



cde

Colorado Department of Education

Standard-Setting Procedures for the Specification of Performance Levels on a Standards-Based Assessment

by

*Vonda L. Kiplinger, Ph.D.
Assessment Unit
Colorado Department of Education*

October 14, 1997

In order to determine performance levels, cut points must be established that reflect what students in each performance level should know and be able to do. In other words, cut points are set such that students whose performance exceeds the particular cut point are inferred to possess sufficient knowledge, abilities, and skills to be regarded as performing at that level relative to the academic achievement standards. Several standard-setting methods are described below.

Overview of Standard-Setting Methods for Setting Performance Levels

Currently, the method most commonly used for setting performance level cut points (i.e., standard-setting) for large-scale assessments is the modified Angoff method (see Angoff, 1971). Another method is the Bookmark procedure, which -- unlike the modified Angoff method -- handles constructed-response (polytomously scored) items and selected-response (dichotomously scored) items equally well (Lewis, Mitzel, & Green, 1996). Although a recent development, this procedure is being well-received in the assessment and psychometric communities.

Several other standard-setting models have been used in varying degrees in the past. These methods, including the Ebel's, Jaeger's, and Nedelsky's procedures, are not as commonly used as the modified Angoff methods. All of these methods may be termed "test-centered" models since they focus on test items or other elements of the assessment. Another class of standard-setting methods are "examinee-centered" in that the judgments focus on examinee qualifications. The two most frequently used standard-setting methods in this group are the borderline-group and contrasting-groups procedures. These methods share several characteristics, including disadvantages, of the modified Angoff method; therefore, they will not be discussed here. The reader is referred to Jaeger (1989) for a discussion of these and other standard-setting models.

Because of the disadvantages and problems associated with most of the older methods, only the popular modified Angoff method and the Bookmarking procedure were considered as appropriate strategies for the setting of performance levels for the Colorado Student Assessment Program (CSAP). Both of these methods are inherently judgmental, as are the older methods; however, only the Bookmark procedure is based on empirical data (i.e., actual student work). In the following section, these two methods are briefly described, followed by discussion of the advantages and problematic issues and concerns relating to each method.

Modified Angoff Method

In the modified Angoff method, the panelists usually begin by drafting descriptions of achievement levels. The panelists then examine the test and may even take the test. In the first round, panelists judge each test item and give their estimate of the difficulty level of each item. Each judge is asked to look at one item at a time and give his or her best guess of the *proportion* of a hypothetical group of borderline, "minimally acceptable" students that he or she would *expect* to answer the item correctly. In the second round, the individual judgments from the first round are discussed among the panelists, and participants may revise their original item difficulty judgments. A third round of item difficulty judgments is included in attempts to produce concurrence, or at least convergence, of the judgements. Actual student performance on the test items (generally, the proportion of students who answered the item correctly, called "p-values") may be introduced at this point or after the second round judgments. In either case, the third round usually is supposed to produce approval or sanction of the standards set by the stakeholder panelists (which typically include non-expert parents and community members, as well as teachers and subject matter experts). According to a National Academy of Education report (Shepard et al, 1993), these item-by-item judgments often are difficult to make and are highly dependent on the particular sample of judges selected.

The final step in the modified Angoff procedure is to aggregate the judgmental p-values for each item; disagreement among the judges is usually "resolved" by applying some type of averaging technique. Note that since p-values for each item are being estimated, the modified Angoff method is appropriate for selected-response (multiple choice) items *only*. Even proponents of this method recognize that a different procedure must be used for assessments, such as the Colorado Assessments of Reading and Writing, that contain constructed-response (open-ended) items (cf. Mehrens, 1994). Advantages and disadvantages of this method are presented following the description of the Bookmark procedure.

Bookmark Procedure

A different procedure, which involves "behavioral anchoring" and "item mapping" using actual assessment results and which handles selected-response and constructed-response items equally well, is being used by a number of assessment programs, including NAEP in their Anchor Levels procedure as a check on the 1992 standard-setting (Lissitz & Bourque, 1995) and in Maryland to set standards for the Maryland State Performance Assessment Program. Other states and jurisdictions that have used this method to set performance standards include Colorado; Florida; Hawaii; Indiana; New Mexico; Missouri; Wisconsin; Douglas County School District, CO; and Sacramento City Unified School District, CA.

The Bookmark procedure is described below in somewhat more detail than the modified Angoff method since it is newer and not as well-known. This procedure is whole-test-based, rather than item-based, since it is derived from Item Response Theory (IRT) item mapping results. Item maps are the locations of test items on the IRT ability scale. The steps in this method of standard-setting require actual student results on either the item pool or the test forms.

Using an IRT procedure that accounts for item discrimination and student guessing and places item locations (difficulties) and student abilities on the same continuum, all items are placed on a common scale. Ordered item booklets are then produced in which the test items are rearranged in order of difficulty, from easiest to hardest. Constructed-response items are included on the same scale as the selected-response items. To accomplish this, each score point for a constructed-response item has its own location on the scale and in the ordered booklet. Thus, a zero-to-three-point constructed-response item would be represented three times on the ability scale and in the booklet. A score point of one would be located closer to the beginning of the booklet than a score point of two, while the score point of three would be located further into the booklet since it is harder to get a three than a two. A zero score is not included since that typically signifies a non-scorable response.

Usually, panels of 15 to 21 judges for each grade and subject area are selected. The panelists should be primarily master teachers and curriculum specialists in terms of subject area content and general knowledge of the students taking the assessment. Panels should contain master teachers of the grade level and master teachers of one grade level higher. Members of the community also may participate in these panels; however, the ratio of educators-to-community participants should be a minimum of two-to-one.

The panelists take the assessment to increase their understanding of what the test measures. The panelists are then given ordered item booklets, item maps, strand maps, the operational tests, scoring rubrics for the constructed-response items, and the specific content standards on which the assessment is based. In addition, it is extremely helpful if panelists are given some preliminary descriptions of the kinds of skills, knowledge, and abilities expected at each performance level. It must be emphasized that these preliminary descriptions are just that -- preliminary. They are based on expectations only, not on actual student performance, and serve as starting points from which judgments based on empirical results proceed. The typical process of this procedure is outlined below.

The large group of 15 - 21 panelists is divided into smaller groups (typically, three) who will make their judgements during three rounds independently. After first studying the ordered item booklet, each panelist makes independent judgments regarding expectations of skills and knowledge that students had to demonstrate in order to be classified as attaining each specified performance level relative to the specific items on the assessment. These judgments are operationalized by the panelists placing "bookmarks" in their ordered item booklets that "separate" the partially proficient student from the proficient student, the student whose performance is unsatisfactory from the partially proficient student, and the advanced student from the proficient student.

In the Colorado standard-setting, four proficiency levels were expected. Panelists were instructed to first place their markers in the following order: Proficient, Advanced, and Partially Proficient. This is because a "proficient" student is usually considered as one who meets the standard, and thus, may be considered the "anchor". Each panelist first places the "Proficient" bookmark at the first point in the ordered item booklet where he or she feels that a student who had a high probability of success in responding correctly to the items up to that point will have demonstrated sufficient skills to infer that the student has mastered the Model Content Standards in the area assessed. Items preceding the bookmark are items reflecting content that *all* proficient students should be likely to know and be able to do. Thus, the Proficient bookmark is placed immediately *after* the last item *all* proficient students should be expected to know. This bookmark represents the *top* of the category of items that students should be able to respond to correctly in order to be classified as proficient. More intuitively, perhaps, if one is describing students, this marker defines the *bottom* of the category of proficient students. The expectation is that a "*just barely proficient*" student would be likely to respond correctly to all questions prior to the marker, though not necessarily to the items beyond the marker. Academically strong students should be able to respond correctly to some of the items beyond the Proficient bookmark, but not in sufficient quantity or quality to "boost" them into the Advanced performance level.

The Advanced and Partially Proficient bookmarks are placed in an analogous manner. Thus, the "Advanced" bookmark separates the "proficient" student from the "advanced student; the Proficient bookmark separates the "partially proficient" student from the proficient one; and the final bookmark placed, the "Partially Proficient" bookmark, defines the minimal performance of a student who would be considered performing at a partially proficient level. If a student is not likely to respond correctly to these items, s/he probably is performing at the "Unsatisfactory" level.

When the bookmarks were placed by Colorado educators, the expectations of student performance may be described as follows:

The items prior to the first bookmark (Partially Proficient) are items that all partially proficient students, with a high probability, should know and be able to do. Students who are not likely to respond correctly to these items are probably performing at the "Unsatisfactory" level. Mastery of these items separates the "partially proficient" student from one who is performing unsatisfactorily.

The items between the first bookmark and the second bookmark (Proficient) are items that all proficient students, with a high probability, should know and be able to do. Students who are not likely to respond correctly to these items are probably "partially proficient" or "unsatisfactory". Mastery of these items separates the "proficient" student from the "partially proficient" student.

The items between the second bookmark and the third bookmark (Advanced) are items that all advanced students, with a high probability, should know and be able to do. Students who are not likely to respond correctly to these items are probably performing at the "proficient," "partially proficient," or "unsatisfactory" level. Mastery of these items separates the "advanced" student from the "proficient" student.

The items after the last bookmark (Advanced) are items that some, but not all, advanced students are likely to know and be able to do.

In Round 2 of the Bookmark procedure, panelists in each group discuss the reasons for their individual selections of bookmarks and may adjust their selections following the discussion. In Round 3, the results of the Round 2 small group judgments, as well as overall impact data (i.e., the proportions of the total student group who would be classified into each of the performance levels -- unsatisfactory, partially proficient, proficient, and advanced -- according to the selected bookmarks), are presented to the entire group and final judgments are made. If group consensus is not reached, the median score at each cut point is used.

The final step in the Bookmark procedure is to develop descriptions of each performance level that reflect what students know and are able to do. By this time, the panelists are intimately familiar with the academic constructs, knowledge, skills, and abilities being measured by the assessment. Based on this knowledge of the skills being measured, the panelists then write descriptors of each performance level that reflect what student should and can do relative to the standards. Thus, the Bookmark procedure facilitates the writing of more valid performance level descriptions since the panelists also write the final performance level descriptions based on a synthesis of the empirical information on actual student performance.

A major strength of the Bookmark method is that it is based on actual student performance on a "whole task" (i.e., a test). It asks for categorical level judgments that are much easier to make than the modified Angoff's *a priori* point estimates of item difficulty levels for each item on the test in isolation. *The combination of empirical results and expert judgment provides a compelling argument for the intuitive usefulness of the Bookmark procedure.* As described above, the panelists are asked to simply place "bookmarks" in an ordered item booklet at the point where they feel the difficulty of the items surpasses the level of what they would reasonably expect a student at that ability level to know and be able to do. Thus, the bookmarks are placed in a way that directly reflects expected student performance, conditioned on actual performance. Use of the item maps, ordered test booklets, and samples of student work provide a basis for panelists to gain an understanding of the overall trait and individual strands being measured by the test. This method also can be a tool for evaluating what the test actually measures, whether the test is aligned with the state's academic achievement standards, and the comparability of "equivalent forms". In other words, this procedure gives information on what the test actually measures, what is missing from the test in terms of measuring standards, and what is included on the test that shouldn't be there (e.g., poor performing items).

Advantages and problematic issues or concerns relating to the modified Angoff and the Bookmark procedures are discussed in the following section.