

Chapter 9

A Mapmark Method of Standard Setting as Implemented for the National Assessment Governing Board

E. Matthew Schulz

Howard C. Mitzel

Pacific Metrics

Introduction

In the course of a project to help the National Assessment Governing Board (NAGB) set achievement levels for the 2005 National Assessment of Educational Progress (NAEP) in Grade twelve mathematics, ACT and its subcontractor, Pacific Metrics, developed an enhanced bookmark method called Mapmark. Mapmark was named for the role of item maps (Masters, Adams, and Loken, 1994) and significant elements of the bookmark standard setting method (Lewis, Mitzel, and Green, 1996; Mitzel, Lewis, Patz, and Green, 2001) in its process. Over the course of the 2005 project and a more recent contract to set achievement levels for the 2006 NAEP in Grade 12 Economics, two additional features have been identified as key

components of the Mapmark method: holistic feedback and independent judgment by panelists. The following section presents background and rationale for these key features of Mapmark.

The Bookmark Method

Until recently, standard setting panelists were required to provide item ratings in order to derive cut scores from the standard setting process. For example, panelists were required to estimate for each item, the proportion of students at the cut score who would answer the item correctly. If the item was polytomously-scored, panelists would estimate the average score that such students would obtain on the item (Loomis and Bourque, 2001). Obtaining cut scores from item ratings can be a simple matter. No test data is required. The cut score is the sum of the item ratings. However, the task of producing item ratings is considered by many experts to be too cognitively complex to be meaningful (Shepard, 1994). Also, the rise of item response theory in testing has made the estimation of cut scores from item ratings a considerably more complex and computer-intensive proposition (e.g., Chen, Loomis, and Fisher, 2000).

Rasch models and item response theory (IRT) models in general allow effective alternatives to item rating methods. When test data is analyzed with an IRT model, items can be placed directly on the scale of student achievement. An item's location on the scale corresponds to a certain probability—called a response probability (RP) criterion—that a student at the same location would answer the item correctly. When the items are assigned scale values using an RP criterion that is sufficiently high to connote “mastery” of the item's content, and the items are presented to panelists in order of their scale value from easiest to hardest, the standard setting panelist's task can be simplified to dividing the items into two groups—those representing content that a student at the cut score should have mastery of, and those too difficult for this expectation. The cut score is the scale value of the hardest item in the mastery group. This is essentially the bookmark method of standard setting.

Developers of the bookmark method (Lewis, et al., 1996; Mitzel, et al., 2001) argue that the task of a standard setting panelist should be focused primarily on the content dimensions of the assessment rather than on estimating how students would perform on individual test items. This argument has been made independently by Stone (2001) in the context of Rasch measurement. Bookmark developers also hold that the semantic component of the panelists' task, i.e., “mastery” is the most meaningful component.

Panelists need to be informed about the RP value, but “mastery,” not the RP value is the concept that should drive their cut score decision. The choice of RP-value should therefore be a policy decision, made in advance of the standard setting meeting. An RP value of 0.67 or two-thirds chance is most commonly used, but other values, such as 0.5, have also been used. ACT investigated different RP criteria (0.67 and 0.5) and task statements (mastery versus “can do”) prior to NAGB’s selecting 0.67 and “mastery” for use in Mapmark (Williams and Schulz, 2005).

No particular IRT model is specified in the bookmark procedure. In ACT’s work for NAGB, a mixture of IRT models was used: 3-PL for multiple choice items, 2-PL for dichotomously scored constructed response items, and the generalized partial credit model for polytomously-scored items. These models were used by NAGB’s contractor for test development and scale construction and therefore had to be used for standard setting.

Choice of IRT model and RP criterion is acknowledged to affect the ordering of items in a bookmark method (Mitzel, et al., 2001). ACT investigated the effects of different RP criteria on item order with the use of NAGB’s IRT models and found it to be small and to have no effect on panelists’ perceptions (Williams and Schulz, 2005). Item order using a 0.5 RP value correlated 0.97 to 0.98 with item order using a 0.67 RP value. Panelists generally perceive some items as being out of order in their Ordered Item Book (OIB) (Mitzel, et al., 2001). In ACT’s research, panelists’ perceptions in this regard did not differ by RP value (Williams and Schulz, 2005).

Item Maps

An item map is a spatially representative display of items on a student achievement scale. Distance from one point to another simplifies probability information in the same way the use of the term “mastery” simplifies a judgment related to probability in the bookmark method. When a student is “even” with the item, the student has a particular probability of success on the item such as 0.5 or 0.67. The probability decreases the farther ‘ahead’ of the student an item is, and increases the farther behind the student the item is.

Although an “item map” is described in the bookmark method (Mitzel, et al., 2001), the map is not spatially representative. It is a list that provides content and statistical information about the items in the same order the items are presented in the OIB (and would appear on an item map from

easy to hard). In this chapter, the term item map will mean a spatially-representative display unless otherwise noted. The distance between items on the map is spatially representative of how much the items differ in difficulty. The distance between an item and an achievement level boundary on an item map is spatially representative of how hard or how easy the skill represented by the item is likely to be for a student at the borderline of the achievement level.

Rasch measurement practitioners have been ahead of the general measurement community in recognizing the usefulness of item maps for explaining test results to parents, teachers, and the general public (Stone, Wright, and Stenner, 1999). In 1992, reports for parents and teachers based on Rasch-model variable maps (Masters, et al., 1994) won an award from the National Council on Measurement in Education (NCME) for the Dissemination of Measurement Concepts to the General Public. Now item maps are used routinely in NAEP reports (e.g., National Center for Education Statistics, 2007).

In standard setting too, Rasch measurement practitioners have been ahead of the general education community in using item maps. One of the first uses of an item map in standard setting was by Grosse and Wright (1986). Rasch model item maps were used in standard setting by Engelhard and Gordon (2000), Stone (2001), and Wang (2003). Shen (2001) had panelists study a set of items and their location on an item map before selecting a scale value to represent the passing standard. Panelists simply drew a “mark” on the item map to represent the passing standard.

Given the wide-spread use of item maps and the nature of standard setting, it is remarkable that item maps have not been used more widely in standard setting. Standard setting is a highly complex process involving potentially many types of statistical information. The importance of graphical displays for representing complex statistical information is widely recognized (Larkin, 1987; Tufte, 2001; Wainer, 1992). Wainer (1992) commented that the field of education is lagging behind scientific disciplines in its use of effective graphics.

Item maps were proposed for NAEP standard setting partly as a matter of procedural validity. Given the use of item maps in NAEP reports, item maps should be used in standard setting. If the item-map information given to the general public to explain the achievement levels is not what the standard setting panelists actually saw, one could question whether the panelists really understood the meaning of their cut scores relative to the

items on the achievement scale and if they would have set the same cut scores were the item map provided to them.

As suggested by the work of Shen (2001), item maps were expected to be a particularly useful and natural extension of the bookmark method. In the bookmark method, panelists spend a great deal of time identifying the knowledge, skills, and abilities (KSAs) required by test items in the context of the OIB—thinking about what KSAs are required by harder items that are not required by easier items representing similar content. This is called the KSA review. The distance information on the item map helps panelists “see” how much more difficult one item is than another and therefore to appreciate the amount of growth, or lack thereof, that separates one item from another. When setting multiple cut scores, it is also reasonable for panelists to consider “how much” growth separates one cut score from another. The distance between cut scores on an item map is more informative and easily seen than the physical separation of multiple bookmarks in an OIB.

Holistic feedback

The use of holistic feedback in Mapmark has roots in the popular belief that panelists should be given multiple perspectives on their cut score recommendations. This notion developed in response to the criticism that different test-centered, criterion-referenced standard setting procedures produce different cut scores (Glass, 1978). Counterarguments pointed to differences between such methods on a range of issues including perspectives on student performance and minimal competence, and argued that these differences were legitimate and defensible (Hambleton, 1978; Popham, 1978). In item rating methods, panelists are primarily concerned with student performance on an item-by-item basis. A more holistic perspective involves student performance on a larger set of test items, such as on an entire test booklet. Results that are synthesized from multiple methods or perspectives are generally considered more defensible (Green, Trimble, and Lewis, 2003). In an Angoff-based method used to set achievement levels for the 1998 NAEP in Civics (Loomis and Hanick, 2000), panelists were shown “whole booklet” feedback—completed booklets where the total score on the booklet translated to achievement at or near the cut score.

Two kinds of holistic feedback have been used in Mapmark. In the 2005 Project (setting achievement levels for the 2005 NAEP in Grade 12 mathematics), “domain score” feedback was used. The use of domain score feedback is described in detail in this chapter. As will be seen, domain score feedback does not involve examples of student work. Domain scores are

expected scores on subsets of items representing more specific areas of content. The domain scores are conditional on achievement scale scores. In a later project to set achievement levels for the 2006 NAEP in Grade 12 economics (the 2007 Project), whole-booklet feedback and domain-score feedback was used. The delivery of both types of feedback in Mapmark methods are described in available reports (ACT, 2005a, 2007a).

The rationale for using domain-score feedback is developed in a paper by Schulz, Lee, and Mullen (2005) (see also Schulz, Kolen, and Nicewander, 1999). This chapter addresses the general problem of supporting criterion-referenced inferences in large scale testing by describing student performance on individual test items (Forsyth, 1991). Panelists in a bookmark method, even one enhanced with an item map, could be misled into thinking that a student at the cut score has mastered a particular skill because an item representing that skill is below their cut score. Domain scores are used in Mapmark to provide a more systematic and reliable basis for inferences than item scores.

Table 1 shows the titles of the “teacher” domains that were ultimately used in the 2005 Mapmark project. The domain development process is described in a publicly-available report delivered to NAGB (ACT, 2005d). The process was similar to the one described for the Grade 8 mathematics NAEP (Schulz, et al., 2005). Multiple “teacher domains” were defined within each content area of the framework. There were four content areas at Grade 12: a) Number Properties and Operations, 2) Measurement and Geometry, 3) Data Analysis, and 4) Algebra. The domains were intended to cover a wide range of difficulty—there should be an easy domain and a hard domain in each content area if possible. There was no limit on the number of teacher domains per content area other than each domain should contain at least three items and the domains as a group should exhibit coherence. Coherence was assessed by the reliability with which teachers could independently classify items into the domains using only domain definitions such as the one shown in Figure 1. This figure shows the domain definition for teacher domain M4 in the Measurement and Geometry content area. A domain definition was developed for each teacher domain. The sample items in the domain definitions were released items on the NAEP website. A total of twenty-three teacher domains were defined.

For use in standard setting, it was desirable to have perhaps fewer domains within each content area, with the domains being distinct in both difficulty and content and still covering a wide range (Schulz, et al., 2005). A smaller number of “score domains” were thus defined. Some teacher

domains were large and distinct enough in difficulty to stand on their own as score domains. Other teacher domains were combined, on the basis of both content and difficulty, into a score domain. The relationship of teacher domains to score domains is shown in Table 1. There were sixteen score domains.

Domains were also developed for the Grade 12 economics assessment (ACT, 2007a). The development process differed somewhat from the process used for mathematics. The domains correspond closely to standards in the economics framework (National Assessment Governing Board, 2006). In a pilot study, Mapmark with domains and Mapmark with whole booklet feedback were both implemented, yielded statistically identical results, and were both judged to be acceptable for use in the ALS meeting (ACT, 2007a). NAGB elected to use whole booklet feedback in the operational

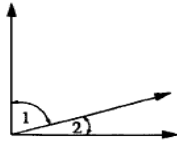
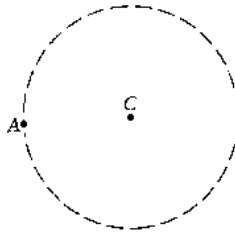
Table 1

Titles of Teacher Domains and the Correspondence between Teacher and Score Domains by Subscale of the 2005 Assessment

Teacher Domain	Teacher Domain Title	Score Domain
Number Properties and Operations		
N1	Perform Basic Operations	N—1
N2	Determine Correct Operations	N—2
N3	Place Value and Notation	N—3
N4	Multistep Problems	N—4
Measurement/Geometry		
M1	Basic Measurement	M—1
M2	Symmetry, Motion, and Proportionality	M—2
M3	Identifying Geometric Objects	
M4	Angles	M—3
M5	Perimeter, Area, and Volume	
M6	Coordinates and Their Applications	M—4
M7	Triangle Properties and Measurements	
M8	Geometric Relationships	M—5
Data Analysis		
D1	Common Data Displays	D—1
D2	Elementary Probability and Sampling	D—2
D3	Central Tendency	D—3
D4	Advanced Data Displays	
D5	Abstract Reasoning	D—4
Algebra		
A1	Reading Tables and Graphs	A—1
A2	Algebraic Expressions, Equations, and Inequalities	
A3	Systems of Equations	A—2
A4	Slope and Rates	
A5	Creating and Recognizing Expressions	A—3
A6	Advanced Functions and Concepts	

Teacher Domain M4. Angles: Items in this domain involve obtaining degree measures direct measurement or through knowledge about degree measures, such as the sum of angles in triangles or regular polygons, or the properties of angles formed by intersecting lines. Students are required to use rulers or protractors to draw figures having specified shapes or measurements.

6. On the circle with center C shown below, use the protractor to locate and label a point B that creates an arc AB with measure 235° . Darken this arc.

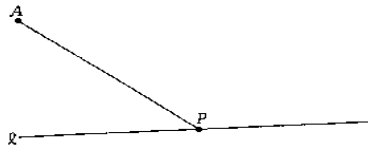


33. The sum of the measures of angles 1 and 2 in the figure above is 90° . What is the measure of the angle formed by the bisectors of these two angles?

A) 60° B) 45° C) 30° D) 20° E) 15°

Key: B

27. In the figure below, use the protractor to draw a line m through point P perpendicular to segment AP . In the answer space provided, give the measure of the smaller angle formed by lines l and m .



Answer: _____

Figure 1. Example of a Domain Definition.

standard setting meeting since whole booklets exist for all subject areas, but domains do not (this also means that whole booklets are more familiar to teachers and that whole booklet feedback is less costly to provide), and NAGB wished to use the same form of holistic feedback for both present and future standard settings. Although domains had been successfully developed in both mathematics and economics, there was concern that other subject areas such as Reading might not be as amenable to domain development.

In order to facilitate the use of domains in Mapmark, items on the item maps were organized into columns. Each column represented a teacher domain. An example of this kind of map is the “Domain Item Map” in Figure 12. A Domain Item Map was created for each content area. Domain Item Maps were used to illustrate concepts and to facilitate panelists’ tasks with domain score feedback as described later in this chapter.

The principal of organizing items into columns of more specific content was adopted more generally in the Mapmark method by organizing items on the “Primary Item Map” into columns representing the content areas in the assessment framework. The Primary Item Map for the Grade 12 Mathematics assessment is shown in Figure 2. Its construction and use is described in detail later in this chapter.

Independent Judgment

Independent judgment is gaining in practice with the use of bookmark and related methods (e.g., Buckendahl, Smith, Impara, and Plake, 2002) but its importance has not been stressed as strongly as it might be for standard setting in general. This is partly because consensus plays different roles and carries different risks in different methods. In item rating methods, panelists need a detailed and coherent concept of minimal competence *prior* to rating items (Loomis and Bourque, 2001). It makes sense to ask panelists to work together to develop this concept. In the bookmark method, the KSA review enables each panelist to develop and specify a coherent concept of minimal competence *on their own* in the process of placing their bookmark. The bookmark placement represents a panelists’ concept of minimal competence directly. Since it is so easy for panelists to see each others bookmark placement, the bookmark process magnifies the general risk that some panelists will unduly influence others.

The value of independent judgment is specifically addressed in the Mapmark process. Panelists are told that standard setting is a process

much like others described in the book, *The Wisdom of Crowds* (Surowiecki, 2004). Certain decisions and predictions in a wide array of human activities—from economic forecasting to predicting the number of beans in a jar—benefit from group diversity and independence among individuals within the group. As long as there is no systematic bias due to human nature in the basic judgment or prediction being made, individual “errors” tend to cancel out and the group average tends to be consistently superior to the judgment of even the most experienced or expert individual.

Another idea from *Wisdom of Crowds* that is presented to panelists is that they should feel free to use information in the process selectively. Mapmark panelists are given permission to use primarily the OIB, holistic feedback, the spatial information on the item maps, or any other information they feel is important and which they understand. They are also given permission to ignore information that they do not understand. If each person bases his or her judgment on the information that they understand best, the average of their independent judgments will be the best possible outcome. Of course, every effort must be made to present standard setting information clearly and to avoid information that is inherently confusing.

Independent judgment in Mapmark is encouraged also because the mean or median of a distribution of independent bookmark placements has better statistical qualities than a single bookmark placement reached by consensus. This is largely due to the presence of gaps between adjacent items’ scale values. The mean of a distribution of discrete values can take on any intermediate value. Even then, the mean can be less reliable and accurate if it is unduly influenced by one panelist.

An Achievement Level Setting Meeting using Mapmark with Domains

This section provides a relatively detailed description of the Mapmark method as implemented in the ALS meeting for the 2005 NAEP in Grade 12 mathematics. More complete details and results are provided in available documentation (ACT, Inc., 2005a, b, c, d). The ALS meeting lasted four days. Sessions generally started at 8:00 AM or 8:30 AM and lasted until 5:00 PM or 6:00 PM, except the last day, which adjourned at 12:30 PM.

NAGB Policy

NAGB policy requires achievement level descriptions (ALDs) to be developed prior to the standard setting meeting. ALDs are content-specific and grade-specific statements of what students should know and be able

to do at each achievement level. These are developed from more general “NAGB Policy Definitions” of the achievement levels. NAGB developed the ALDs in advance of the ALS meeting in a process described in the Process Report (ACT, 2005a). The ALDs developed for the 2005 NAEP in Grade 12 mathematics are published in *The Nation’s Report Card* (National Center for Education Statistics, 2007).

Panelists

Thirty-one panelists participated in the ALS meeting. The percentage of panelists by type were very close to targeted percentages of 55% teachers, 15% non-teacher educators, and 30% general public. The ALS panelists were nationally recruited. Panelists came from a total of 23 states. Thirty percent of the panelists belonged to an ethnic minority group (Black, Hispanic, or Asian). Forty-two percent were female.

Staffing

The ALS meeting was staffed by two process facilitators, two content facilitators, a psychometrician, a meeting manager, and a data entry/clerical staff person. Also attending the meeting at various times were members of ACT’s technical advisory committee, members of the National Assessment Governing Board (NAGB), technical consultants to NAGB, and ACT managers. One of the content facilitators had served on the committee to develop the assessment framework. The other content facilitator was a mathematics content expert on the staff of Pacific Metrics and had participated in domain development for the project. One of the process facilitators was an employee of Pacific Metrics and had experience with the bookmark method. The other facilitator was an ACT employee and had experience with item maps and domain development.

Meeting Design and Preparation

All technical procedures, meeting materials, and design aspects of the ALS meeting are documented in a Technical Report (ACT, 2005b). All items in the 2005 assessment were used in the meeting. There were 180 items of which 119 were multiple choice (M), 24 were dichotomously-scored constructed response (D), and 37 were polytomously-scored constructed response (P). The polytomously-scored items represented a total of 95 score points, or 40% of the points in the item pool. For the meeting, the items were divided into two overlapping pools (A and B). Each pool contained about 60% of the items. The pools were equivalent in key respects such

as difficulty, content, and even the dispersion of item difficulty within content area.

Panelists were assigned to one of two Groups (A and B). There were 15 panelists in Group A and 16 panelists in Group B. Groups A and B used item pools A and B respectively. There were three tables per group. Care was taken to balance tables and groups with regard to panelist type, gender, region, and race/ethnicity.

Materials were created using the scaling procedures of NAGB's contractor. The achievement scale is a linear composite of four unidimensional subscales, one for each content area. Item parameters corresponding to subscales had been estimated by the contractor using 3-PL, 2-PL, and generalized partial credit models in a multidimensional framework. Items had been calibrated using field test data collected in the spring of 2004. The item parameters and regression methods based on subscale covariances and scaling constants, were used to map item characteristic curves to the composite score scale. A 0.67 RP value was used to construct the item maps. Domain characteristic curves were computed as described in Schulz, et al. (2005). To disguise the actual scale used for NAEP reporting, the achievement scale used in the ALS meeting was a linear transformation of the reporting scale.

Panelist Orientation

Materials mailed to panelists before the ALS meeting included a briefing booklet, the framework for the assessment, and the ALDs. The briefing booklet described the tasks and materials of the ALS meeting. At the beginning of the meeting, panelists were given a presentation on NAEP and NAGB by a NAGB staff member. They were also given an overview of NAEP achievement level setting in general. They took a form of the NAEP exam. Panelists were then given an overview of the Mapmark method. The overview explained item maps, domains, and the contents of the OIB. Panelists were informed of the RP value underlying the term "mastery" and item positions on the maps.

Figure 2 shows sections of the "Primary Item Map" that was used in the orientation and throughout the meeting. Items are represented on the map by a handle consisting of a character followed by a number. The character indicates item type (P = polytomously-scored, D = dichotomously-scored constructed response, and M = multiple choice). The number indicates the easiness rank of the item (1 = easiest within item type). Handles for



Figure 2. Primary Item Map on which score levels for polytomously-scored item P6 (P6_1 and P6_2) are marked by circles.

polytomously-scored items include an underline ‘_’ followed by the score level. Polytomously-scored items were ordered by the scale value of their last score level. The scale value of a score level was the point where a student would have a 0.67 chance of scoring at that level or higher.

Circles on the map in Figure 2 show the score locations of a two-point polytomously-scored item, P6. It can be seen that P6 is an item in the Algebra and Functions content strand, that the scale value of the first score point, P6_1, is in the map score interval whose midpoint is 252, and that the scale value of the second score point, P6_2, is in the interval whose midpoint is 279. Score intervals on the item map were three points wide.

The reader may be able to see that the item handles on the item maps are shaded differently. The different shading is due to the use of different colors on the map, which do not show in the reproduction here. Items exclusively in Pool A were tan, items exclusively in Pool B were green, and common items were yellow. Common items, such as item P6 were discussed in “whole group” activities.

Round 1

KSA review

Round 1 began with a 40-minute presentation on the NAEP framework by the NAEP content facilitator. The purpose of this review was to orient panelists to the knowledge, skills, and abilities that the framework is intended to represent and to the specific terminology used. Following this presentation, panelists spent the next nine hours of meeting time identifying the KSAs required by the test items in the context of the Ordered Item Book (OIB) and Primary Item Map. The KSA review was divided into four parts. Each part is described briefly below, followed by a detailed description of materials.

KSA activity 1. The bookmark content facilitator led the entire panel (whole group) through a process of identifying the KSAs in the common constructed response items. They began with common dichotomously-scored items, then moved to items with more than two score levels (polytomously-scored items). For each polytomously-scored item, the activity involved identifying the *additional* KSAs needed to earn successively higher scores on the item.

KSA activity 2. Panelists continued to the review constructed response items, but now the items were group-specific (A or B) and the review was conducted in table groups. Panelists took turns leading this activity at their table. Content and process facilitators circulated among the tables to keep the process on time and to answer questions.

KSA activity 3. In this activity, each panelist reviewed all items in his/her pool independently using the OIB and the Primary Item Map. Items were reviewed sequentially in the OIB, beginning with the first, or easiest item. An important part of this task was to think about the additional KSAs that an item might require that were not required by earlier, easier items representing similar content.

KSA activity 4. This was basically a table-group repetition of KSA activity 3. Panelists were asked to share their ideas about the KSAs and to record other panelists KSAs if they agreed with them.

Materials. Materials for KSA Activities 1 and 2 were the Constructed Response Ordered Item Book (CROIB) and a Note-template. The CROIB contained all the polytomously-scored items in a Group item pool, plus the common dichotomously scored (constructed response) items. The

dichotomously-scored items appeared first in the booklet, and were the first items covered in KSA Activity 1.

Figure 3 illustrates the contents of the CROIB. Unlike the OIB, all the information about a polytomously-scored item was contained together, on consecutive pages. Items were separated by tabbed pages, with the tab showing the item handle (minus the score points). The CROIB included scoring rubrics and examples of student responses at each score level, including zero. The first page showed the item, the information-box, and the page number(s) where the item's score point(s) could be found in the OIB.

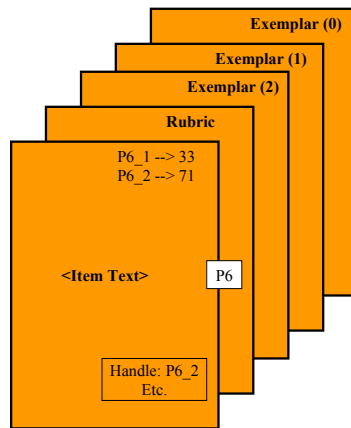


Figure 3. Slide illustrating contents of the CROIB (Constructed Response Ordered Item Book)

Because the KSAs identified by going through the CROIB would need to be recorded in the OIB, panelists used large yellow post-its to record their notes on the KSAs. Their notes were for their own use. They used one post-it for each score point. When panelists were finished with an item, they placed their notes on a Note-template. This was a stapled set of legal size pages that accommodated six post-its per page (See ACT, 2005a for illustration). The Note-template was organized so panelists could transfer their post-its into the OIB in one pass without flipping pages back and forth.

As noted earlier, the OIB contained all items, including the constructed response items that panelists had used in KSA activities 1 and 2. Figure 4 shows how score levels of polytomously-scored items were treated as separate items in the OIB. When panelists see score points of polytomously-scored items relative to the difficulty of all other items in their pool in KSA

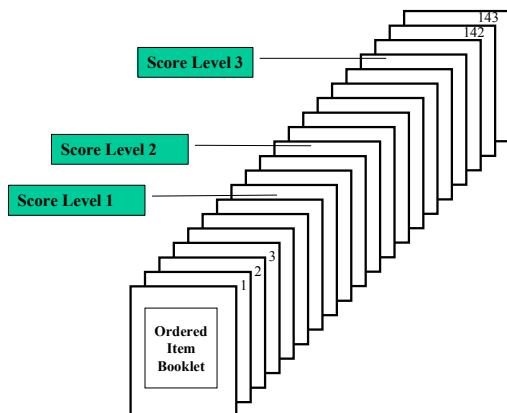


Figure 4. Score levels of a polytomously-scored item are treated as separate items and appear at different places in the OIB.

Activity 3, they typically notice additional KSAs the score point may require that previous, easier items and score points did not require. Panelists record additional notes directly on the pages of the OIB.

Panelists checked items off on their Primary Item Map as they progressed through the OIB. The item check-off process helped panelists see “how much” more difficult one item was than another and which items were in the same content area.

Understanding the Achievement Level Descriptions

Panelists had been instructed to study the ALDs prior to the meeting. To reinforce this learning, the NAGB content facilitator presented the ALDs on slides and provided a clear explanation of how the ALDs were related to both the framework and to the NAGB policy definitions. Panelists were asked to identify KSAs that appeared to be required by each achievement level, and to identify what additional KSAs appeared to be required by a higher achievement level (e.g., Proficient) compared to a lower achievement level (e.g., Basic).

To help panelists see the connection to their OIB and Primary Item Map, panelists at each table were asked to think of a task, preferably in the form of an item, for each achievement level that exemplified a knowledge, skill, or ability that students at that level should have. Some tables shared their tasks/items with the whole group and there was discussion. Panelists

were asked to avoid discussing items in their pool as this might interfere with independent judgment.

Bookmark Placements

A carefully scripted presentation was used to explain this task. The ALD should be thought of as representing a *range* of performance on the achievement scale, and the panelist's job is to decide what the lower *borderline* of that range should be. Panelists were told to think of the lower borderline in terms of a student who was "just qualified" to be in the achievement level and to decide for themselves what "just qualified" means in the process of placing their bookmarks.

Panelists were instructed to page through the OIB, starting with the easiest item, until they come to an item that they judge to be too difficult for mastery by the borderline student. The instructor tells panelists that they might not be sure where to place their bookmarks because 1) they may not feel there is a noticeable or meaningful difference between adjacent items in terms of difficulty, and 2) they may feel that a few items in the OIB are out of order with their own expectations of relative difficulty. Panelists are instructed to go beyond the first item they judge to be too difficult, to see if there are any later items that they feel the borderline student should have mastery of.

These instructions are reinforced graphically with a slide showing a "range of uncertainty." This slide is shown in Figure 5. Panelists were told

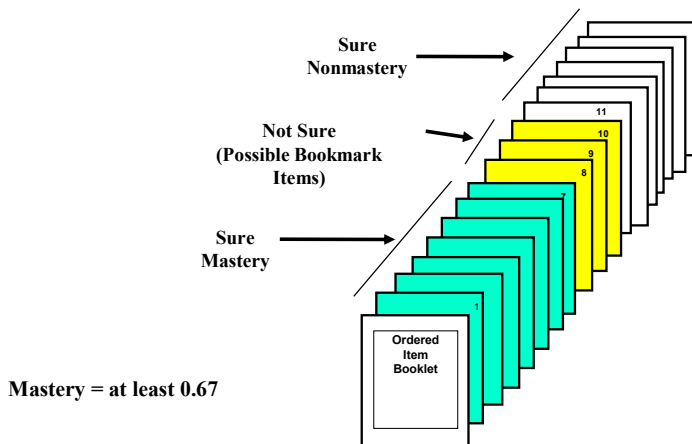


Figure 5. Slide illustrating range of uncertainty in bookmark placements.

that any item in the range of uncertainty would be an acceptable choice for placing the bookmark and that if they had difficulty choosing, the middle of the range would be fine. The final bookmark placement is illustrated in Figure 6.

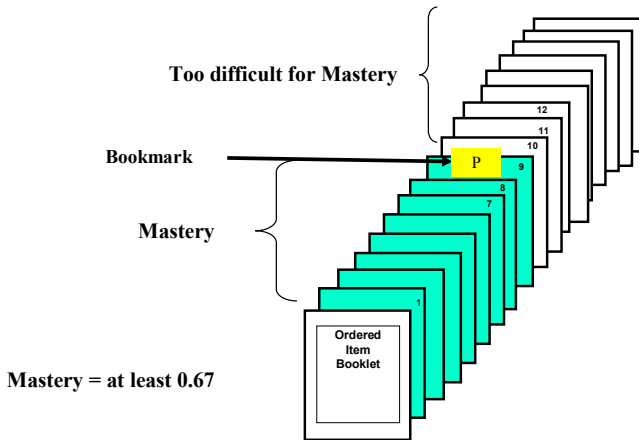


Figure 6. Bookmark placement task simplified.

For each achievement level, beginning with Proficient, then Basic, and then Advanced, panelists read the ALD and used the content of the ALD to place the corresponding bookmark. The next achievement level is not started until all panelists finish their placements for the previous one and key ideas and procedures in the task have been reviewed.

After placing all bookmarks, panelists were given an opportunity to adjust their bookmark placements. Panelists were encouraged to look at all of the ALDs together and to consider whether the differences between their bookmark placements were consistent with the increments of achievement implied by the ALDs. They were instructed to find their bookmarked items on the item map, to consider the vertical distance between the bookmarked items and to make any adjustments they felt were consistent with the ALDs.

Panelists recorded the page number of their bookmark placements on a special form designated for this purpose and circled the handle of their bookmarked item on their Primary Item Map. Page numbers were entered into an interactive computer program that returned the scale value of the

item on the bookmarked page. The scale value was written beneath the bookmarked page number on the panelist's form.

Round 2

Feedback

Feedback at the beginning of Round 2 consisted of showing, for each achievement level, a) the median cut score, b) the high and low cut scores, c) rater-location, and d) domain scores. The median, high, and low cut scores were for the overall distribution of panelists' Round 1 cut scores. These reference points were marked on panelists' materials as described below. The purpose of showing the median was to provide a common focal point for discussing the criterion referenced meaning of potential cut scores. The purpose of showing the high and low was to provide a range for panelists to focus on in selecting their Round 2 cut scores. Panelists were not discouraged from considering cut scores outside this range.

The "rater location" feedback consisted of panelists' marking and observing in these same materials the location of their Round 1 cut score. Although the rater location feedback enabled panelists to evaluate their Round 1 cut score strictly in terms of where other panelists' cut scores were generally located (median, high, and low), its purpose was to help panelists see and process the difference in criterion-referenced meaning between their cut score and the median cut score. To reflect this intent, and to encourage panelists to maintain independent judgment, no other information concerning the distribution of panelists' cut scores was given to panelists, and all the feedback that was given was presented only on materials that allowed criterion-referenced interpretations.

Figure 7 shows how the median cut scores and bookmarked items looked on the Primary Item Map. Panelists were given rulers and pencils and instructed to draw the median cut score lines on their maps. The lines were drawn beneath the midpoint of the intervals containing the cut scores.

Domain Presentations and Tasks

Before panelists were shown domain score feedback, they were given a presentation that explained the domains. The presentation included a brief overview of the domain development process (see ACT, 2005d for a full description of the domain development process). Then, expected domain characteristic curves based on content areas (subscales) were shown (Figure 8). The vertical lines on this plot correspond to the Round 1 cut scores and

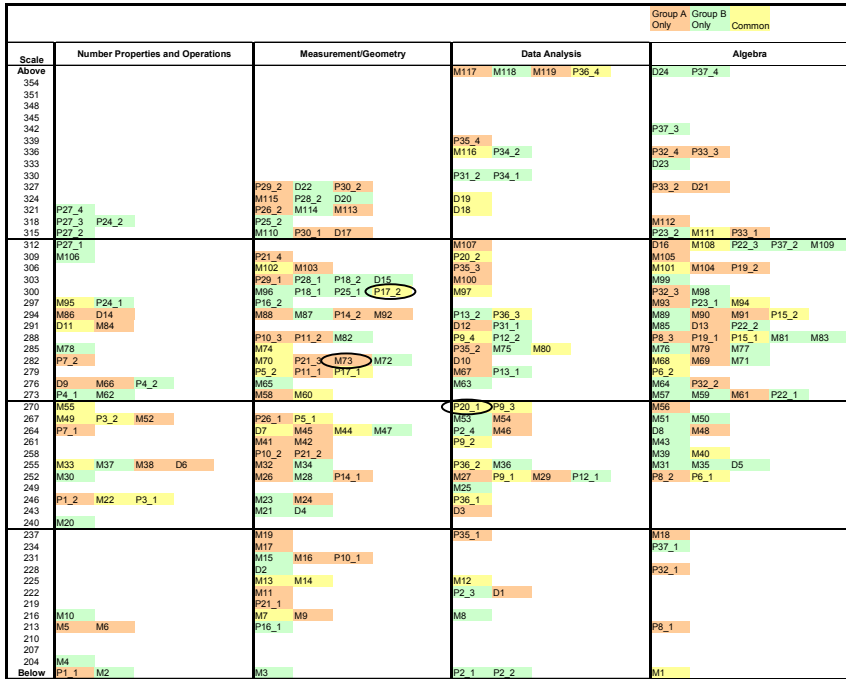


Figure 7. Primary Item Map showing Round 1 median cut scores (horizontal lines) and the location of panelist A1201's bookmarked items (circled).

the dashed horizontal line represents a 67% correct criterion for mastery. It is apparent in this figure that the major content areas (e.g., Number Properties and Operations) are not very useful for understanding growth in achievement as a sequential mastery of skills. The characteristic curves for the content areas are not widely separated in difficulty. Consequently, they do not provide a very rich picture of what students can and cannot do at each achievement level. Students at the Basic cut score are not even close to mastery of any content area, while students at the Advanced cut score are well beyond mastery of all content areas. Moreover, while it is not absolutely clear with these particular curves, when curves cross the sequence in which domains are mastered depends on the mastery criterion.

Plots of score domain curves were then shown to illustrate how the domains defined in the domain development process are more useful for understanding growth in achievement as a sequential mastery of skills. Figure 9 represents the score domains in the Data Analysis content area. The wide range of difficulty covered by these domains allows panelists to make one relatively reliable inference of mastery at the Basic cut score and one

Subscale Percent Correct Curves

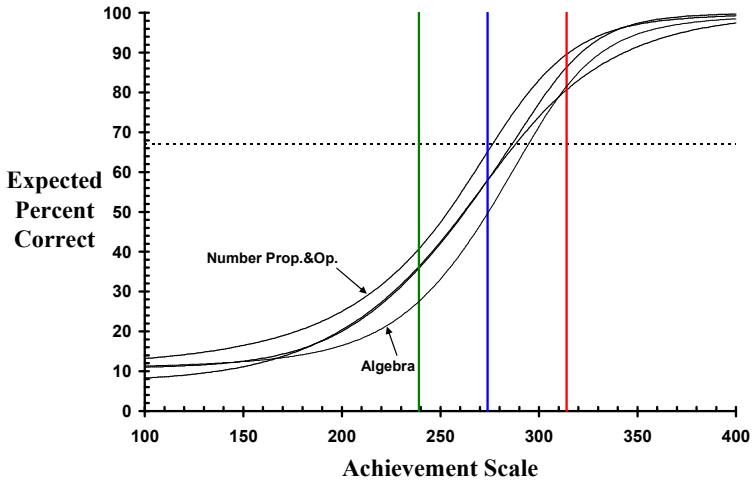


Figure 8. Expected percent correct curves based on subscales of the Assessment Framework.

Data Analysis

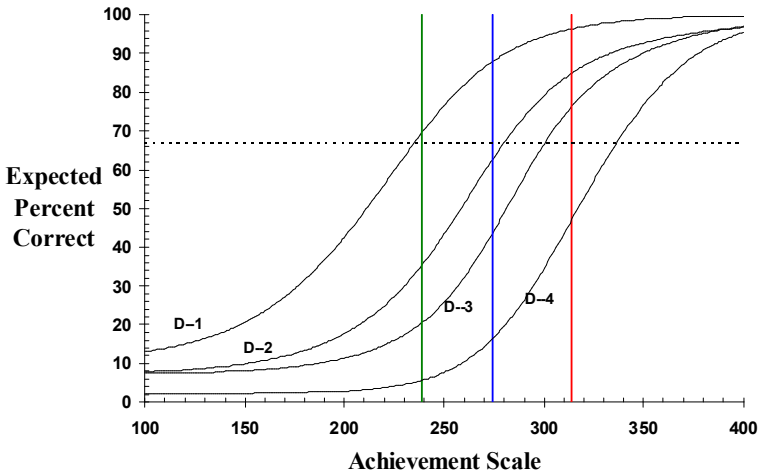


Figure 9. Percent correct curves for score domains in Data Analysis subscale, with vertical lines showing location of Round 1 cut scores and a horizontal line representing a 67% criterion for mastery.

reliable inference of nonmastery at the Advanced cut score. These inferences of mastery are based on a 67% criterion. Panelists are encouraged to use any percent criterion they choose for a particular combination of domain and achievement level, although a lower percent, such as 30%, would not be indicative of “mastery” and a higher percent may be indicative of more than just mastery. Since the domain characteristic curves are non-crossing, the domains are mastered in the same order regardless..

A Percent Correct Table (PCT) was used to show domain scores at all three achievement level cuts. The PCT for Round 1 is shown in Figure 10. Panelists were told that Round 2 cut score recommendations should be based on judgments of whether the domain scores were too low, OK, or too high for the borderline of an achievement level. Tasks in Round 2 were designed to help them understand the domain scores and to make these judgments. The highest, lowest, and closest-to-67% domain scores for the Proficient level were circled (see Figure 10) to illustrate the range of difficulty across domains and to draw panelists’ attention to a domain that easily moves from mastery to nonmastery or vice versa depending on their Round 2 recommendations.

Panelists were next shown a Domain Score Chart (DSC). A DSC shows the expected percent correct score on each score domain for every scale score within a range of 10 points below the Round 1 “low” to 10 points above the Round 1 “high.” Figure 11 shows the DSC for the Proficient Achievement Level. The location of Panelist A1201 is indicated by a circle on the score scale. Panelists were instructed to mark the location of their Round 1 cut score in this fashion. The median, high, and low cut scores were already marked for panelists, as seen in Figure 11. Circles were also pre-drawn around 67% domain scores within the range of the high and low cut scores. The percent correct scores in the row marked “median” of the DSC match the percent correct scores shown in the Percent Correct Table.

The only information that panelists added to the DSC themselves, aside from the location of their Round 1 cut score, was the location of their Round 2 recommended cut score. By seeing their Round 1 cut score on the DSC, panelists could see how much difference there was between their cut score and the median in terms of student performance on domains. In similar fashion, by seeing circles around their bookmarked items on the Primary Item Map together with the horizontal lines representing the median cut scores, panelists could see the distance between their bookmarks and the median and could interpret this distance in terms of intervening item content.

Subscale	Teacher Domain	Score Domain	Expected Percent Correct on Score Domain at Lower Borderline of...		
			Basic	Proficient	Advanced
Number Properties and Operations	N1. Perform Basic Operations	N--1	79%	90%	96%
	N2. Determine Correct Operations	N--2	56%	81%	95%
	N3. Place Value and Notation	N--3	39%	69%	95%
	N4. Multistep Problems	N--4	17%	45%	82%
Measurement/Geometry	M1. Basic Measurement	M--1	62%	83%	97%
	M2. Symmetry, Motion, and Proportionality	M--2	52%	77%	93%
	M3. Identifying Geometric Objects				
	M4. Angles	M--3	35%	61%	89%
	M5. Perimeter, Area, and Volume				
	M6. Coordinates and Their Applications	M--4	22%	41%	80%
	M7. Triangle Properties and Measurements				
	M8. Geometric Relationships	M--5	3%	8%	62%
Data Analysis	D1. Common Data Displays	D--1	70%	88%	96%
	D2. Elementary Probability and Sampling	D--2	35%	63%	85%
	D3. Central Tendency	D--3	21%	44%	76%
	D4. Advanced Data Displays				
	D5. Abstract Reasoning	D--4	6%	16%	47%
Algebra	A1. Reading Tables and Graphs	A--1	44%	73%	93%
	A2. Algebraic Expressions, Equations, and Inequalities				
	A3. Systems of Equations	A--2	26%	49%	86%
	A4. Slopes and Rates				
	A5. Creating and Recognizing Expressions	A--3	19%	37%	74%
	A6. Advanced Functions and Concepts				

Figure 10. Percent Correct Table highlighting expected percent correct scores at Round 1 cut score for Proficient.

Domain Task 1: Understanding domain scores. Domain Task 1 was designed to help panelists understand percent correct scores on the domains by looking at a sample of items from which the domain score was derived and seeing the difficulty of this sample in relation to other items on which the domain score was based. Secondary benefits of this exercise are that it helps panelists 1) gauge the reliability of the domain score by seeing the number of items used to compute the score, 2) see how a single item may not be a reliable measure of a more general skill, and 3) interpret the meaning of distance on the item map.

Scale Score	Number Sense				Measurement					Data Analysis				Algebra			
	N-1	N-2	N-3	N-4	M-1	M-2	M-3	M-4	M-5	D-1	D-2	D-3	D-4	A-1	A-2	A-3	
312	96	95	94	81	97	92	88	78	59	96	84	75	45	93	85	72	
311	96	95	94	80	96	92	88	77	57	96	84	75	45	92	84	72	
310	96	94	93	80	96	92	87	76	55	96	84	74	44	92	83	71	
309	96	94	93	79	96	91	87	76	53	96	83	73	43	92	83	70	
308	96	94	93	78	96	91	86	75	51	96	83	73	42	91	82	69	
307	95	94	92	77	96	91	86	74	49	96	82	72	41	91	81	68	
306	95	94	92	77	96	90	85	73	47	95	82	71	40	91	80	67	
305	95	93	91	76	95	90	85	72	45	95	82	71	39	90	79	66	
304	95	93	91	75	95	90	84	71	44	95	81	70	38	90	79	65	
303	95	93	90	74	95	90	84	70	42	95	81	69	37	90	78	64	
High	302	95	93	90	73	95	89	83	68	40	95	80	68	36	89	77	63
	301	95	92	89	73	94	89	82	67	38	95	80	68	36	89	76	62
	300	95	92	89	72	94	89	82	66	36	95	79	67	35	89	75	61
	299	94	92	88	71	94	88	81	65	34	94	79	66	34	88	74	60
	298	94	92	88	70	94	88	80	64	32	94	78	65	33	88	73	59
	297	94	91	87	69	93	88	80	63	31	94	78	64	32	87	72	58
	296	94	91	87	68	93	87	79	62	29	94	77	63	31	87	71	57
	295	94	91	86	67	93	87	78	61	27	94	77	63	30	86	70	56
	294	94	90	85	66	92	87	77	60	26	93	76	62	30	86	69	55
	293	94	90	85	65	92	86	77	59	24	93	76	61	29	85	68	54
	292	93	90	84	64	92	86	76	58	23	93	75	60	28	85	67	54
	291	93	89	83	63	91	85	75	57	21	93	74	59	27	84	66	53
	290	93	89	83	62	91	85	74	56	20	93	74	58	27	84	65	52
	289	93	89	82	61	91	85	74	55	19	92	73	57	26	83	64	51
	288	93	88	81	60	90	84	73	54	18	92	73	56	25	83	63	50
	287	93	88	80	59	90	84	72	53	17	92	72	56	24	82	62	49
	286	92	87	80	58	89	83	71	52	16	92	71	55	24	82	61	48
	285	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	284	92	86	78	56	88	82	70	50	14	91	70	53	22	80	59	46
Panelist X ->	283	92	86	77	55	88	82	69	49	13	91	69	52	22	80	58	45
	282	92	86	76	54	88	81	68	48	12	91	69	51	21	79	57	44
	281	91	85	75	53	87	81	67	47	11	90	68	50	20	78	56	43
	280	91	85	75	51	87	80	66	46	11	90	67	49	20	78	55	42
	279	91	84	74	50	86	80	65	45	10	90	66	48	19	77	54	41
	278	91	84	73	49	86	79	64	44	9	89	66	47	19	76	53	41
	277	91	83	72	48	85	79	64	43	9	89	65	46	18	76	52	40
	276	90	83	71	47	84	78	63	43	8	89	64	45	17	75	51	39
	275	90	82	70	46	84	78	62	42	8	88	63	45	17	74	50	38
Median	274	90	81	69	45	83	77	61	41	8	88	63	44	16	73	49	37
	273	90	81	68	44	83	77	60	40	7	88	62	43	16	73	49	37
	272	89	80	68	43	82	76	59	39	7	87	61	42	15	72	48	36
	271	89	80	67	42	82	75	58	39	6	87	60	41	15	71	47	35
	270	89	79	66	41	81	75	57	38	6	86	60	40	14	70	46	34
	269	89	79	65	40	81	74	57	37	6	86	59	39	14	69	45	34
	268	88	78	64	39	80	74	56	37	6	86	58	38	14	69	44	33
	267	88	77	63	38	79	73	55	36	5	85	57	38	13	68	43	32
	266	88	77	62	37	79	72	54	35	5	85	56	37	13	67	43	32
	265	88	76	61	36	78	72	53	35	5	84	55	36	12	66	42	31
	264	87	75	60	35	78	71	52	34	5	84	55	35	12	65	41	30
	263	87	75	59	34	77	70	52	33	5	83	54	34	12	64	40	30
	262	87	74	58	33	76	69	51	33	4	83	53	34	11	64	39	29
	261	87	73	58	32	76	69	50	32	4	83	52	33	11	63	39	29
	260	86	72	57	31	75	68	49	32	4	82	51	32	10	62	38	28
	259	86	72	56	30	74	67	48	31	4	82	51	32	10	61	37	28
	258	86	71	55	29	74	67	48	30	4	81	50	31	10	60	37	27
	257	85	70	54	28	73	66	47	30	4	81	49	30	10	59	36	27
	256	85	70	53	28	73	65	46	29	4	80	48	29	9	58	35	26
	255	85	69	52	27	72	64	45	29	4	79	47	29	9	57	35	26
	254	85	68	51	26	71	64	45	28	4	79	47	28	9	56	34	25
	253	84	67	51	25	71	63	44	28	3	78	46	28	8	56	33	25
	252	84	66	50	25	70	62	43	27	3	78	45	27	8	55	33	24
	251	84	66	49	24	69	61	43	27	3	77	44	26	8	54	32	24
	250	83	65	48	23	69	61	42	26	3	77	43	26	8	53	31	23
	249	83	64	47	23	68	60	41	26	3	76	43	25	8	52	31	23
	248	83	63	46	22	68	59	40	25	3	75	42	25	7	51	30	23
	247	82	62	46	21	67	58	40	25	3	75	41	24	7	50	30	22
	246	82	62	45	21	66	57	39	25	3	74	40	24	7	49	29	22
	245	81	61	44	20	66	57	39	24	3	74	40	23	7	49	29	21
Low	244	81	60	43	20	65	56	38	24	3	73	39	23	7	48	28	21
	243	81	59	42	19	64	55	37	23	3	72	38	22	6	47	28	21

Figure 11. Domain Score Chart showing Round 1 results and location of panelist A1201 for Proficient achievement level.

The materials used in Domain Task 1 were a) a Domain Ordered Item Book, or DOIB, b) Domain Item Maps, and c) the Domain Task 1 form. Panelists were given one Domain Item Map for each subscale. Figure 12 shows the Domain Item Map for the Data Analysis and Probability subscale. The vertically-upward trend of item locations as one goes from left to right in the Domain Item Map illustrates the general trend of increasing difficulty in the teacher and score domains from left to right. Facilitators drew panelists’ attention to this trend and to the variability of item difficulty within the teacher and score domains. This variability means that no single item is a very reliable indication of the difficulty of a more general skill.

Scale	D1 Common Data Displays	D2 Elementary Probability and Sampling	D3 Central Tendency	D4 Advanced Data Displays	D5 Abstract Reasoning
Above		P36_4	M118		M117 M119
354					
351					
348					
345					
342					
339			P35_4		
336				M116	P34_2
333					P31_2 P34_1
330				D19	
327					D18
324					
321					
318					
315					
312		M107			
309		P20_2	P35_3		
306					
303		M100			
300			M97		
297		P36_3		P13_2	
294					P31_1
291			D12		
288		P9_4 P12_2			
285		M75	P35_2	M80	
282		D10			
279			M67	P13_1	
276		M63			
273					
270		P20_1 P9_3			
267		M53 M54			
264	P2_4 M46				
261		P9_2			
258					
255	M36	P36_2			
252	M27 M29	P9_1 P12_1			
249	M25				
246		P36_1			
243	D3				
240					
237			P35_1		
234					
231					
228					
225	M12				
222	P2_3 D1				
219					
216	M8				
213					
210					
207					
204					
Below	P2_1 P2_2				
Border Adv.: 96 %		85 %		76 %	47 %
Border Prof.: 88 %		63 %		44 %	16 %
Border Basic: 70 %		35 %		21 %	6 %

Figure 12. Domain Item Map for Data Analysis and Probability Subscale.

The DOIB was like the OIB in containing one page per item. Items were presented in order of difficulty, within teacher domain. Polytomously-scored items were shown only once, and were located by their highest score. Teacher domain definitions (see Figure 1 for example) were at the front

of each set of items. The teacher domains were in order of their columns from left to right on the Domain Item Map.

Panelists responded to the question, “I see how this item is like other items in its domain” for each item in their pool. Before answering this question, panelists read the narrative of the Teacher Domain definition and looked at the sample items in the domain definition. In answering this question for polytomously-scored items, panelists were told to think of the KSAs needed to attain the highest score on the item. Figure 13 shows a section of the Domain Task 1 Form for Group A. The complete form was four pages, one for each subscale, and included all teacher domains.

Teacher Domain	Item Handle	I see how this item is like other items in its domain. (Check ✓)		
		Yes	Not Sure	No
N1) Perform Basic Operations	M5			
	P1_2			
N2) Determine Correct Operations	M6			
	M22			
	M33			
	P3_2			
	D9			
N3) Place Value and Notation	M66			
	D6			
	M49			
	M52			
	M84			

Figure 13. Section of Domain Task 1 Form for Group A.

As panelists worked through the items within a teacher domain, they noted the items’ locations on their Domain Item Map. The expected percent correct scores shown at the bottom of the Domain Item Map were conditional on the cut scores represented by horizontal lines across the map. [These percent correct scores matched those shown in the Percent Correct Table and in the median row of the Domain Score Chart.] Facilitators drew panelists’ attention to the following:

- The expected percent correct scores were based only on the items shown on the map.
- The items in each panelist’s pool are only a sample of items on which the expected percent correct score was based. Group A’s items were tan and yellow. Group B’s items were green and yellow. (These colors are not reproduced in the figures.) Panelists could see whether their items

were more or less difficult than all of the items put together within a score domain.

- All of the items on the map are only a sample of the items that could be included in the domain.

The following points were made to help panelists understand domain scores in terms of the spatial information on their DOIB:

- when most items in a domain lie below the cut score, the expected percent correct score on the domain is higher than 67%
- when most items in the domain lie above the cut score, the expected percent correct score on the domain is less than 67%
- when items are about equally above and below the cut score, the expected percent correct score on the domain is about 67%

When panelists finished reviewing items by teacher domain within a given subscale, they were shown a plot of expected percent correct curves on the score domains within that subscale. Figure 9 shows the plot that was presented for the Data Analysis and Probability subscale. The plots were used to show that there is a range of achievement within each achievement level and to focus panelists on deciding what the lower boundary of the range should be.

The KSA review in Round 1 proved to be a useful precursor to Domain Task 1. Panelists were prepared to see why one item within a domain was easier than another, yet why items belonged to the same domain. Panelists may have identified similar KSAs for items in the same domain.

Domain Task 2: Evaluating the domain scores. In Domain Task 2, panelists made judgments about whether the domain scores associated with the Round 1 median cut score should be higher, lower, or are OK as a standard of lower borderline performance for a given achievement level. Figure 14 shows the form that was used to collect panelists' judgments about domain scores associated with the Round 1 median cut score for Proficient. Similar forms were used for the other achievement levels.

Although panelists could answer the Domain Task 2 question strictly on the basis of whether they thought the domain score should be higher or lower than 67% (i.e., mastery), they were encouraged to think more generally. They were told to think of what was acceptable borderline performance on a scale ranging from guessing to 100% correct. This was like an Angoff-based task at a domain-level except that they already had information about domain difficulty, and they did not have to provide an

Subscale	Teacher Domain	Score Domain	Expected Percent Correct Borderline PROFICIENT	I think the percentage correct score at the PROFICIENT borderline should be... (check the appropriate cell)		
				lower	OK	higher
Number Properties and Operations	N1. Perform Basic Operations	N--1	90%			
	N2. Determine Correct Operations	N--2	81%			
	N3. Place Value and Notation	N--3	69%			
	N4. Multistep Problems	N--4	45%			
Measurement/Geometry	M1. Basic Measurement	M--1	83%			
	M2. Symmetry, Motion, and Proportionality	M--2	77%			
	M3. Identifying Geometric Objects					
	M4. Angles	M--3	61%			
	M5. Perimeter, Area, and Volume					
	M6. Coordinates and Their Applications	M--4	41%			
	M7. Triangle Properties and Measurements					
	M8. Geometric Relationships	M--5	8%			
Data Analysis	D1. Common Data Displays	D--1	88%			
	D2. Elementary Probability and Sampling	D--2	63%			
	D3. Central Tendency	D--3	44%			
	D4. Advanced Data Displays					
	D5. Abstract Reasoning	D--4	16%			
Algebra	A1. Reading Tables and Graphs	A--1	73%			
	A2. Algebraic Expressions, Equations, and Inequalities					
	A3. Systems of Equations	A--2	49%			
	A4. Slopes and Rates	A--3	37%			
	A5. Creating and Recognizing Expressions					
	A6. Advanced Functions and Concepts					

Figure 14. Domain Task 2 Form for Proficient Achievement Level.

expected score on each domain—they only had to indicate whether a borderline score should be higher, lower, or about equal to the domain score associated with the Round 1 median.

Round 2 Cut Score Recommendations

Panelists used the Domain Score Chart to select scale values for their Round 2 cut score recommendations. So that panelists would focus their attention on the domains, the OIB was not available to panelists in Round 2. Instructions for selecting scale values began by directing panelists to consider the pattern of checks on their Domain Task 2 form. If all of the checks were in the “OK” column, they should recommend a cut score

close to the median on the DSC. If all of the checks were in the “higher” column, they should select a cut score higher than the Round 1 median. If there were checks in both the “higher” and “lower” columns, they should use their judgment and consider which information they understood best and which domains they considered most important. They were advised to look to the ALDs for guidance. Panelists were also advised to give less importance to domains represented by smaller numbers of items, other things being equal, and also to give less importance to domains with very high or very low expected scores.

Panelists were also told that their Round 1 cut scores could be a factor in their Round 2 cut score recommendation. If the domain scores associated with their Round 1 cut score were consistent with the pattern of “higher/lower” checks on their Domain Task 2 form, or if they did not feel comfortable with their understanding of the domain scores, they could simply recommend the scale value/cut score that had been derived from their Round 1 bookmark placement.

In making their scale value selections, panelists were instructed to work independently and to start with the Proficient cut score, then recommend a Basic cut score, and finally an Advanced cut score. Panelists recorded their cut scores on a form and circled their choices on their DSC. They also marked the location of their cut scores on their Primary Item Map (by circling the number closest to their cut scores in the Scale Value column).

Round 3

Feedback

Feedback at the beginning of Round 3 was similar to that provided in Round 2, except that the OIB was also included in the materials. Panelists were shown the numerical values of the Round 1 and Round 2 medians. Panelists could see the change in the median from Round 1 to Round 2.

For each achievement level, panelists were given the OIB page numbers that corresponded to the easiest and hardest items within the range of the highest and lowest cut scores recommended in Round 2. They placed flags on these pages. Different colored flags were used for each achievement level in case the high flag of a lower level overlapped with the low flag of a higher level.

Panelists were instructed to draw horizontal lines across a new Primary Item Map to indicate the location of the Round 2 medians. They circled the

midpoint of the map-interval that contained their Round 2 cut score recommendations. The Domain Score Chart was marked as shown in Figure 11 except that the Round 2 medians, highs, and lows were shown. Panelists circled their Round 2 cut score on the new Domain Score Chart.

Whole-Group Discussion of General Concepts

Following presentation of feedback, there was a whole-group discussion to check for understanding of general concepts in the standard setting process. The concept of borderline performance was illustrated with Domain Score Plots showing cut scores as vertical lines (e.g., Figure 9). Panelists were asked if they were comfortable with the range of performance represented by an achievement level and with the lower borderline represented by the cut score. Panelists were reminded that they should not place too much importance on where their cut score lay with respect to a single item. A Domain Item Map was shown as a reminder that a usefully general skill (domain) cannot be represented by a single item. Panelists were encouraged to consider the distance between achievement levels on their item map.

Rater Group Discussion of Criterion-Referenced Meaning

Most of the time in Round 3 was spent on a “Rater Group Discussion” of the criterion-referenced meaning of panelists’ own cut scores. Within each group, tables were pulled together and panelists took turns sharing the following: 1) the criterion-referenced meaning of their Round 1 bookmark placement, 2) the criterion-referenced meaning of their Round 2 cut scores, and 3) what information they considered important in recommending a Round 3 cut scores.

To encourage independent judgment, the purpose of the discussion was characterized as “sharing,” not “persuading”. Facilitators circulated to keep the discussion on track, to keep the ALDs in focus, and to encourage all panelists to participate. The discussion began with the Proficient level, then moved to Basic, and finished with Advanced. The discussion lasted about 90 minutes. Panelists had available all of the key materials: the ALDs, OIB, Primary Item Map, Domain Item Maps, Domain Descriptions, DSC, and PCT (based on Round 2 median cut score).

Round 3 Cut Score Recommendations

Panelists again selected scale values for their Round 3 cut score recommendations. They were instructed to use the information that they

understood best. They were asked, however, to record their cut score recommendations in all of their materials. To facilitate this recording, they were given a special table showing the correspondence between OIB page numbers and scale values in their Domain Score Chart. Panelists were instructed to work independently and to reflect on the feedback and the Round 3 discussion in making their selections.

Round 4

Feedback

Feedback at the beginning of Round 4 was similar to that provided at the beginning of Round 3. Panelists were given a new Primary Item Map, Domain Score Chart, and Percent Correct Table for marking their Round 3 recommendations in the context of the Round 3 medians, highs, and lows. A table of the median cut scores from Rounds 1 to 3 was presented to show panelists how the cut scores were changing (or not) over rounds and what the current cut scores were. The consequences data described below were also a form of feedback.

Consequences Data and Discussion

“Consequences” data are the percent of students in each achievement level and the percent at or above each achievement level. The percent of students below Basic is also shown. This information was presented in the form of a bar graph and pie chart. The consequences data was based on the Round 3 median cut scores. The consequences data are also referred to as “achievement level percentages.” As a lead-in to discussing these percentages, panelists were reminded that student performance is estimated from tests like the ones they took, which were taken by students under similar time constraints. Panelists were told that the consequences data was nationally representative. Although they were told that they could consider the effect of student motivation and time constraints on student performance, they were reminded that evidence of what students at the cut scores “can do” should ultimately be consistent the ALDs and the concept of borderline.

Round 4 Cut Score Recommendations

Panelists again selected scale values for their cut score recommendations. Panelists were instructed to work independently, study the feedback from Round 3, reflect on the discussion of the consequences data, and to choose and record a scale value for their cut score recommendations. Pan-

elists recorded their cut score recommendations in the OIB, domain score chart, and Primary Item Map, as they had in Round 3.

Post-Rounds Activities

Feedback

Feedback after Round 4 was similar to that provided at the beginning of Round 3. A new Primary Item Map, Domain Score Chart and Percent Correct Table were distributed, but only the Round 4 median cut scores were pre-marked or added to these materials by panelists. Highs and lows were not given. Panelists did not mark their own Round 4 cut scores on these materials. The feedback included achievement level percentages based on the Round 4 medians.

Consequences Questionnaire

A consequences questionnaire was given to panelists to assess their reactions to the cut scores after viewing the consequences data. Panelists were told that the Round 4 medians would be reported to NAGB as one of the key outcomes of the ALS meeting. Panelists were asked to evaluate the cut scores based on the match between the criterion-referenced feedback, the ALDs, and their concept of borderline performance. For each level, panelists could endorse the Round 4 cut score or recommend a different cut score. Twenty-three of the panelists endorsed the Round 4 cut score at each level.

Cut Score Results

Table 2 shows median cut scores by round within achievement level. There was very little change in the median cut score over rounds. This is not always the case with the Mapmark method (ACT, 2005c, 2007a). Based on process evaluation results considered in the next section and on results of a pilot study conducted in this project (ACT, 2005c), the most likely reason

Table 2

Median Cut Scores

Round	Basic	Proficient	Advanced
1	239	274	314
2	240	276	314
3	241	275	314
4	241	275	314

why cut scores did not change much in the ALS meeting is that already at Round 1 they were in reasonable agreement with the feedback and tasks that occurred in later rounds.

The final cut scores (Round 4) were 241 for Basic, 275 for Proficient, and 314 for Advanced. Student achievement on this scale had a mean of approximately 250 and a standard deviation of approximately 30. The achievement level percentages corresponding to these cut scores were: 2% at Advanced, 22% at Proficient, 39% at Basic, and 37% below Basic. These results differ from those shown in *The Nation's Report Card* due to slight differences in item parameters and student distribution data between field test and operational testing and the linear transformation of scale mentioned earlier.

Table 3

Number of Panelists Increasing, Decreasing, or Not Changing Their Cut Scores by Achievement Level and Round

Round	Basic			Proficient			Advanced		
	No			No			No		
	Increase	Change	Decrease	Increase	Change	Decrease	Increase	Change	Decrease
2	15	5	11	15	3	13	12	7	12
3	11	9	11	8	14	9	6	14	11
4	7	18	6	6	17	8	5	20	6

Although the median cut scores showed little change over rounds, a substantial proportion of panelists changed their cut score each round. Table 3 shows that no less than one third of the panelists changed their cut scores from one round to the next. In each case, however, about as many panelists increased as decreased their cut scores.

Panelists' cut scores were further studied to assess variability, reliability, and other properties. Reliability analyses of data from other Mapmark studies as well as the ALS meeting show that Mapmark cut scores are reliable (Yin, Schulz, and Sconing, 2005; Yin and Sconing, 2008). The standard error of cut scores in the ALS meeting was estimated to be approximately 2 to 3 points (ACT, 2005b). The average absolute difference of individual cut scores from the median fell from approximately 14 in Round 1 to approximately 7 in Round 2. The mean absolute difference remained about the same after Round 2 even though individual panelists continued to change their cut score, and about as many panelists raised as lowered their cut scores each round (Table 3). This indicates that panelists were not changing their cut scores just to be closer to the group median.

An analysis of panelists' bookmark placements in Round 1 showed that panelists did not tend to bookmark one type of item more than any other (ACT, 2005a). For example, score levels of polytomously-scored items were bookmarked only in proportion to the total number of points they represented in the assessment (40%).

Process Evaluations

Process validity data includes panelists' responses to process evaluation questionnaires and to the Domain tasks. Panelists responded to a total of six questionnaires throughout the ALS meeting. Some questions were used in NAEP standard setting projects dating back to the 1992 NAEP mathematics assessment. New questions were specific to Mapmark. On a 5-point rating scale where 5 is highest and 1 is lowest, an average of 3.5 has historically been considered acceptable. Averages at or above 4.0 have been considered good, and averages at or above 4.5 have been considered very good.

Process Evaluation Questionnaire Responses

Table 4 shows the mean ratings of Mapmark and of previous NAEP ALS methods on the key process evaluation questions. Both of the previous NAEP ALS methods represented in this table were Angoff-based. Both were used to set achievement levels for NAEP assessments. Statistical significance tests were not performed on the differences among methods, but it can be seen that the average rating for Mapmark were generally as high as or higher than for the other two methods. It should be noted that a "3" was an optimum rating for the amount of time allocated for tasks.

As shown in Table 5, panelists had good understanding of concepts and procedures essential to the Mapmark method. Panelists understood, for example, how to use their item map (Primary) and OIB, the concept of domains scores, and how to use Domain Item Maps, the Domain Ordered Item Booklet (Domain Task 1), and the Domain Score Chart.

Panelists indicated good levels of comfort with tasks such as working through the OIB on their own (KSA Activity 4), and deciding whether an item was like other items in its domain (Domain Task 1). They were also comfortable with a two-thirds chance representing "mastery" and with using this probability to interpret item locations on their maps. Panelists expressed only an acceptable level of confidence in deciding whether domain scores should be higher or lower (3.84; Domain Task 2) and in choosing a scale value rather than a bookmark placement to recommend a cut score (3.90),

Table 4

Mean Ratings of Mapmark and Previous ALS Methods on Key Process Evaluation Questions

Question	Meeting	Mean
The most accurate description of my level of confidence in the cut score recommendations I provided was... (5=Totally confident)	Mapmark ALS	4.37
	1998 Civics	4.04
	1992 Math	4.12
I would describe the effectiveness of the achievement level setting method as... (5=Highly effective)	Mapmark ALS	4.28
	1998 Civics	3.59
	1992 Math	4.07
This ALS process provided me an opportunity to use my best judgment to recommend cut scores (5=To a great extent)	Mapmark ALS	4.57
	1998 Civics	4.11
	1992 Math	4.46
The instructions on what I was to do during each round were... (5=Absolutely clear)	Mapmark ALS	4.17
	1998 Civics	4.18
	1992 Math	4.13
My understanding of the tasks I was to accomplish during each round was... (5=Totally agree)	Mapmark ALS	4.27
	1998 Civics	4.11
	1992 Math	4.24
The amount of time I had to complete the tasks I had to accomplish was generally... (3=About right)	Mapmark ALS	3.03
	1998 Civics	3.21
	1992 Math	3.12

Table 5

Understanding of How to Use Materials

Round	Activity	Average Rating
1	How to use my Primary Item Map and Ordered Item Book	4.42
2	How to use the Domain Item Maps	4.19
2	How to use the Domain Ordered Item Booklet	4.52
2	How to use the Domain Score Chart	4.39

but these ratings were obtained in Round 2. It is reasonable to suppose that panelists’ confidence in these tasks increased over subsequent rounds.

Panelists found the materials and information in the Mapmark process to be useful (see Table 7). As shown in Table 7, panelists found the OIB most useful, followed by rater location feedback, the achievement level descriptions, item maps, and domain score feedback. This rank ordering has varied slightly across Mapmark studies, but the OIB has almost always been the most useful (ACT, Inc., 2005a, c, 2007a). Item Maps have moved up in usefulness as facilitators gained experience (ACT, Inc., 2007a). Consequences data is generally found to be least useful, as one would expect in a process highly oriented to criterion-referenced standard setting. The

Table 6

Comfort and Confidence

Round	Activity	Average Rating
1	How to use my Primary Item Map and Ordered Item Book	4.42
1	Using a 2/3 or 0.67 probability to interpret the location of an item on my map	4.23
1	Working through the ordered item booklet on my own	4.39
1	Using a 0.67 probability to define mastery in placing my bookmarks	4.00
2	Thinking about whether an item was like other items in its domain (Domain Task 1)	4.39
2	Deciding whether domain scores should be higher or lower	3.84
2	Choosing scale values instead of placing bookmarks to recommend cut scores	3.90
4	Using the consequences data to recommend cut scores	4.30

relatively high average rating panelists' gave for usefulness of rater location data, 4.46 is associated with the median, high, and low cut scores indicated on panelists' item maps and OIB, from which criterion-referenced meaning can be taken. Although panelists may have given even higher ratings to feedback if additional normative information, such as the entire distribution of cut scores, had been provided, we believe the feedback rating is sufficiently high and that additional normative information, if presented outside of a clear, criterion-referenced context, could send a confusing message to panelists about the importance of independent judgment.

Panelists' respected the value of independent judgment in the process. Panelists tended to disagree with the statement that they felt pressure to recommend cut scores that were close to those of another panelist. At the conclusion of Round 1, the average response to the question, "I feel that my perspective is being heard by others in my table group" was 4.5 (5 = "totally agree"). At the conclusion of the meeting, the average response to the statement, "I felt my input was valued and considered by others in my group" was 4.32 (5 = "to a great extent").

Domain Task 1 Responses

Panelist's responses to Domain Task 1 indicated that the domains were coherent. Recall that they were asked whether they saw how each item fit into its particular domain (yes, no, not sure). The overall percentage of "Yes" responses across all items and panelists was 93%. By panelist type, the percentage was 96% for teachers, 91% for non-teacher educators, and 89% for general public representatives. By individual panelist, the percentage of "Yes" responses ranged from 62% (a general public representative)

Table 7
Usefulness of Materials and Information

Information/materials	Average Rating
The Achievement Level Descriptions	4.38
The Ordered Item Booklet	4.76
The Primary Item Map	4.24
The Domain Ordered Item Maps	4.24
The rater location data	4.46
The domain score feedback	4.21
The consequences data	4.07

Note: These questions were asked at conclusion of the meeting.

to 100% (for two teachers). It seems natural for the percentage of ‘yes’ responses to be highest among teachers, next highest among non-teachers educators, and lowest for the general public representatives.

Domain Task 2 Responses

The relative frequency of checks in the higher/OK/lower categories of Domain Task 2 was consistent with the direction and amount of change in panelists’ cut scores. Table 8 shows the percentage of checks in each category by achievement level in the ALS meeting. At all three levels, the majority of checks (averaged across panelists and domains) were in the “OK” category. The largest percentage of OK checks was for the Advanced achievement level (70%) where the cut score did not change at all over rounds (see Table 2). At the Basic and Proficient levels, most checks were in the OK category (56% and 54% respectively), but there were significantly more checks in the “higher” category than in the “lower” category. These results are consistent with the slight increase in the median cut score for both of these achievement levels from Round 1 to Round 2 (1 and 2 points respectively) as shown in Table 2. In other Mapmark meetings, fewer checks were in the “OK” column, cut scores tended to change more, and the direction of change was consistent with the balance of checks in the “higher” and “lower” categories (ACT, 2005a, c).

Table 8
Domain Task 2 Results

Category	Percent of Checks in Category		
	Basic	Proficient	Advanced
Higher	25	27	12
OK	56	54	70
Lower	19	18	15

Discussion and Conclusions

Of the four components judged to be most important in the Mapmark method (bookmark, item maps, holistic feedback, and independent judgment), the bookmark component is probably the most fundamental. Order information is more fundamental than distance information, and easier to use. When bookmark was first conceived, the developers had hoped to include a spatially-representative item map in the method, but cost and time constraints in state assessments generally made their routine use prohibitive. Yet the method has been very successful. Since the bookmark method was introduced in 1996 (Lewis, et al., 1996), it has become the most widely used standard setting method in state assessments (Council of Chief State School Officers, 2001).

While Mapmark could be called an enhanced bookmark procedure, it was given a separate name to stress the importance of item maps in standard setting generally. Results from the ALS meeting and ACT's research over the past two NAEP standard setting contracts (ACT, 2005a, b, c; ACT, 2007a, b) show that panelists like item maps and find the distance information on them useful when performing tasks such as the KSA review, and placing bookmarks. The columns on item maps were effective for representing content dimensions in panelists' tasks.

Holistic feedback in Mapmark was effective largely because of item maps. Bookmark developers were unsure how panelists would deal with inconsistency in holistic feedback, such as when a student gets an easy item wrong and a hard item right (Mitzel, et al., 2001). When "whole-booklet" feedback was presented in Mapmark (ACT, 2007a, b), panelists were able to consider inconsistency in a student's work by marking the item-responses on an item map. They could evaluate inconsistency in terms of "how far" an item was located above or below the student's location on the item map. In the domain score feedback described in this chapter, there was no "inconsistency" to be shown, but Domain Item Maps were helpful to panelists for understanding and effectively using this form of holistic feedback.

Independent judgment was effective in Mapmark also largely because of item maps. Distance information on the map gave meaning to any scale value in terms of "how far" it was from any given item. This allowed panelists to meaningfully select scale values for their cut score recommendations. They did not place bookmarks after Round 1. The information on the item map gave panelists more choice and gave their cut score judgments more

nuance. More choice and nuance increases the likelihood that panelists will exercise independent judgment. When the meaning of a particular score is not clear, a panelist may feel anxious about their cut score recommendation, and might therefore be less likely to exercise independent judgment.

In the future, standard setting planners should consider adding normative and external validity information to item maps. For example, columns could be added to an item map to show the probability of earning an “A” in an advanced placement course and to show the cumulative percentage of students at each scale value. Panelists could then consider test-centered, norm-referenced, and external validity information simultaneously while choosing a scale value for their cut score recommendation.

On its face, the Mapmark method might not seem easier or less cognitively complex for panelists than other methods. Panelists are given a great deal of information. But the information is structured through the ordering of items by difficulty. The spatial representation of difficulty information on an item map makes the information even more assessable. The value placed on independent judgment allows each panelist to select a cut score using only the information they understand and have experience with. The common frame of reference provided by item maps for all the information in the process—from item-level to holistic—allows panelists to discuss and understand the information in the process even better. The resulting cut score is broadly representative, reliable, and responsive to all of the information presented.

Acknowledgement

Work described in this chapter was conducted under contract with the National Assessment Governing Board. Contract No. ED-03-CO-0099.

References

- ACT, Inc. (April, 2005a). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Process report*. Iowa City, IA: Author.
- ACT, Inc. (April, 2005b). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Technical report*. Iowa City, IA: Author.
- ACT, Inc. (April, 2005c). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Special studies report*. Iowa City, IA: Author.

- ACT, Inc. (April, 2005d). *Developing achievement levels on the 2005 national assessment of educational progress in grade twelve mathematics: Domain development report*. Iowa City, IA: Author.
- ACT, Inc. (April, 2007a). *Developing achievement levels on the 2006 national assessment of educational progress in grade twelve economics: Process report*. Iowa City, IA: Author.
- ACT, Inc. (April, 2007b). *Developing achievement levels on the 2006 national assessment of educational progress in grade twelve economics: Technical report*. Iowa City, IA: Author.
- ACT, Inc. (April, 2007c). *Developing achievement levels on the 2006 national assessment of educational progress in grade twelve economics: Special studies report*. Iowa City, IA: Author.
- Buckendahl, C. W., Smith, R. W., Impara J. C., and Plake, B. S. (2002). A comparison of Angoff and Bookmark standard setting methods. *Journal of Educational Measurement*, 39, 253-263.
- Chen, W., Loomis, S. C., and Fisher, T. (2000, December). *Developing achievement levels on the 1998 NAEP in civics and writing: Technical report*. Iowa City, IA: ACT.
- Council of Chief State School Officers. (2001). *State student assessment programs annual survey. Data volume II*. Washington, DC: Author.
- Engelhard, G., Jr., and Gordon, B. (2000). Setting and evaluating performance standards for high stakes writing assessments. In M. Wilson and G. Engelhard, Jr. (Eds.), *Objective measurement: Theory into practice* (Volume 5, pp. 3-14). Stamford, CT: Ablex Publishing Corporation.
- Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, 10(3), 3-9.
- Glas, G. V. (1978). Standards and criteria. *Journal of Educational Measurement*, 15, 237-261.
- Green, D. R., Trimble, C. S., and Lewis, D. L. (2003). Interpreting the results of three different standard setting procedures. *Educational Measurement: Issues and Practice*, 22(1), 22-32.
- Grosse, M. E., and Wright, B. D. (1986). Setting, evaluating, and maintaining certification standards with the Rasch model. *Evaluation and the Health Professions*, 9, 267-285.
- Hambleton, R. K. (1978). On the use of cutoff scores with criterion-referenced tests in instructional settings. *Journal of Educational Measurement*, 15, 277-290.
- Larkin, J. H. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11, 65-99.

- Lewis, D. M., Mitzel, H. C., and Green, D. R. (1996, June). Standard setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behavioral anchoring*. Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.
- Loomis, S. C., and Bourque, M. L. (2001). From tradition to innovation: Standard setting on the National Assessment of Educational Progress. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Lawrence Erlbaum.
- Loomis, S. C., and Hanick, P.L. (2000). *Developing achievement levels for the 1998 NAEP in civics: Final report*. Iowa City, IA: ACT.
- Masters, G. N., Adams, R., and Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, 21, 595-609.
- Mitzel, H. C., Lewis, D. M., Patz, R. J., and Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum.
- National Assessment Governing Board. (2006). *Economics framework for the 2006 National Assessment of Educational Progress*. Washington, DC: Author.
- National Center for Education Statistics. (2007). *The nation's report card. 12th grade reading and mathematics 2005. National Assessment of Educational Progress*. NCES 2007-468. Washington, DC: U.S. Department of Education.
- Popham, W. J. (1978). As always provocative. *Journal of Educational Measurement*, 15, 297-300.
- Schulz, E. M., Kolen, M., and Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement*, 23, 347-362.
- Schulz, E. M., Lee, W., and Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement*, 42, 1-26.
- Shen, L. (2001, April). *A comparison of Angoff and Rasch model based item map methods in standard setting*. Paper presented at the Annual Meeting of the American Educational Research Association, Seattle, WA.
- Shepard, L. A. (1994, October). Implications for standard setting of the National Academy evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the Joint Conference on Standard Setting for Large-Scale Assessment*. Washington, DC: National Assessment Governing Board and National Center for Education Statistics.
- Stone, G. E. (2001). Objective standard setting (or truth in advertising). *Journal of Applied Measurement*, 2, 187-201.

- Stone, M. H., Wright, B. D., and Stenner, A. J. (1999). Mapping variables. *Journal of Outcome Measurement*, 3, 308-322.
- Surowiecki, J. (2004). *The wisdom of crowds*. New York: Doubleday.
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.
- Wainer, H. (1992). Understanding graphs and tables. *Educational Researcher*, 21(1), 14-23.
- Wang, N. (2003). Use of the Rasch IRT model in standard setting: An item-mapping method. *Journal of Educational Measurement*, 40, 231-253.
- Williams, N. J., and Schulz, E. M. (2005, April). *An investigation of response probability (RP) values used in standard setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada.
- Yin, P., and Sconing, J. (2008). Estimating variability of cut scores for item rating and Mapmark procedures: A generalizability theory approach. *Educational and Psychological Measurement*, 68, 25-41.
- Yin, P., Schulz, M., and Sconing, J. (2005, April). *A comparison of cut scores and cut score variability from Angoff-based and bookmark-based procedures in standard setting*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego, CA.