controversy and, in fact, a heavy barrage of criticism (Cizek, 1993; Kane, 1993; Linn, Koretz, Baker, & Burstein, 1991; National Academy of Education, 1993a, 1993b; Stufflebeam, Jaeger, & Scriven, 1991; U.S. General Accounting Office, 1993). The simplest single-stage methods, such as Angoff (1971), Ebel (1979), and Nedelsky (1954), which were extremely popular and reportedly effective tools for setting cut scores in a variety of applications in the 1980s and early 1990s, are now being pinned to the mat. The National Academy of Education (1993a) recommended that NAEP's modification of the Angoff method as well as any other item-judgment method be discontinued in favor of the contrasting-groups approach (Livingston & Zieky, 1982), which should be "used to evaluate the current set of achievement levels" (p. xxx). The reason given was that such methods are fundamentally flawed because they require judges to perform a nearly impossible cognitive task, i.e., estimating the probability that hypothetical students at the borderline of each achievement level will answer each item in the pool correctly.
When I first read this recommendation, it blew me off my beach chair. The Angoff method with all of its modifications has been one of the most trusted names in standard setting. And now, it seems as though it's fighting for its life along with the other item-judgment methods. However, despite the specific advantages and disadvantages of each cut-score method, which have been debated extensively, NAE (1993a) concluded that:

The most consistent finding from the research literature on standard setting is that different methods lead to different results. Not only do judgmental and empirical methods lead to different results, . . . but different judgmental methods lead to different results. In fact, judgmental methods appear to be sensitive to slight and seemingly trivial differences in the procedures used to implement a given method (p. 24).

Jaeger's (1989) analysis of 32 comparisons of methods led to virtually the same conclusion. Probably the only point of agreement among standard-setting gurus is that there is hardly any agreement between results of any two standard-setting methods, even when applied to the same test under seemingly identical conditions. This conclusion is based on empirical evidence, no less.

So, where do we go from here? There are at least four alternatives we might consider. First, we can pick just one method that seems to fit the requirements of a specific application. In other words, we can bite the bullet and simply acknowledge the psychometric limitations of the chosen method. Second, we can select one method but analyze the range of cut scores by the judges, rather than just the average, in conjunction with external validity evidence to guide the choice of a cut score. Unfortunately, it is unclear how these different types of information should be integrated to arrive at a final standard. Third, several of our fearless standard-setting leaders have suggested using two or more methods in a given application and then consider all of the results, together with extraterrestrial factors, when determining a final cut score. Since it is reasonable to forecast that these results will not converge on the same cut score, how does one reconcile the difference? This strategy might create somewhat of an unsettling feeling among a few practitioners. Finally, there's one other not so obvious option. We can scrutinize the components of all methods at our disposal and the available evidence of their effectiveness in
order to build an eclectic approach that capitalizes on the best and the brightest elements in standard-setting history. Of course, this approach would have to be tested, but it might be worth the effort.

To provide a different direction for future standard-setting practices in large-scale assessments, I have chosen the deviant course of constructing an eclectic model. (What a surprise!) However, in order to embark on this construction job, two key questions need to be answered: (1) Where have we been? and (2) What have we learned? Once these have been answered, we can answer the most critical question: So what's a practitioner to do?

Since there is some evidence that my previous papers on standard setting have produced drowsiness and, in rare instances, nausea and vomiting, readers should take heed of the caution at the beginning of this chapter. In an effort to minimize the occurrences of these side effects, the vehicle for presenting this material will be three "Top 10 Lists." This strategy also capitalizes on the "three" phenomenon in history. Several historians, all of whom wish to be disassociated with this chapter, have observed that many great events and turning points in history have occurred in "threes," such as the Three Musketeers, measures of dispersion (range, variance, standard deviation), and, of course, Angoff-Ebel-Nedelsky. This chapter will emulate that distinguished model.

**Where Have We Been?**

Numerous summaries of the standard-setting literature have been reported throughout the 1980s (Berk, 1986; Hambleton, 1980; Hambleton & Eignor, 1980; Jaeger, 1989; Livingston & Zieky, 1982; Pulakos, Wise, Arabian, Heon, & Delaplane, 1989; Shepard, 1980, 1984), and one review has appeared in the '90s (Kane, 1994). Most of these method-by-method critiques are quite lengthy. In fact, I can be accused of writing one of the longest pieces on the topic, and I have regretted the pain I have inflicted on the readers of my "consumer's guide" for quite some time. In order to redeem myself, I recently proposed the shortest summary of standard setting ever recorded (Berk, 1995). A revision of that summary will be presented here.

There is a remarkable similarity in structure among the approximately 30 methods that utilize a test-centered, judgmental review process, which has been the most popular, feasible, and credible
approach in many large-scale assessments to date. While there are numerous variations in what judges rate, how they rate, and the steps required to reach the cut score, the research and experience with these methods have taught us more about what works in the standard-setting process than in the less frequently used and studied examinee-centered, empirical methods. Therefore, the summary presented here is derived primarily from the test-centered methods and, to a lesser extent, from the examinee-centered methods. Rather than reviewing the specific methods, which has already been done, this presentation lists the pieces in the standard-setting puzzle, which have been extracted from those methods.

So, from the home office in Baltimore, Maryland, here is Top 10 List Number One.

**Berk's Top 10 Characteristics of the Judgmental Standard-Setting Process**
(a.k.a. Where Have We Been?)

10. Given enough time, judges can be trained to do just about anything.

9. Judges can rate objective or constructed-response items, clusters of items, work samples, profiles of behavior, or whole tests.

8. Judges can weight the importance, difficulty, or complexity of objectives or behavioral dimensions.

7. Judges can match the content in any item format or an examinee's item/work sample/test performance to behavioral descriptions of achievement levels.

6. Judges can use feedback on individual ratings, panel ratings, and/or pertinent quantitative performance information on examinees after their initial ratings.

5. Judgments about standards can be determined by judges independently or through consensus after group discussion.

4. The judgmental process can be single stage, multistage, or iterative.
3. The pattern of ratings structured by objective or dimension can be compensatory, conjunctive, or disjunctive.

2. The standard can be computed based on consensus of the panel of judges or from the average or median percentage/proportion across judges, weighted average proportion across judges, median percentage across samples of judges, or any of those measures in conjunction with an examinee-performance-based scale.

And the number 1 characteristic of the judgmental standard-setting process:

1. The standard that emerges from any method will probably be adjusted higher or lower based on judgments about the political, economic, social, and/or educational consequences of the decisions.

These characteristics can be viewed as ingredients or options available in the process. They have been applied in a variety of combinations since the early 1970s to test scores used in criterion-referenced or competency tests of student performance; teacher and professional licensing and certification exams; personnel evaluation of teachers, counselors, and administrators; and performance appraisal of employees and managers in private industry and the military. The pivotal question is: What optimal combination of ingredients produces an effective standard-setting procedure? As I've already suggested, we don't know! The problem is that the measurement community has never reached consensus on a set of criteria that can operationally define the "effectiveness" of any single standard-setting method. That's what makes research on this topic so much fun! There are NO RULES!

Lists of criteria for evaluating the various methods have been proposed previously (see, for example, Berk, 1986; Hambleton & Powell, 1983; Plake 1995). The 1985 edition of the Standards for Educational and Psychological Testing (AERA, APA, & NCME Joint Committee, 1985) provides standards that focus more on the technical outcomes in terms of reliability and validity evidence, such as Standards 1.23, 1.24, 2.10, 2.12, and 11.3, rather than on the standard-setting process itself. Only Standard 6.9 states that the
method, rationale, and technical analyses should be reported along with the qualifications of the judges involved in the process:

When a specific cut score is used to select, classify, or certify test takers, the method and rationale for setting that cut score, including any technical analyses, should be presented in a manual or report. When cut scores are based primarily on professional judgment, the qualifications of the judges also should be documented (Primary) (p. 49).

Although these published sets of criteria and the Standards do not provide the level of specificity the profession seems to need to guide standard-setting practices, there are high expectations among many measurement professionals that the current revision of the 1985 Standards will set explicit requirements for the entire standard-setting process.

What Have We Learned?

By default, with no professionally agreed-upon criteria at this time, how do we determine the "best" elements or attributes of the standard-setting process? The previous Top 10 List presents a confusing array of options. Furthermore, techniques that were effective during the 1980s may not be applicable now. Throughout the '80s, item-judgment methods, such as Angoff's and its numerous variations, were the predominant choices of practitioners for setting standards in a variety of applications. At the outset of the 1990s, two trends in the large-scale assessment business—(1) the increased use of multipoint item formats and (2) the requirement of multiple cut scores—seriously questioned the utility of these "most-loved" methods. Either the older methods had to be retooled or new approaches had to be developed. Let's examine some of the recent developments in these areas.

Recent Developments

Multipoint item formats. The performance assessment mania sweeping the country has resulted in a meteoric rise in the use of constructed-response item formats, essays, writing samples, oral discourse, exhibitions, experiments, portfolios, and assessment centers for student, teacher, and administrator examinations (Berk, 1993; Office of Technology Assessment, 1992). This translates into a variety of scoring protocols and multipoint or polytomous (pronounced polly toe’mus) (a.k.a. polychotomous) items.
Exercises that are scored on a range of 0–2 or greater necessitate a different cut-score strategy from the item probability judgments recommended by most standard-setting methods for dichotomous items (see, for example, Hambleton & Plake, 1995).

New approaches are currently being investigated to move beyond the simple item-by-item judgments of yesteryear. These approaches include the following: (a) setting cut scores on each multipoint item and the total test (Delaware Department of Public Instruction, 1993; Freed, 1993), (b) setting scale score bands or intervals around achievement level thresholds (Atash, Ferrara, & Bayless, 1994; Kahl, Crockett, & DePascale, 1994), (c) requesting judges to specify score performance distributions (Poggio & Glasnapp, 1994), and (d) stating decision policies across an entire assessment package (Putnam, Pence, & Jaeger, 1995). These efforts concentrate judgments on scores derived from aggregates of items, exercises, or work samples. Yet they also capitalize on many of the aforementioned elements that have proven successful in past standard-setting projects, such as using an iterative process (Jaeger, 1982), and are equally applicable to the traditional dichotomously scored item format. However, further research is needed to evaluate the most effective "elements" in these applications.

**Multiple cut scores.** In another twist on the passing score theme, there has been a trend by a few agencies to set multiple cut scores, such as two, three, or four, to categorize examinees by levels of achievement, competency, or proficiency. For example, NAEP specifies "Advanced," "Proficient," and "Basic" (American College Testing, 1993; National Assessment Governing Board, 1990); the National Board for Professional Teaching Standards recommends "Highly Accomplished," "Accomplished," "Competent," and "Novice" (Jaeger, 1995); the Kansas State Board of Education uses "Excellent," "Strong," "Progressing," "Borderline," and "Inadequate" (Poggio & Glasnapp, 1994). Several other state education agencies (SEAs), including Delaware, Kentucky, Maine, Maryland, Massachusetts, and New Hampshire, have proposed similar schemes (see, for example, Atash et al., 1994; Delaware Department of Public Instruction, 1993; Freed, 1993; Kahl et al., 1994; WESTAT, 1993).

As if setting one passing score weren't challenging enough to defend psychometrically, these multiple cutoffs complicate the judgmental process even more. This complication begins with defining every level of achievement for which a cut score must be
determined. The achievement level definitions serve as behavioral descriptions or examples of performance. The procedure used to describe these levels is referred to as "behavioral anchoring." Selected points on a scale, called anchor points, are described in terms of the knowledge, skills, and abilities (KSAs) exhibited by students who are near those points. These descriptions are similar to the "benchmarks" used on behaviorally anchored rating scales (BARS) (Smith & Kendall, 1963) and behavior summary scales (BSS) (Borman, Hough, & Dunnette, 1976) originally designed to rate the job performance of employees in industrial, educational, and governmental organizations in the early 1960s and 1970s (for examples of scales, see Borman, 1986). An example of behavioral anchoring for four achievement levels is shown in Figure 2. Once the achievement levels are explicitly defined, anchor or prototype exercises, behaviors, work samples, or other item formats are assigned to the designated levels by the judges.
### Figure 2
Example Of Behavioral Anchoring
For Four Achievement Levels

| ULTRA STUPENDOUS  
(Gold Medal) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaps tall buildings in a single bound.</td>
</tr>
</tbody>
</table>

| KINDA BETTER THAN ORDINARY  
(Silver Medal) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaps short buildings in two or more bounds, maybe.</td>
</tr>
</tbody>
</table>

| MIDDLING GARDEN-VARIETY  
(Bronze Medal) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Leaps short buildings with a running start and a strong tailwind.</td>
</tr>
</tbody>
</table>

| PUTRID  
(Pet Rock) |
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Barely leaps over a Port-O-Potty</td>
</tr>
</tbody>
</table>
This strategy for defining three or more achievement levels and sorting student behaviors into these levels has been tested in conjunction with different item response theory (IRT) methods in several state assessment programs. For example, the one-parameter partial-credit model has been used to scale test scores, item difficulty values, and examinees' ability estimates onto a common scale called an "activity difficulty scale." Researchers at the Maryland State Department of Education (Atash et al., 1994; WESTAT, 1993) have explored the use of a judgmental process to set achievement levels with narrow bands of points on this difficulty scale that discriminate between performance at two adjacent levels. The behaviors and skills assigned to these levels by the judges provide the foundation for deriving descriptions of the achievement levels.

Other IRT approaches have been investigated by Kahl et al. (1994) in Maine's and New Hampshire's assessment programs. A Student-Based Constructed-Response (SBCR) method, where students are placed on a Rasch ability scale based on their scores on all of the common questions they answered, requires judges to review a complete set of responses for every student whose work they examine. Alternatively, an Item-Based Constructed-Response (IBCR) method places score points for individual items on the ability scale. In this method, judges review responses sorted by score point by item. Both the SBCR and IBCR methods request judges to participate in "range-finding" activities to minimize their review load. They match student work to predetermined definitions of four levels of achievement ("Distinguished," "Proficient," "Apprentice," "Novice") and three cut scores on the scaled score continuum.

Other variations of the "behavioral anchoring" method were explored by the National Board for Professional Teaching Standards. Jaeger (1995) and Putnam et al. (1995) investigated two "judgmental policy capturing" strategies adapted from a procedure developed by Hobson and Gibson (1983) for the performance appraisal of employees. It focuses on the pattern of responses or the policy used by judges to arrive at their final standard. The judgmental process involved defining multiple levels of competency, rating profiles of teacher performance, and then rating hypothetical candidates with specified profiles of exercise scores from "Novice" to "Highly Accomplished." The policy that emerges to describe the judges' responses can take one of three possible forms: (1) compensatory, where the cut score is simply
the total score across all items and objectives, (2) conjunctive, where objectives/dimensions are weighted by importance, and different cut scores may be set across objectives in order to pass, or (3) disjunctive, where a mixture of (1) and (2) may be designated with cut scores set for some objectives and total scores computed across other objectives.

This analysis of how judges' decisions are made by item and objective to produce a single cut score or pattern of cut scores for a test was considered for the student competency and teacher licensing tests in the 1970s and 1980s. However, the compensatory model with its inherent simplicity (i.e., just add 'em up) reigned supreme. The research by the National Board has refocused attention on these decision-rule options to more accurately and sensitively reflect how judges think and feel about what candidates should be able to perform.

Considering all of these developments within the past four years in conjunction with the previous decade of research and experience with item judgment methods, here is Top 10 List Number Two.

**Berk's Top 10 Picks for the Judgmental Standard-Setting Process**
(a.k.a. What Have We Learned?)

(Note: The following picks are for professional discussion only and should not be used as the basis for any actual cash wager.)

10. Select a broad-based sample of the most qualified and credible judges you can find.

9. Train these judges to perform the standard-setting tasks to minimize "instrumentation effect" and maximize "intrajudge consistency."

8. Use a multistage iterative process whereby judges are given one or two opportunities to review/refine their original decisions based on new information to maximize "interjudge consistency."

7. Require judges to provide explicit behavioral descriptions for each achievement level with corresponding anchor items.

6. Determine the judges' decision policy based on the objectives or dimensions measured.

5. Provide judges with feedback on their individual and the panel's decisions.
4. Supply judges with meaningful performance data on a representative sample or appropriate subsample of examinees to "reality-base" the ratings.

3. Allow judges the opportunity to discuss their decisions and pertinent data without pressure to reach consensus.

2. Solicit judges' content-related decisions about achievement levels via consensus but all item and test score decisions via independent ratings to avoid pressuring "outlier judges" into alignment or the influence of "dominant judges."

And the number 1 pick for the standard-setting process:

1. Compute the cut score(s) from the mean/median item or test scores based on the judges' ratings.

So What's A Practitioner To Do?

If the elements in this preceding Top 10 List could be synthesized into a package plan to be manufactured as a generic brand of standard setting, what would it look like? It should incorporate the best of the past with the most promising new techniques in order to handle any item or test format and single or multiple cut scores, and yet be feasible for large-scale assessments. The structure should permit the user to choose between test-centered and examinee-centered components, and it should be applicable to student as well as teacher assessment. The steps in this Generic Eclectic Method or GEM are listed next.

Preliminary Steps

Two giant steps need to be completed before the standard-setting games begin: (1) select samples of judges and (2) train the judges.

Select samples of judges. Pick the most credible judges you can dig up. For most large-scale assessments it is advisable to choose two samples of judges: (1) a broad-based, diverse group representative of the population of educators (such as teachers, curriculum experts, and local- and state-level administrators) and noneducators (such as parents, professionals in a variety of disciplines, and the general public) and (2) a specialized group of content experts by grade level corresponding to the test content areas and grade levels. Beyond the professional composition of the samples, both should be representative of all sociodemographic characteristics. This assures sampling precision and diversity and avoids the appearance of an all-dweeb judgmental panel. The
broad-based sample should be selected first; then the content experts can be drawn as a subsample or as a separate sample.

The broad-based, rather heterogeneous sample should be employed to make decisions about achievement level definitions and items within the limits of its expertise. For tests measuring specialized content across a range of cognitive levels at different grade levels, one or more homogeneous subsamples should be composed of appropriate content experts with the qualifications to render valid decisions at the various steps in the process. The composition of each subsample of judges should accurately reflect the types of content and levels of cognitive complexity being measured, the difficulty of the items, and the grade levels. Unqualified judges who are required to make uninformed decisions at any stage of the process will tend to contribute inaccurate information and error variance. The validity and generalizability of the decisions made by the judges depend on their competence, representativeness, and credibility.

Jaeger's (1991) suggestions regarding sample representativeness and size should be considered in this selection process:

Judges should be selected through procedures that permit generalization of their collective recommendations to well-defined populations. The number of judges . . . should be sufficient to provide precise estimation of the standard that would be recommended by an entire population of judges . . . such that the standard error of the mean or median recommended standard is small, compared to the standard error of measurement of the test for which a standard is sought (p. 10).

**Train the judges.** Train these judges 'til it hurts. The effectiveness of the training program can determine the effectiveness of the entire standard-setting process. The major objectives of the training are to provide judges with a thorough understanding of the concepts of achievement or proficiency levels and to promote competence in the tasks required by the standard-setting method. The research on factors that influence judges' ratings during training and the stages of standard setting deal almost exclusively with item judgments and defining minimal competence (Melican & Mills, 1986, 1987; Mills, Melican, & Ahluwalia, 1991; Norcini, Shea, & Kanya, 1988; Plake, Melican, & Mills, 1991; Pulakos et al., 1989; Smith & Smith, 1988). Despite the limited focus of these studies, the criteria proposed by Reid (1991) for evaluating training effectiveness...
based on their findings can be generalized to the more recent standard-setting methods: (1) judgments should be stable over time, (2) judgments should be consistent with item and test score performance, and (3) judgments should reflect realistic expectations.

When conducting the training, how much practice is enough? For example, suppose judges are matching nontest items from the item domain to previously defined achievement levels. A judge should be able to establish a high level of consistency (intrajudge) in matching items at different levels of difficulty or complexity to the appropriate achievement levels. The judge should not waiver during the process in his/her understanding of the items, the achievement levels, or in the matching itself. Once judges have reached this level of competence in the specific matching or rating task and are confident in the stability of their decisions, SOCK IT TO 'EM!

Since the background characteristics of the judges can affect the speed with which each judge masters the task, individualized training using an interactive computerized presentation should be considered. A computerized procedure will also permit constant monitoring of the consistency of the judges' performance during training and facilitate intrajudge consistency during the standard-setting process (Plake et al., 1991). The judges' confidence in their decisions can also be assessed and correlated with their consistency. A rating scale can be used to measure confidence at the conclusion of training and after the final decisions have been made during standard setting. (To reward the judges for all of their efforts during training, take 'em to lunch or give 'em gift certificates for Extra Value Meals™ at McDonald's®!)

Are you ready for this? Another Top 10 List. Talk about beating a dead horse! For those of you who watch Letterman, you'll probably never watch him again! Well, here comes Top 10 List Number Three, plus criteria to evaluate each step.

**Berk's Top 10 Steps to Standard-Setting Salvation** *(a.k.a. Generic Eclectic Method or GEM)*

10. **Broad-based sample of judges defines achievement levels by subject area and grade level based on consensus.**

Criteria: This diverse group of judges should reach consensus on general policy and content definitions of what types of examinee performance are expected at each designated achievement level. If policy definitions have
already been set by a governing board, then either the judges can extend those definitions to clarify their intent or skip this step entirely, pass GO, and collect $200.

9. **Subsample (content experts) of judges provides amplified, explicit behavioral descriptions of achievement levels by subject area and grade level based on consensus.**

**Criteria:** This merry band of content experts should build on the general definitions produced in step 10. They should reach consensus on specific, operational descriptions of the knowledge, skills, and abilities (KSAs) that must be demonstrated at the different achievement levels. The meaning of those levels and the interpretation of the final cut scores hinge on the clarity of the behavioral definitions at this step. (The descriptions will most likely need revisions once the entire process has been completed. The actual items selected in subsequent steps of the process may suggest some refinements in the original behavioral descriptions to improve their meaning.)

8. **Judges select anchor/prototypic unscored items (test centered) or scored items/work samples/tests (examinee centered) at upper and lower ends of achievement level categories based on consensus.**

**Criteria:** This step is the standard-setting version of "boot camp," which gives judges the opportunity to "BE ALL THAT THEY CAN BE!" It is intended to train the judges in item content and difficulty, achievement level definitions, and the matching process. Non-test items should be used for this practice exercise until competence and confidence are attained. This training may be conducted on an individualized basis via computer.

At this step, decisions must be made on the unit of judgment (item, work sample, or total test), item-scoring format (dichotomous, polytomous, or a combination of both), and test-centered (unscored unit) or examinee-centered (previously scored unit) approach. For example, the units may consist of multiple-choice and constructed-response items that have not been administered to any examinees. Matching the content and perceived difficulty of these items to the achievement levels is a test-centered approach. Alternatively, if the units have already been administered and scored based on a sample of examinees, then the scored
level of performance on each item becomes part of the matching criteria. In other words, since the examinees' scores are used to assign the items to the achievement levels, the method is examinee centered.

Once each judge has completed the training, all judges should be convened to select anchor items/work samples from the practice pool for upper and lower ends of the achievement levels. These anchors should provide concrete representations of the descriptions and include all item formats used on the final test. The judges should reach consensus on a set of anchors for each achievement level. (The anchors may need relocation or replacements after they have been tested in the process.)

7. Judges independently match unscored items (test centered) or scored items/work samples/tests (examinee centered) to achievement-level categories based on behavioral descriptions and anchors.

Criteria: After the training and final agreement on anchors at step 8, all of the judges should independently match all test items/work samples/tests to the appropriate achievement levels based on content, perceived difficulty, and, if scored, level of performance. If this process could be conducted on microcomputers, each judge could be monitored to detect "drift" in the degree of interjudge inconsistency (Plake et al., 1991). For unscored multipoint items, judges must select item cut scores (e.g., 3 out of 4) corresponding to the achievement-level classifications.

6. Judges independently rate the importance, difficulty, or complexity of each objective/dimension.

Criteria: If the test is structured by objectives or other content categories/dimensions, an absolute, rather than relative, scale should be used to weight the objectives to allow any possible decision policy to emerge. A range of anchors for the importance scale should be presented for each objective, for example, "Essential," "Extremely Important," "Very Important," "Somewhat Important," "Not Very Important," and "Unimportant."
5. Judges are given both feedback on their individual and the panel's decisions as well as meaningful performance data AND requested to revise their initial decisions independently.

Criteria: Feedback should occur in the form of item distributions by achievement levels and estimated cut scores based on those distributions for each judge and across the panel. Performance data from a representative sample or appropriate subsample of examinees should include item difficulties, IRT scale item values if applicable, and estimated score distributions and cut scores based on the panel's decisions. In addition, the percentage of examinees at and above each cut score should be presented to communicate the potential impact or consequences of the cut-score decisions. Such information may help judges understand their classifications and ratings in terms of the results. If gaps exist between what they intended and the projected cut scores, the judges have the opportunity to eliminate those gaps by adjusting their ratings. (Note: Depending on whether a test-centered or examinee-centered approach is used, p-values or an IRT difficulty scale may be introduced to the judges at steps 7 or 8 when the items are matched to the achievement levels. There is a sparsity of evidence on whether such data are useful to judges earlier in the process [Atash et al., 1994].)

Once all of this information is presented to each judge on a microcomputer, if possible, he/she should be encouraged to revise the initial item or work-sample classifications and importance ratings. These individual decisions maintain confidentiality and privacy in the judgmental process.

4. Judges discuss their item or work-sample classifications and importance ratings without pressure to reach consensus.

Criteria: This open discussion provides the only opportunity for the judges to interact and process each other's explanations and justifications for their decisions. This step may help judges clarify discrepancies in their own classifications and ratings and affirm or disaffirm their decisions. (It should be noted that for this discussion session dress is usually casual, although a few judges have shown up wearing lederhosen.)
3. Judges render their final, independent revision of their classifications and ratings based on the discussion (step 4) as well as accumulated insights.

Criteria: As a follow-up to the discussion and all prior information, each judge should be able to privately make any final adjustments in his/her classifications and ratings without the direct influence of the panel. This step terminates the judges' involvement in the process. (Before they disband, throw them a going-away party or, depending on the time of the year, invite their families to your house for the holidays, buy gifts for their children, or send them on a cruise!)

2. Determine the judges' decision policy rule from the designated weights finally assigned to the objectives/dimensions (step 3).

Criteria: If the judges' median weights of importance across objectives do not yield a distinctive conjunctive or disjunctive pattern, then the compensatory model should serve as the default. Certainly all objectives rated as equally important would indicate a compensatory policy as well.

And the number 1 step to standard-setting salvation:

1. Compute the cut scores for the total test or by objective/dimension based on the decision policy rule (step 2) and the mean/median item or test scores from the judges' final classifications at each achievement level (step 3) in terms of the chosen test score scale.

Criteria:  
- **Compensatory policy**—Single or multiple cut scores for the total test are computed from the judges' mean/median scores.
- **Conjunctive policy**—Cut scores are computed from the judges' mean/median scores by objective.
- **Disjunctive policy**—Cut scores are computed from the judges' mean/median scores for the most important objectives and aggregated for the less important objectives.

This 10-step process is an iterative behavioral-anchoring approach. The unit of judgment may be the item in any format, a cluster of items, a work sample (such as a portfolio), or the total test. Judges are not requested to dream up item probabilities for any
hypothetical sample of examinees. They are asked to draw on their content expertise to provide behavioral descriptions of one or more achievement levels, select anchor items/work samples/tests for those levels, and then match the test items or the test itself to the levels. If the unit being matched has been scored previously (examinee centered), those scores represent cut scores corresponding to the achievement levels (in the minds of the judges). The mean/median score for the sample of judges at the end of the process is the cut score for a particular level.

If the unit of judgment has not been scored (test centered), either the point value of the item or separate cut scores on the multipoint items can be summed and averaged across judges to establish cut-score levels. Alternatively, IRT scaled scores based on the assignment of items or work samples to difficulty values (examinee centered) can be used to transform the judges' decisions into cut scores at the specific achievement level(s).

Decisions must be made on six issues in order to apply this generic method to a specific cut-score situation.

1. examinee target population (students, teachers, administrators, etc.)
2. unit of judgment (item, cluster of items, work sample, or total test)
3. item scoring format (dichotomous, polytomous, or a combination of both)
4. test-centered (unscored unit) or examinee-centered (previously scored unit or IRT scale) approach
5. number of achievement levels or cut scores
6. weighting of objectives for decision policy analysis

The two iterations at steps 3 and 5 are designed to refine the judges' decisions and improve the likelihood of a high degree of interjudge consistency at step 3.

**What's Wrong with this Standard-Setting Picture?**

So far, not a single quantitative index of reliability or estimate of misclassification error has been mentioned. Time's up! Reliability and validity evidence related to the judgmental process and the consequences of the decisions should be collected to determine whether the method works.
Reliability Evidence

There are three indices that can be computed to assess the degree to which the judges behaved themselves or, at least, behaved according to the psychometric expectations: (1) intrajudge reliability between steps, (2) intrajudge reliability within steps, and (3) interjudge reliability. Although these indices will be discussed separately here, the analysis of the different sources of measurement error can be conducted in one generalizability study, where the variance components for each source can be isolated and indices calculated for a given application.

Intrajudge reliability between steps. The iterative judgmental process recommended in the Top 10 Steps to Standard-Setting Salvation may yield high intrajudge variance across the steps (or occasions) inasmuch as the procedure encourages each judge to revise, refine, and reconsider his/her decisions based on new information (step 5) and input from other judges (step 4). This process affords the judges two chances to change their earlier item/work sample classifications and ratings of objectives. Variability across the three sets of decisions (steps 7, 5, and 3) can be expected to be relatively high if judges are integrating the information presented and making the appropriate adjustments in their ratings. Consequently, an estimate of intrajudge reliability across the steps of the judgmental process should yield a low coefficient.

Intrajudge reliability within steps. The estimate of intrajudge reliability that’s of greater concern is the consistency within steps 8, 7, 5, and 3. Key elements in the process are intended to maximize this type of consistency: (a) requesting judges to provide explicit behavioral descriptions of the achievement levels to improve clarity in interpretation (step 9); (b) training judges to match items/work samples/tests to achievement levels at acceptable levels of consistency and confidence (step 8); (c) using anchor items to improve the accuracy of item achievement level classifications (step 7); (d) providing judges feedback on their own decisions and performance data to adjust decisions to be consistent with realistic expectations (step 5); (e) permitting discussion of decisions among judges for clarification (step 4); and (f) iterations at steps 3 and 5 for judges to revise or refine their decisions. Other strategies to improve intrajudge consistency have been identified by Plake et al. (1991).
Reliability checks should be conducted during and after training (step 8), after the first formal item matching process (step 7), and after the final decisions are rendered (step 3). One cannot assume that high levels of consistency attained in the early stages will be sustained throughout the process. Estimates need to be computed at several checkpoints to monitor each judge's consistency.

**Interjudge reliability.** In contrast to the above sources of intra-judge variance, the estimate of interjudge reliability measures the degree of homogeneity or internal consistency of the final decisions by the judges (step 3). It is essentially an index of convergence. However, despite the iterations and specific steps in the process designed to increase the likelihood of convergence, interjudge consistency may be relatively low due to any number of factors (e.g., ambiguity in definitions of achievement levels, format of items or exercises, competence of judges, background characteristics of judges). Mean/median scores can still be translated into cut scores for the objectives and total test in the presence of considerable interjudge variability.

A generalizability coefficient computed from the among-judges-variance component would furnish evidence of the dependability of the cut score(s) generated from the specific sample of judges. It would indicate the extent to which the standard is replicable or generalizable to other samples. This is an important criterion for evaluating the success of the process.

These three measures of "reliability" reflect the characteristics of the process itself; that is, they are based on the internal mechanism for generating the standard. They should not be confused with the other indices of reliability reported in the literature (Berk, 1984) and required by Standards 2.10, 2.12, and 11.3, including decision consistency and standard errors of measurement, which are external to the judgmental process. Those indices are calculated from the score distributions of examinees on one or more occasions.

**Validity Evidence**

The internal validity of the process hinges on the qualifications of the judges and the procedure used to solicit their judgments. After digesting Berk's Top 10 Lists and scrutinizing every procedure that can be used in the standard-setting process, it can be concluded that the final standard is "whatever the judges say it is." This is certainly not a compelling argument for validity evidence, but the credibility of the group of content experts and procedural fidelity
are the only available internal criteria. What's missing in this rather circular thinking is an external criterion (Berk, personal communication while jogging, August 1994; I talked to myself).

**Consequential.** No matter how well the internal mechanism is polished, only external evidence can provide insights into the consequences of the decisions or whether the correct decisions were made. Step 1 of the first Top 10 List stated how consequences have been incorporated into standard setting. However, judges can be given the opportunity, as illustrated in step 3 of the process (Top 10 List Number Three), to consider the consequences of their decisions based on estimated cut-process and the percentages of examinees who attained them. Adjustments at this step make it possible to contextualize the judges' decisions in terms of consequences before the final standard is determined.

Any standard emerging from the judgmental process can be adjusted to account for a variety of what Messick (1989) refers to as "social consequences and side effects of the decisions" (p. 21). Evaluating these consequences, intended or unintended, of student and teacher testing is essential to the validation of the decision-making uses of test scores. Criterion measures of the political, economic, social, and/or educational outcomes of decisions about examinees must be obtained to determine whether or to what extent cut scores should be raised or lowered. This analysis of consequences should encompass all pertinent issues, such as school improvement and adverse impact.

**Evidential.** In addition to evidence related to the consequential basis of test use, the evidential basis of test interpretation and score use must be addressed. Specifically, evidence of the success of candidates who take the test can indicate the accuracy of the decisions made based on the cut score(s). Decision validity or accuracy (Hambleton, 1980) is the acid test of the worth of a standard-setting method. Many of the empirically based standard-setting methods of the 1970s and 1980s required the collection of this type of evidence in the form of traditional "hit rates" (a.k.a. probability of correct decisions) and false positive and false negative errors. Such evidence is also stipulated by Standard 1.24.

Decision validity evidence is based on predictive studies of the relationship between performance status on the test (e.g., proficient-basic) and actual performance in a subsequent position (Kane, 1982). For professional licensing and certification
examinations, defining the criterion of success and measuring it without bias or distortion have been the bane of most test-criterion relationships. Messick (1989) has identified numerous sources of bias that can "operate differentially on the three major forms that criterion measures typically take: namely job samples, supervisors' ratings, and production records in industrial settings or achievement tests, teacher judgments, and career data in the case of educational outcomes" (p. 73).

Despite these psychometric limitations and the real-world roadblocks to gathering predictive evidence, a commitment to investigate the accuracy and meaningfulness of the decisions made on the basis of cut scores is long overdue. Achievement levels at the upper grade levels, in particular, can be linked to entry-level job performance data and success in college. Linkage models developed in military applications (Wise, 1994) can be adapted to large-scale assessment standards. Studies to determine the relationships between achievement levels in different content areas and the skills from a representative sample of jobs should be conducted to evaluate some of the consequences of pass-fail decisions.

An Afterthought or Ten

What implications do the preceding Top 10 Lists and challenges to collecting reliability and validity evidence have for future standard-setting practices? I don't have a clue! (Just kidding!) Actually, we've probably learned more about what doesn't work than what does. Although you probably don't want to read anything that's packaged in "10s" by now, my summary of the main points of this chapter is an awards presentation, expressed in (what else?):

A Top 10 MOST Awards List

10. Most Promising Old Approach: Multistage Iterative Process
8. Most Politically Correct Procedure: Selecting a Broad-Based Panel of Judges
7. Most Politically Incorrect Procedure: Using Only Content Experts to Set Standards
6. Most Confusing New Term: Polychotomous or Polytomous
5. **Most Psychometrically Incorrect Procedure:** Asking Unqualified Judges to Make Uninformed Decisions

4. **Most Challenging Old Procedure:** Consensus Building Among Judges

3. **Most Challenging New Complication:** Setting Multiple Cut Scores

2. **Most Neglected Technical Topic:** Evidential (Predictive) Validity

1. **Most Difficult to Defend:** All of the Above!

As Sergeant Phil Esterhaus of *Hill Street Blues* used to say to his officers every morning, "Hey, let's be careful out there!"

**References**


Acknowledgment

The ideas expressed in this chapter were partially funded by Moisha & Izzy's Bagel Shop (No. POPPYSEED 12-94); the rest of the dough was donated by the National Society for the Prevention of Cruelty to Psychometricians.

The author gratefully acknowledges the assistance of Ellen Spies in the preparation of this chapter and all of the materials for the presentations upon which the chapter was based.