Running head: MAPMARK

The Mapmark Standard Setting Method

E. Matthew Schulz

ACT, Inc.

Howard C. Mitzel

Pacific Metrics

Portions of this research paper have been presented to the National Assessment Governing Board for the National Assessment of Educational Progress.

Correspondence concerning this article should be addressed to E. Matthew Schulz, Department of Statistical Research, ACT, Inc., Iowa City, IA, 52243-0168. Email: matt.schulz@act.org

A new standard setting method, Mapmark, was recently developed by ACT Inc. in the course of a contract with the National Assessment Governing Board (NAGB) to set achievement levels for the 2005 National Assessment of Educational Progress (NAEP) in Grade 12 mathematics. Mapmark includes elements of the bookmark method (Lewis, Mitzel, & Green, 1996; Mitzel, Lewis, Patz, & Green, 2001), item maps (Masters, Adams, & Loken, 1994), and expected percent correct scores on clusters of assessment items representing areas of knowledge, skills, and abilities called "domains" (Schulz, Lee, & Mullen, 2005). The present paper provides a detailed description of the Mapmark method, as implemented in the Achievement Level Setting (ALS) meeting conducted for NAGB, and presents results bearing on the procedural validity of the method. Procedural validity is assessed through panelists' responses to process evaluation questionnaires and by other data collected in the meeting. It is concluded that the Mapmark process exhibits evidence of procedural validity.

INTRODUCTION

Achievement levels on the National Assessment of Educational Progress (NAEP) are intended to help teachers, parents, educators, and the general public understand how students in the United States are performing on the NAEP relative to what students should know and be able to do. Public Law 100-279 mandates the National Assessment Governing Board (NAGB) to identify "appropriate achievement goals for each grade or age in each subject area to be tested…" under the National Assessment. NAGB policy specifies three achievement levels—Basic, Proficient, and Advanced—and states that their purpose is to make NAEP data more understandable to the general user, parents, policymakers, and educators alike.

In the course of a recent 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 developed a new standard setting procedure called Mapmark. Mapmark was named for the role of item maps (Masters, Adams, & Loken, 1994) and significant elements of the bookmark standard setting method (Mitzel, Lewis, Patz, & Green, 2001) in its process. As implemented in this project, however, Mapmark also incorporates domain score feedback (Schulz, Lee, & Mullen, 2005). The use of domain score feedback was initially regarded as optional, to be used only if ACT's research indicated that it could be successfully incorporated into a more basic bookmark method supplemented with item maps. ACT's research showed that domains could be successfully incorporated into the standard setting process. The role of domain score feedback in Mapmark may ultimately be regarded as the most significant, new feature of the method.

The remainder of this section presents a general rationale for the use of item maps, elements of the bookmark method, and domains in the Mapmark standard setting method and explains how these components were related in the method.

A "map," broadly speaking, is a spatially representative display by which one can interpret one's distance from a destination and from points of interest along a journey. An item map essentially represents the journey to higher achievement, with items being 'markers' or points of interest along the way. Direction indicates whether a student has 'passed' the item or not. Distance on the number line represents "how far" the student has traveled since passing the item or how far the student has to go before passing the item. In educational terms, "how far" is synonymous with "how easy" or "how hard" the skill represented by the item is expected to be for the student. "Passing" the item is synonymous with "mastery" of or "being able to do" the item.

Item-response theory models are used to construct item maps. Items are located on a number line that represents both item difficulty and student achievement. The criterion for locating an item on the number line, or scale, can be the item's difficulty statistic in an item response theory (IRT) model (Wright & Masters, 1982), the value of the IRT student ability parameter (θ) at which the item information function is maximal (Huynh, 1998), or the θ associated with a given probability of answering the item correctly (Kolstad, Cohen, Baldi, Chan, DeFur, & Angeles, 1998; Zwick, Senturk, Wang, & Loomis, 2000). All mapping criteria ultimately translate to a certain probability that a student has of answering the item correctly when the student's location on the scale is the same as the item's. "Mastery" or "being able to do" the item, or the skill represented by the item, thus corresponds to having a certain probability or higher of answering the item correctly.

Item maps have been used in standard setting previously (Gross & Wright, 1986; Engelhard & Gordon, 2000; Stone, 2001; Wang, 2003; Shen, 2001). In the applications cited, items were located by their difficulty parameter in the one-parameter Rasch model. A student whose achievement scale value is the same as the value of an item's difficulty parameter in the Rasch model has a 0.5 probability of correctly answering the item. Aside from this commonality, however, no uniformity has emerged in how item maps have been used in standard setting.

In bookmark, items are presented in an ordered item book (OIB) in the same sequence they would be found on an item map using a given mapping criterion. The OIB structures test content from easy to difficult for panelists, to facilitate judgments of what students should know and be able to do. A convention is emerging around using a 0.67 or 2/3 probability as the "Response Probability" criterion (RP or mapping criterion) in bookmark. This probability is associated with "mastery" of the skill or item. In the bookmark kernel, each panelist divides the items in the ordered item book into two groups—those that he/she feels a student at the passing standard, or lower borderline of an achievement level, should have mastery of and those that are too difficult for this expectation.

Bookmark and item mapping approaches to standard setting are often essentially the same. In one item mapping procedure (Shen, 2001), panelists studied the progression of knowledge, skills, and abilities as one goes from a low-to-high achievement direction on the item map and selected a scale value to represent the passing standard. Panelists simply drew a "mark" on the item map to represent the passing standard. The scale value/standard was selected with the intention that a student at the passing standard should have at least a 0.5 probability of correctly answering any item that maps below the standard. This item mapping method has the same kernel as the bookmark method, differing only in whether a scale value or an item is used to represent the passing standard. Whether the response probability used to order the items in bookmark, or locate items on the item map, is 0.5, 0.67, or some other value may be considered an open choice in either method.

Proponents of item mapping methods, like bookmark, and similar approaches to standard setting generally believe that panelists' efforts should be focused primarily on content rather than on individual item-level judgments that are statistically aggregated to imply a cut score on the test scale. Stone (2001) criticizes approaches that require panelists to make probability judgments for each item, based on their understanding of how a borderline student "would" perform. He claims that panelists are not experts in making probability judgments, and that their judgments should be primarily content-based. In Stone's standard setting method, panelists classify items as essential or non-essential "can do" items according to the performance standard. [Interestingly, this classification does not have to agree with the empirical difficulty-order of the items.] The passing standard is the average scale value of the essential items. This method does not require any sort of probability judgment but, like the mapping criterion in item mapping and the RP criterion in bookmark standard setting, does require examinees to demonstrate a certain level of performance on the "essential" items. No standard setting method can circumvent such a requirement.

ACT preferred to use a method more like that of Shen (2001), because it is more compatible with the bookmark method and it allows the possibility of incorporating the response probability criterion—the criterion that determines what level of performance is required on items in order to meet the performance standard—more explicitly into panelists' judgments. The bookmark probability judgment is greatly simplified when the RP criterion and student performance data are used to order the items on the item map and/or in the OIB. It is not necessary for panelists to make a probability estimate for each and every item. Rather, panelists can focus on a range of items that they feel are appropriate for the borderline. ACT also supposed that a non-extreme probability equal to a simple fraction such as 1/2 (0.5) or 2/3 (0.67) would be easier for panelists to understand and work with. ACT's reasoning was that probability judgment in the task should not overwhelm the content component, which involves aligning the performance standard with the progression of knowledge, skills, and abilities (KSAs) in the assessment.

Since the bookmark method was introduced in 1996 (Lewis, Mitzel, & Green, 1996) it has become the most widely used standard setting method in state assessments (CCSSO, 2001). One reason for the popularity of the bookmark method is its dependence on a KSA review of test items. In the KSA review, panelists identify and discuss the knowledge, skills, and abilities required by test items in the context of the OIB. In this context, panelists develop an understanding of student achievement as a progression of increasing knowledge, skills, and abilities. This understanding is useful in aligning performance standards with performance on a test, particularly when working with a series of achievement levels (e.g., Basic, Proficient, and Advanced) that also represent a progression of KSAs.

The KSA review also prepares panelists for the use of domains and domain-score feedback in the Mapmark process. The KSA review is performed in Round 1 in Mapmark and in the standard bookmark method. In Mapmark, domains and domain score feedback are introduced in Round 2. To fully understand the criterion-referenced meaning of a test or domain score, one must look at a representative sample of the items that were used to obtain the score. Panelists therefore perform a task in Round 2 of Mapmark, in which they consider how representative samples of items fit into the particular domain into which they have been classified. This task, like the KSA review, focuses panelists' attention on item content, particularly with regard to similarities and differences among test items and general patterns of content in the test. Similarities among items classified into the same domain may have already been noted by panelists in the KSA review. In practical terms, having performed the KSA review in Round 1, Mapmark panelists can perform the domain-content review in Round 2 (also called Domain Task 1) relatively quickly.

Item maps were added to the Mapmark process to provide a more tangible representation of differences in item difficulty and achievement level boundaries. The OIB is sufficient for dividing test items into two groups using a response probability or item-mapping criterion. The traditional bookmark method supplements the OIB with a table that contains additional information about the items, in the order that items appear in the book. The table contains item scale values (the location they would have on an item map). But difficulty differences between items are more difficult to keep track of and can easily be ignored by panelists in their tasks unless an item map is used. The spatial representation of difference, or lack of difference, between items helps panelists keep track of the magnitude of progression in KSAs represented by test items. When multiple achievement levels are being set (e.g., Basic, Proficient, and Advanced), the representation of achievement level boundaries on the map also helps panelists keep track of the magnitude of progression in KSAs represented by the achievement levels.

In a novel development for standard setting, item maps in the Mapmark method are also used to help panelists keep track of similarities in item content. This is done by arranging items into columns corresponding to more specific areas of content. One element of the KSA review involves identifying what additional KSAs may be required by an item that were not required by easier items that represent similar content. The "Primary Item Map," which is used by panelists in the Mapmark KSA review, facilitates this task by organizing items into columns representing subscales of the assessment. For the Grade 12 mathematics assessment in NAEP, the subscales are 1) Number Properties and Operations, 2) Measurement and Geometry, 3) Data Analysis and Probability, and 4) Algebra and Functions. The Measurement and Geometry Subscale is a combination of two content areas in the assessment framework. The other subscales correspond to individual content areas.

The motivation for incorporating domain-score feedback into the Mapmark method came from a previous study (Schulz, Lee, & Mullen, 2005). The Schulz, et al., study addressed the criticism that individual test items do not provide reliable support for inferences about mastery of skills and areas of content more general than a single test item (Forsyth, 1991). This criticism implies that panelists in a pure bookmark procedure could be misled into thinking that the cut score they have recommended requires mastery of a particular skill because an item representing that skill lies below their cut score on an item map or below their bookmark in the OIB. Domains are used in the Mapmark process to help panelists appreciate the unreliability of inferences they typically attach to individual test items and to provide a more reliable basis for inference.

Table 1 shows the titles of the teacher domains that were ultimately used in Mapmark Grade 12 standard setting activities. A total of twenty-three teacher domains were defined using methods similar to those of Schulz, Lee, and Mullen (2005). Teacher domains were defined within subscales of the assessment framework. The number of teacher domains per subscale ranged from four (in Number Properties and Operations) to eight (in Measurement and Geometry). These were organized into a total of sixteen score domains as shown in the table.

No more than two teacher domains were combined into the same score domain. Many teacher domains were large enough and/or distinct enough to stand alone as score domains.

Figure 1 shows the domain definition for teacher domain M4 in the Measurement and Geometry subscale. Mapmark panelists read and referred to the domain definitions for various purposes. For easier reference, the panelists were given a table that consisted of only the domain titles and narratives. But the sample items were helpful to panelists when answering the question, "I see how this item fits with other items in this domain." To answer this question, panelists referred not only to other items in the 2005 assessment that were classified into the same domain, but also to the sample items.

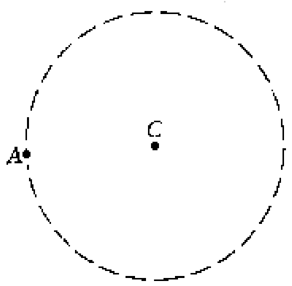
Table 1: Titles of Teacher Domains and the Correspondence between Teacher and Score Domains by Subscale of the 2005 Assessment

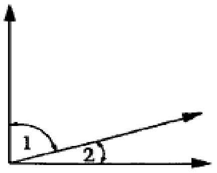
Teacher Domain	Title	Score Domain
Number Properties and Operations		
N1	Perform Basic Operations	N--1
N2	Determine Correct Operations	N--2
N3	Place Value and Notation	N--3
N4	Multistep Problems	N--4
Measurement/Geometry		
M1	Basic Measurement	M--1
M2	Symmetry, Motion, and Proportionality	M--2
M3	Identifying 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	

Domain M4: Angles

Items in this domain involve obtaining degree measures of angles through direct measurement or through knowledge about degree measures, such as the sum of angle measures in triangles or regular polygons, or the properties of angles formed by intersecting lines. Some items may require students to use rulers or protractors to draw figures having specified shapes or angle 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.



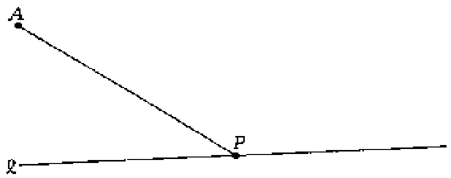


33. The sum of the measures of angles 1 and 2 in the figure above is 90°. What is the measure of the angle formed by the bisectors of these two angles?

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. Domain Definition for Teacher Domain M4.

THE ACHIEVEMENT LEVEL SETTING MEETING

This section provides a relatively detailed description of the Mapmark method as implemented in the Achievement Level Setting (ALS) meeting for the 2005 NAEP in Grade 12 mathematics. More complete details and results will be provided in forthcoming documentation required by ACT's contract with NAGB. Actual results of the process, including achievement level descriptions, cut scores, and the percentage of students in the achievement levels are not presented here because they must be kept confidential until the achievement levels are set by formal action of the NAGB. To protect this confidentiality, the achievement scale used in the ALS meeting and in the figures and tables presented in this paper is not the same scale that will be used to report the assessment results.

Methods not Specific to Mapmark

NAGB Policy

The achievement level descriptions used in the ALS meeting were developed prior to the meeting. The ALS meeting is viewed as a process of "translating" the ALDs into cut scores. This is in keeping with current NAGB policy, which specifies two stages to the NAEP Achievement Level Setting (ALS) process. In Stage 1, grade-specific and subject-specific achievement level descriptions (ALDs) are developed from general policy definitions. In Stage 2, the ALDs are translated into cut scores.

Panelists

Thirty-one panelists participated in the ALS meeting. The percentage of panelists by type were very close to targeted percentages of 55%, 15%, and 30% for, respectively, teachers, non-teacher educators, and general public. The ALS panelists were nationally recruited by methods that included stratified random sampling of school districts and consideration of qualifications such as professional accomplishment, teaching excellence, and community service. 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.

Design Factors

Groups and Tables were design factors in the ALS meeting. Group A and Group B worked with different but equivalent and overlapping item pools. Each pool contained about 60% of the items in the 2005 assessment pool. Combined, they represented 100%. There were 15 panelists in Group A and 16 panelists in Group B. Each group was further divided into three tables of five or six panelists each. The demographic attributes of panelists were considered when assigning members to groups and tables; otherwise the assignments were random. The goal was to have groups as equal as possible with respect to panelist type, gender, region, and race/ethnicity.

Schedule

The ALS meeting lasted four days, November 12-15, 2004 (Friday to Monday). It was conducted at the Westin Hotel in St. Louis. 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. The schedule is shown in Appendix B.

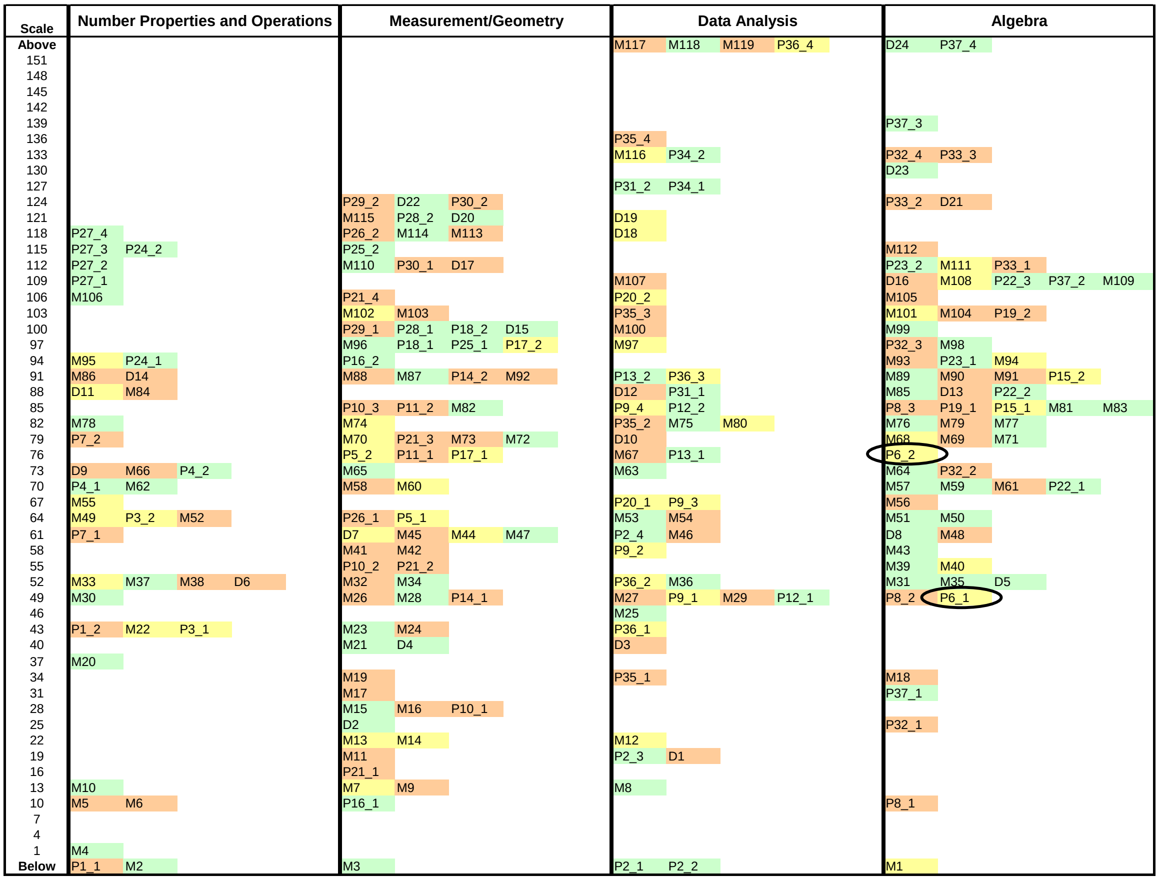
General Orientation

Orientation activities not specific to the Mapmark process began with mailings to panelists before the meeting and included all activities conducted the morning of the first day. Advance materials included a briefing booklet (that described the tasks and materials of the ALS meeting), the framework for the assessment, and the achievement level descriptions. Orientation activities during the meeting included a presentation on NAEP and NAGB by a NAGB staff member, an overview of achievement level setting, and taking a form of the NAEP exam.

Mapmark Methods

Orientation to Method, Study Design, and Materials

Panelists received an overview of Mapmark methods and materials in a 60 minute presentation. The presentation described item maps, domains, and the ordered item booklet (OIB). The study design, though not specific to Mapmark, was also reviewed. The role of the mapping criterion in the process was explained and its value and interpretation (mastery) was made clear.



Scale	Number Properties and Operations	Measurement/Geometry	Data Analysis	Algebra
Above			M117 M118 M119 P36_4	D24 P37_4
151				
148				
145				
142				
139				P37_3
136			P35_4	
133			M116 P34_2	P32_4 P33_3
130				D23
127			P31_2 P34_1	
124		P29_2 D22 P30_2		P33_2 D21
121		M115 P28_2 D20	D19	
118	P27_4	P26_2 M114 M113	D18	
115	P27_3 P24_2	P25_2		M112
112	P27_2	M110 P30_1 D17		P23_2 M111 P33_1
109	P27_1		M107	D16 M108 P22_3 P37_2 M109
106	M106	P21_4	P20_2	M105
103		M102 M103	P35_3	M101 M104 P19_2
100		P29_1 P28_1 P18_2 D15	M100	M99
97		M96 P18_1 P25_1 P17_2	M97	P32_3 M98
94	M95 P24_1	P16_2		M93 P23_1 M94
91	M86 D14	M88 M87 P14_2 M92	P13_2 P36_3	M89 M90 M91 P15_2
88	D11 M84		D12 P31_1	M85 D13 P22_2
85		P10_3 P11_2 M82	P9_4 P12_2	P8_3 P19_1 P15_1 M81 M83
82	M78	M74	P35_2 M75 M80	M76 M79 M77
79	P7_2	M70 P21_3 M73 M72	D10	M68 M69 M71
76		P5_2 P11_1 P17_1	M67 P13_1	P6_2
73	D9 M66 P4_2	M65	M63	M64 P32_2
70	P4_1 M62	M58 M60		M57 M59 M61 P22_1
67	M55		P20_1 P9_3	M56
64	M49 P3_2 M52	P26_1 P5_1	M53 M54	M51 M50
61	P7_1	D7 M45 M44 M47	P2_4 M46	D8 M48
58		M41 M42	P9_2	M43
55		P10_2 P21_2		M39 M40
52	M33 M37 M38 D6	M32 M34	P36_2 M36	M31 M35 D5
49	M30	M26 M28 P14_1	M27 P9_1 M29 P12_1	P8_2 P6_1
46			M25	
43	P1_2 M22 P3_1	M23 M24	P36_1	
40		M21 D4	D3	
37	M20			
34		M19	P35_1	M18
31		M17		P37_1
28		M15 M16 P10_1		
25		D2		P32_1
22		M13 M14	M12	
19		M11	P2_3 D1	
16		P21_1		
13	M10	M7 M9	M8	
10	M5 M6	P16_1		P8_1
7				
4				
1	M4			
Below	P1_1 M2	M3	P2_1 P2_2	M1

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

Figure 2 shows sections of an actual Primary Item Map that was used to illustrate key information about materials. The scale is altered to maintain confidentiality of cut scores. Items were represented on item maps 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 polytomously-scored items include an underline '_' followed by the score level. Polytomously-scored items were ordered by the difficulty of their last score level.

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 color of an item handle on the map indicates whether it is in the Group A pool only (tan), the Group B pool only (green) or in both item pools (yellow). Item P6 was in both item pools. Items in both pools are "common" items.

Round 1

Round 1 began with a presentation on the NAEP framework, followed by a review of the knowledge, skills, and abilities required by test items (KSA review), a presentation on the achievement level descriptions, and finally the bookmark placement task.

KSA Review Panelists spent the next nine hours of meeting time identifying the knowledge, skills, and abilities students must have in order to earn successively higher scores on the test. There were four components to this activity.

1. *KSA Activity 1*. This was a whole group KSA review, led by the bookmark content facilitator, in which panelists were trained in the process of identifying KSAs required by constructed response items. They began with a few dichotomously-scored items common to both Group item pools, then proceeded to look at polytomously-scored items common to both item pools. For each polytomously-scored item, the activity involved identifying the *additional* KSAs needed to earn successively higher scores on the item.
2. *KSA Activity 2*. This was a table-group KSA review in which panelists continued to apply the process begun in the whole group to the remaining polytomously-scored items, unique to their item pool. Panelists took turns 'leading' this activity at their table. Content and process facilitators circulated among the tables.
3. *KSA Activity 3*. This was an independent KSA review in which panelists identified the KSAs required by all of the items in their pool in the context of their Ordered Item Booklet (OIB). They considered items sequentially, 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.
4. *KSA Activity 4*. This was a table-group discussion of the KSAs in the context of the OIB. Again, items were considered sequentially, beginning with the easiest. Panelists shared their ideas about the KSAs and recorded additional notes.

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 were presented first in the booklet, and were the first covered in KSA Activity 1. Within each type, items were listed in order of difficulty.

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 within the CROIB. Items were separated by tabbed pages, with the tab showing the item handle (minus the score points). Item information included the scoring rubric 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.

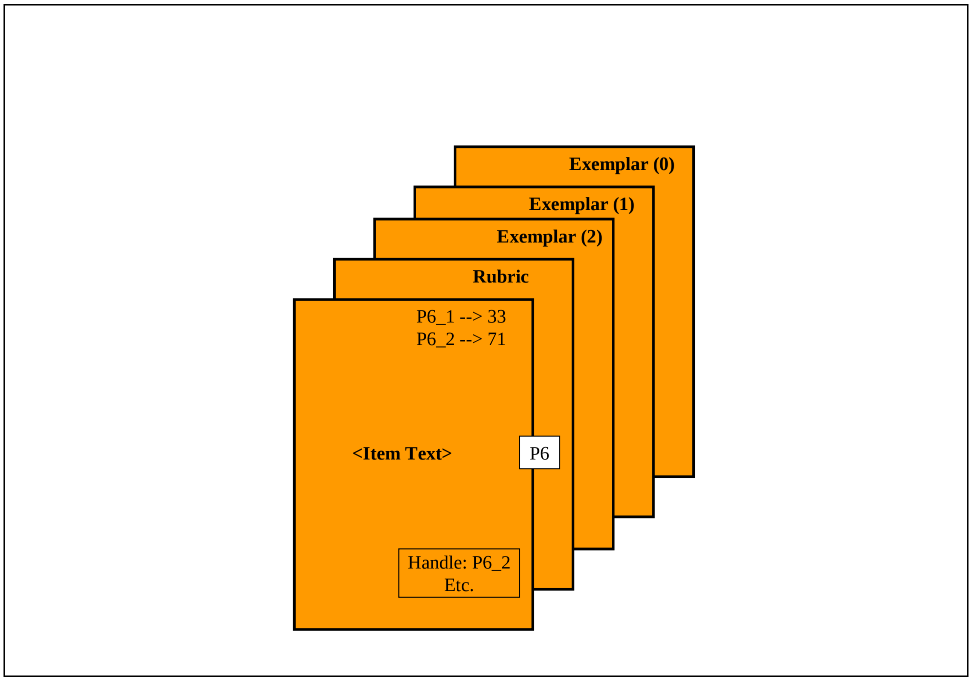


Figure 3. Slide illustrating contents of the CROIB (Constructed Response Ordered Item Book)

Panelists used large yellow stickies to record their notes on the KSAs. They were told that their notes were for their own use. They used one sticky for each score point. When panelists were finished with an item, they placed their notes in the Note-template. This was a stapled set of legal size pages with outlines for accommodating six stickies per page. Within each sticky-outline was an item handle and OIB page number identifying the sticky that was to be placed there. Stickies were positioned in the Note-template in order of the OIB page number on which it was to be placed at the beginning of KSA Activity 3.

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. The use of the Note-template allowed panelists to place their notes on the polytomously-scored item steps on the correct OIB page numbers with just one pass through the OIB.

When panelists see score points of polytomously-scored items relative to the difficulty of all other items in their pool in KSA Activity 3, they can add to their notes observations about what KSAs the score point may require that previous, easier items and score points did not require. Panelists recorded further notes directly on the pages of the OIB.

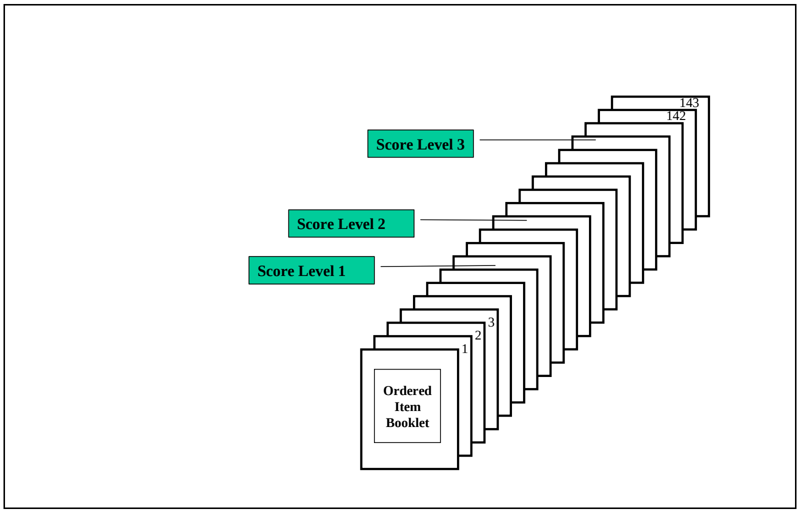
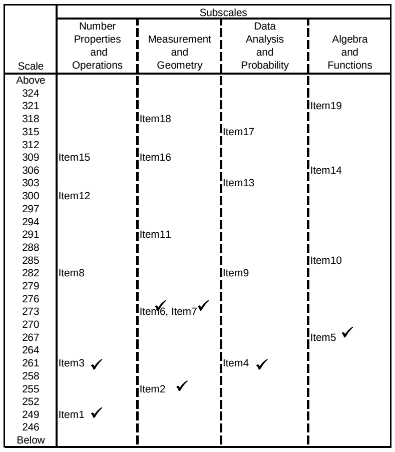


Figure 4. Score levels of a polytomously-scored item are treated as separate items and appear at different places in the OIB.

Panelists checked items off on their Primary Item Map as they progressed through the OIB. Figure 5 is a simplified illustration of the item check-off process on the Primary Item Map. The item check-off process helped panelists see "how much" more difficult one item was than another and which items were related in terms of the general KSAs that distinguished different subscales.



Scale	Subscales			
	Number Properties and Operations	Measurement and Geometry	Data Analysis and Probability	Algebra and Functions
Above				
324				
321				Item19
318		Item18		
315			Item17	
312				
309	Item15	Item16		
306				Item14
303			Item13	
300	Item12			
297				
294				
291		Item11		
288				
285				Item10
282	Item8		Item9	
279				
276				
273		Item6 ✓, Item7 ✓		
270				
267				Item5 ✓
264				
261	Item3 ✓		Item4 ✓	
258				
255		Item2 ✓		
252				
249	Item1 ✓			
246				
Below				

Figure 5. Simplified item map illustrating results of item check-off procedure as panelist progresses through OIB up through Item 7 in KSA Activity 3.

In the Table-group discussion (KSA Activity 4) panelists shared their ideas about the KSAs and added the ideas of other panelists to their notes. Panelists took turns leading the table discussion. The process was monitored by facilitators to reinforce the idea that all panelists have something valuable to contribute to the process.

When the KSA review was complete, panelists had a detailed, *structured* understanding of the assessment and student achievement. Structure is provided by the difficulty-order of knowledge, skills, and abilities required by test items as shown in the OIB and on the Primary Item Map. This structure prepares panelists to understand the continuum of increasing knowledge, skills, and abilities represented by the achievement level descriptions—Basic, Proficient, and Advanced.

<u>Understanding the Achievement Level Descriptions</u> Panelists had been instructed to study the achievement level descriptions prior to the meeting. To reinforce this learning, the primary 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 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 to 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 for reasons of maintaining independence of judgment.

Placing the Bookmarks The bookmark placement task began with a carefully scripted presentation on the following points:

- The ALD should be thought of as representing a *range* of performance on the achievement scale,
- 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. The *structure* provided by the OIB and Primary Item Map made it possible for panelists to develop and apply a concept of borderline in the *process* of placing their bookmarks.

The bookmark placement task is initially described to panelists as a process of going through the OIB, beginning with the easiest item, until they come to an item that they judge to be too difficult for mastery by the borderline student. Mastery is defined as having at least a 0.67

probability of answering the item correctly. The bookmark is placed on the item immediately preceding the "too difficult" item. Figure 6 illustrates a bookmark placement.

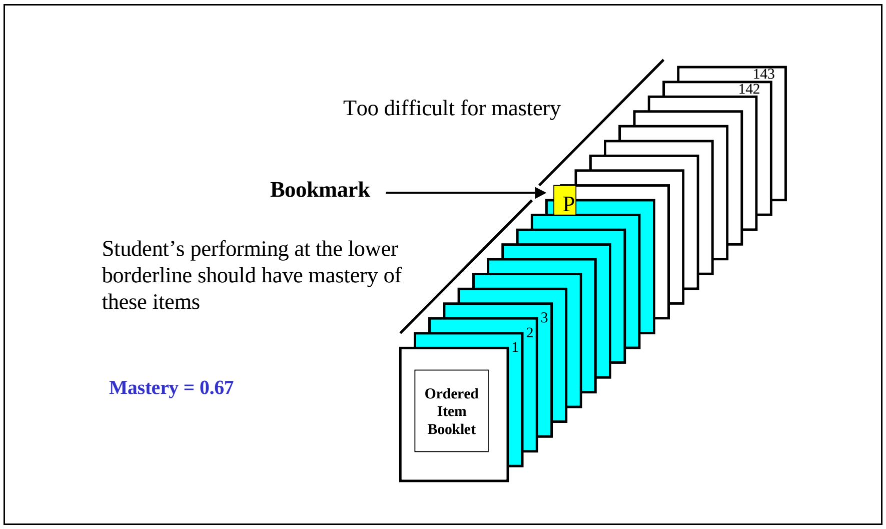

Too difficult for mastery
Bookmark
P
Student's performing at the lower borderline should have mastery of these items
Mastery = 0.67
143
142
3
2
1
Ordered Item Booklet

Figure 6. Bookmark placement task simplified.

Once panelists have this basic idea, 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.

The initial description of the process is then supplemented with the instruction to go further, 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. This instruction is represented to panelists visually by showing a "range of uncertainty" in a slide-depiction of the OIB. All items below this range are "sure mastery" items. All items above this range are "sure non-mastery" items. Figure 7 shows a slide that was used to illustrate this concept for panelists.

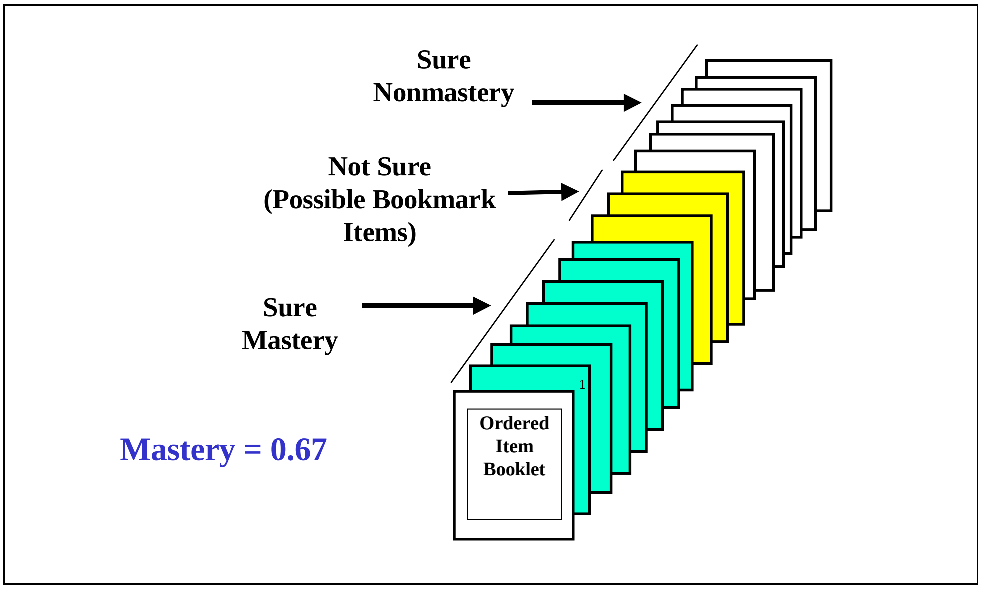


Figure 7. Slide illustrating range of uncertainty in bookmark placements.

Bookmark placements were done one achievement level at a time starting with Proficient, then Basic, then Advanced. Panelists read the ALD for the given level and used only that ALD to place the corresponding bookmark. The next achievement level was not started until all panelists had finished their placements for the previous one.

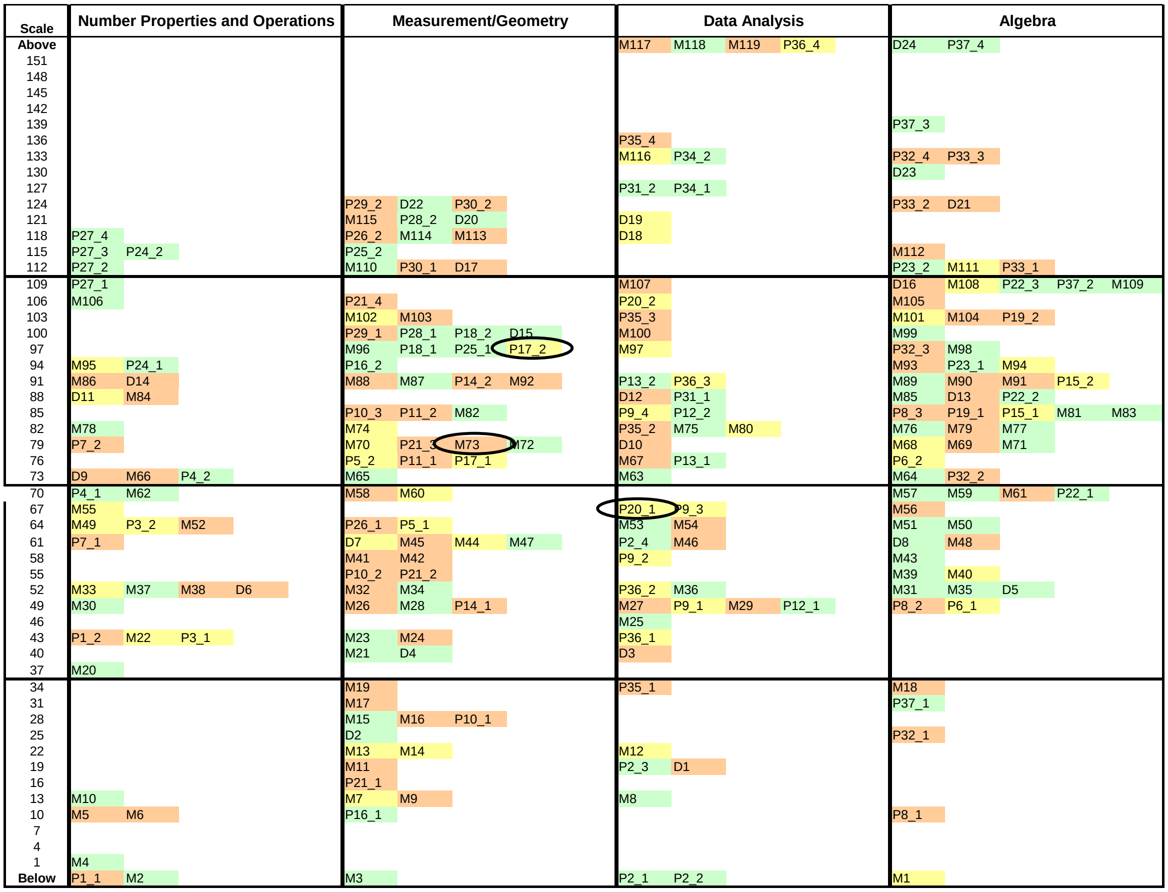
After placing all bookmarks, panelists were given an opportunity to adjust their bookmark placements. Panelists were encouraged to look at all of the ALDs together and to consider whether the differences between their bookmark placements were consistent with the increments of achievement implied by the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

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. The computer program computed the median cut score for each achievement level.

Round 2

Feedback. Feedback after Round 1 consisted of a) median cut scores, b) high and low cut scores, c) rater-location, and d) domain scores. In addition to providing the numerical values of cut scores, feedback was shown on item maps and domain score charts to focus panelists' attention on the intended, criterion-referenced meaning of cut scores.

Figure 8 shows how the median cut scores and a panelists' bookmarked items were marked on the Primary Item Map. Panelists were instructed to draw the median cut score lines on their maps. Lines were drawn beneath the midpoint of the interval containing the cut score.



Scale	Number Properties and Operations	Measurement/Geometry	Data Analysis	Algebra
Above			M117 M118 M119 P36_4	D24 P37_4
151				
148				
145				
142				
139				P37_3
136			P35_4	
133			M116 P34_2	P32_4 P33_3
130				D23
127			P31_2 P34_1	
124		P29_2 D22 P30_2		P33_2 D21
121		M115 P28_2 D20	D19	
118	P27_4	P26_2 M114 M113	D18	
115	P27_3 P24_2	P25_2		M112
112	P27_2	M110 P30_1 D17		P23_2 M111 P33_1
109	P27_1		M107	D16 M108 P22_3 P37_2 M109
106	M106	P21_4	P20_2	M105
103		M102 M103	P35_3	M101 M104 P19_2
100		P29_1 P28_1 P18_2 D15	M100	M99
97		M96 P18_1 P25_1 P17_2	M97	P32_3 M98
94	M95 P24_1	P16_2		M93 P23_1 M94
91	M86 D14	M88 M87 P14_2 M92	P13_2 P36_3	M89 M90 M91 P15_2
88	D11 M84		D12 P31_1	M85 D13 P22_2
85		P10_3 P11_2 M82	P9_4 P12_2	P8_3 P19_1 P15_1 M81 M83
82	M78	M74	P35_2 M75 M80	M76 M79 M77
79	P7_2	M70 P21_3 M73 M72	D10	M68 M69 M71
76		P5_2 P11_1 P17_1	M67 P13_1	P6_2
73	D9 M66 P4_2	M65	M63	M64 P32_2
70	P4_1 M62	M58 M60		M57 M59 M61 P22_1
67	M55		P20_1 P9_3	M56
64	M49 P3_2 M52	P26_1 P5_1	M53 M54	M51 M50
61	P7_1	D7 M45 M44 M47	P2_4 M46	D8 M48
58		M41 M42	P9_2	M43
55		P10_2 P21_2		M39 M40
52	M33 M37 M38 D6	M32 M34	P36_2 M36	M31 M35 D5
49	M30	M26 M28 P14_1	M27 P9_1 M29 P12_1	P8_2 P6_1
46			M25	
43	P1_2 M22 P3_1	M23 M24	P36_1	
40		M21 D4	D3	
37	M20			
34		M19	P35_1	M18
31		M17		P37_1
28		M15 M16 P10_1		
25		D2		P32_1
22		M13 M14	M12	
19		M11	P2_3 D1	
16		P21_1		
13	M10	M7 M9	M8	
10	M5 M6	P16_1		P8_1
7				
4				
1	M4			
Below	P1_1 M2	M3	P2_1 P2_2	M1

Figure 8. Primary Item Map showing Round 1 median cut scores (horizontal lines) and the location of panelist A1201's bookmarked items (circled).

Before panelists were shown domain score feedback, they were given a presentation on how and why the teacher domains and score domains were defined. The presentation included a brief overview of the domain development process and described the intended attributes of the teacher and score domains. (This information is described in detail in a forthcoming *Domain Development Report.*)

Expected percent correct curves based on subscales (Figure 9) were shown to illustrate that the subscales were not as widely separated in difficulty as desired for purposes of defining and differentiating achievement levels. Vertical lines in Figure 9 correspond to the Round 1 cut scores for Basic, Proficient, and Advanced. A hypothetical percent correct criterion for mastery (67%) is illustrated by the horizontal dashed line. One can see that the subscale domains do not differ enough in difficulty to distinguish among achievement levels and or to provide a very rich understanding of what students at each achievement level can or cannot do.

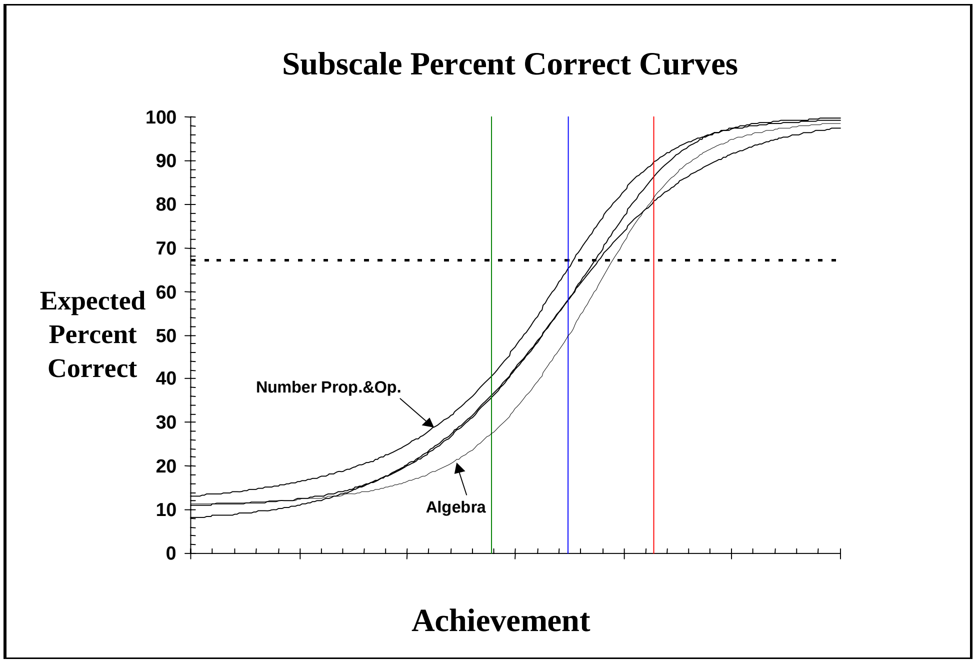


Figure 9. Expected percent correct curves based on subscales of the Assessment Framework.

Expected percent correct curves based on score domains defined within each subscale were shown to illustrate that the attributes of teacher and score domains were more useful for understanding the criterion-referenced meaning of the cut scores. Figure 10 shows the percent correct curves for the Data Analysis score domains. It can be seen that at least one domain is mastered (at a 67% criterion), and at least one domain is not mastered at each achievement level boundary.

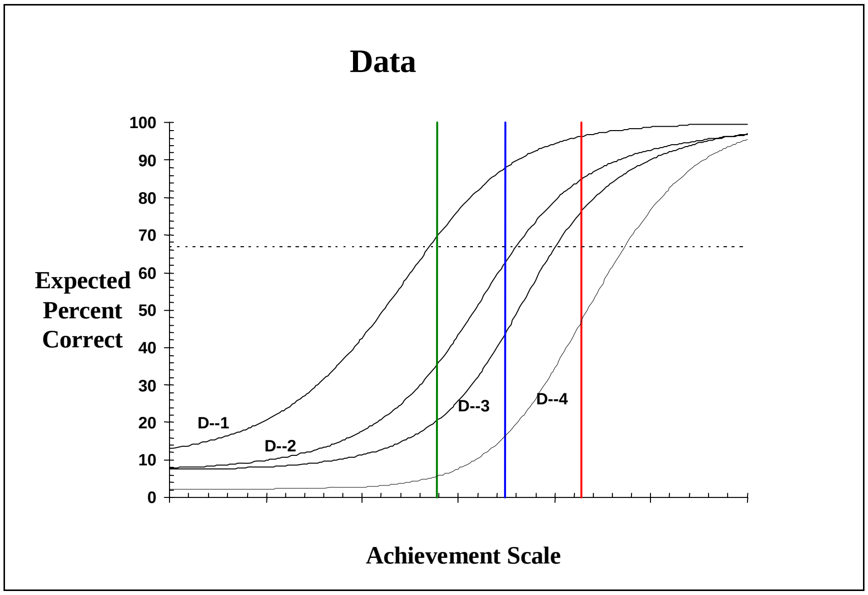


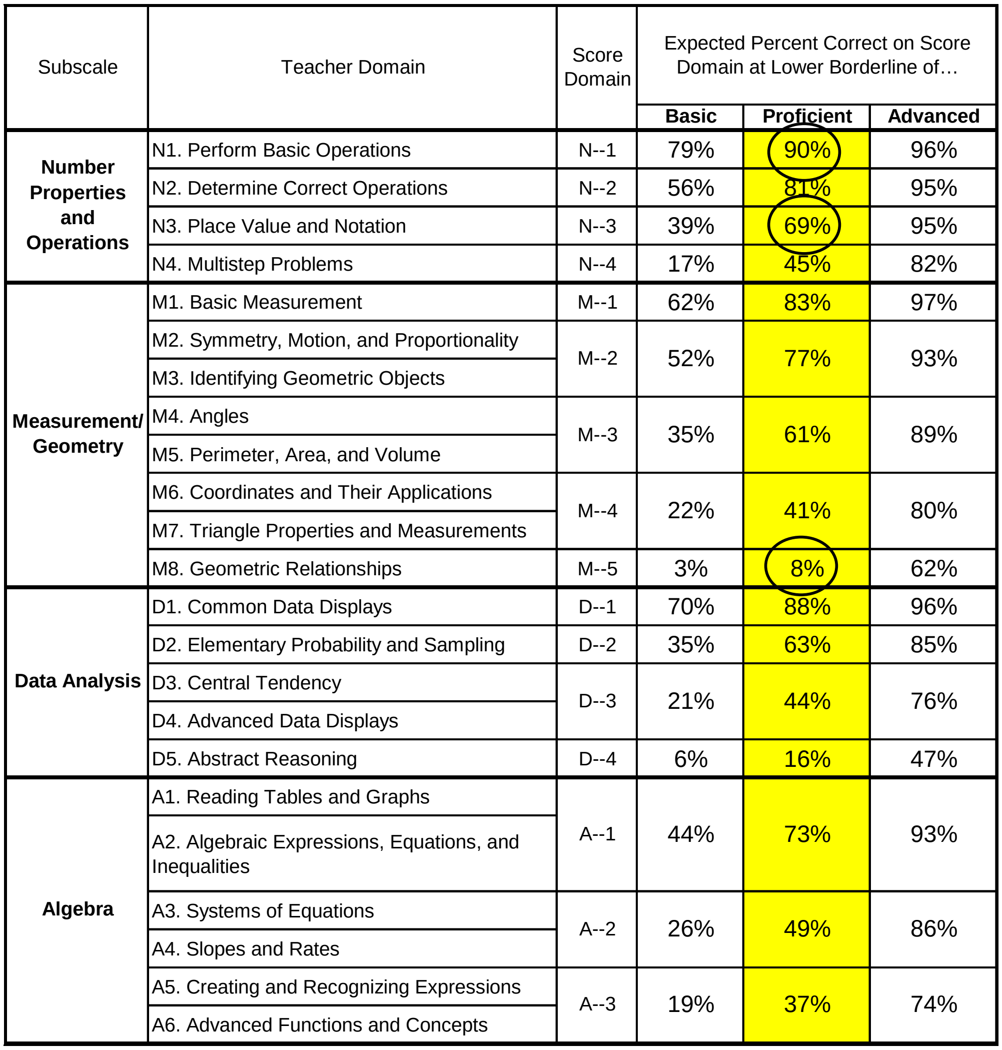
Figure 10. Percent correct curves for score domains in Data Analysis subscale, with vertical lines showing location of Round 1 cut scores and a horizontal line representing a 67% criterion for mastery.

A Percent Correct Table (PCT) was used to show the expected percent correct scores corresponding to the cut scores. The PCT for Round 1 cut scores is shown in Figure 11. This table shows the teacher domain titles and, for each score domain, the expected percent correct scores conditional on the lower boundary of the Basic, Proficient, and Advanced achievement levels, as defined by the median cut scores.

Panelists were told that their Round 2 cut score recommendations would be based on judgments of whether the domain scores were too low, OK, or too high for the borderline of an achievement level and that activities in Round 2 were designed to help them understand the domain scores and make judgments about whether the cut scores should be higher or lower than the Round 1 medians, based on the domain scores in the PCT.

The highest, lowest, and closest-to-67% domain scores for the Proficient cut score in the PCT were circled (see Figure 11) to draw panelists' attention to the fact that in one of their

Domain Tasks, they would be asked to make the "higher/OK/lower" judgment for each domain score in the table.



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

After panelists were told that they would be recommending cut scores based on whether they felt the domain scores in the PCT should be higher, lower, or were OK, they were 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 that goes from 10 points below the "low" cut score to 10 points above the "high" cut score from the previous round.

Figure 12 shows the DSC for the Proficient Achievement Level with the location of Panelist A1201 marked by a circle on the score scale. The median, high, and low cut scores were marked for panelists in the DSC as shown in the figure. Circles were also drawn around 67% domain scores within the range of the high and low cut scores. The percent correct scores in the "median" row correspond to the percent correct scores in the Percent Correct Table.

	Scale	Number Sense				Measurement					Data Analysis				Algebra		
	Score	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
	92	95	94	92	77	96	90	85	73	47	95	82	71	40	91	80	67
	91	95	93	91	76	95	90	85	72	45	95	82	71	39	90	79	66
	90	95	93	91	75	95	90	84	71	44	95	81	70	38	90	79	65
	89	95	93	90	74	95	90	84	70	42	95	81	69	37	90	78	64
High	88	95	93	90	73	95	89	83	68	40	95	80	68	36	89	77	63
	87	95	92	89	72	94	89	82	67	36	95	79	67	35	89	75	61
	86	94	92	88	71	94	88	81	65	34	94	79	66	34	88	74	60
	85	94	91	87	69	93	88	80	63	31	94	78	64	32	87	72	58
	84	94	91	86	67	93	87	78	61	27	94	77	63	30	86	70	56
	83	94	90	85	66	92	87	77	60	26	93	76	62	30	86	69	55
	82	93	90	84	64	92	86	76	58	23	93	75	60	28	85	67	54
A1201	81	93	89	83	63	91	85	75	57	21	93	74	59	27	84	66	53
	80	93	89	82	61	91	85	74	55	19	92	73	57	26	83	64	51
	79	93	88	80	59	90	84	72	53	17	92	72	56	24	82	62	49
	78	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	77	92	86	77	55	88	82	69	49	13	91	69	52	22	80	58	45
	76	91	85	75	53	87	81	67	47	11	90	68	50	20	78	56	43
	75	91	85	75	51	87	80	66	46	11	90	67	49	20	78	55	42
	74	91	84	73	49	86	79	64	44	9	89	66	47	19	76	53	41
	73	90	83	71	47	84	78	63	43	8	89	64	45	17	75	51	39
Median	72	90	81	69	45	83	77	61	41	8	88	63	44	16	73	49	37
	71	90	81	68	44	83	77	60	40	7	88	62	43	16	73	49	37
	70	89	80	67	42	82	75	58	39	6	87	60	41	15	71	47	35
	69	89	79	66	41	81	75	57	38	6	86	60	40	14	70	46	34
	68	88	78	64	39	80	74	56	37	6	86	58	38	14	69	44	33
	67	88	77	62	37	79	72	54	35	5	85	56	37	13	67	43	32
	66	88	76	61	36	78	72	53	35	5	84	55	36	12	66	42	31
	65	87	75	59	34	77	70	52	33	5	83	54	34	12	64	40	30
	64	87	73	58	32	76	69	50	32	4	83	52	33	11	63	39	29
	63	86	72	56	30	74	67	48	31	4	82	51	32	10	61	37	28
	62	85	70	54	28	73	66	47	30	4	81	49	30	10	59	36	27
	61	85	69	52	27	72	64	45	29	4	79	47	29	9	57	35	26
	60	84	67	51	25	71	63	44	28	3	78	46	28	8	56	33	25
	59	84	66	49	24	69	61	43	27	3	77	44	26	8	54	32	24
	58	83	64	47	23	68	60	41	26	3	76	43	25	8	52	31	23
	57	82	62	46	21	67	58	40	25	3	75	41	24	7	50	30	22
	56	82	62	45	21	66	57	39	25	3	74	40	24	7	49	29	22
Low	55	81	61	44	20	66	57	39	24	3	74	40	23	7	49	29	21
	54	81	60	43	20	65	56	38	24	3	73	39	23	7	48	28	21
	53	81	59	42	19	64	55	37	23	3	72	38	22	6	47	28	21

Figure 12. Domain Score Chart showing Round 1 results and location of panelist A1201 for Proficient achievement level.

The only information that panelists added to the DSC themselves was the location of their recommended cut score. Panelists were asked to draw a circle around their recommended cut score, as illustrated in the figure. For their cut score, they referred to the form they used to record their bookmark page number. Staff at the conclusion of Round 1 had written the corresponding scale values, and the form was returned to panelists at the beginning of the Round 1 feedback.

By circling their own cut score on the DSC, panelists were able to see how much difference there was between their cut score and the median both numerically and in criterion-referenced terms. Likewise, panelists could see the criterion-referenced meaning of the high and low cut scores and compare this to their own cut score and the median.

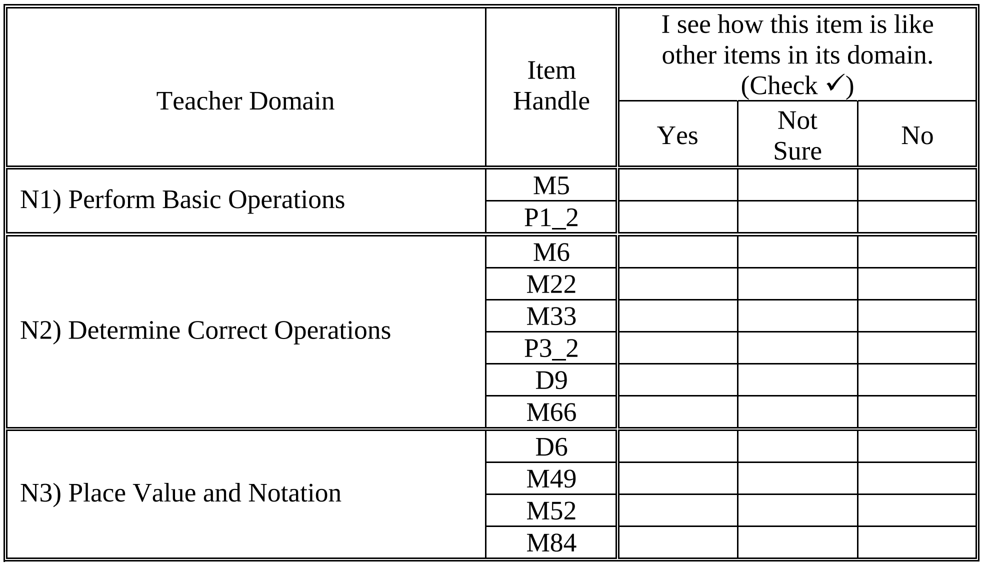
Similarly, the circles around a panelist's bookmarked items on the Primary Item Map, together with the horizontal lines representing the median cut scores, enabled each panelist to see how much difference there was between their individually-recommended cut score and the median cut score in terms of both scale distance and KSAs represented by test items.

Domain Task 1: Understanding Domain Scores One cannot understand a score on a test from the title and a description of the test alone. To truly understand a test score, one must look at the items or exercises that were used to obtain the score. 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, 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. All of these points help panelists understand their essential task of recommending cut scores.

The principal materials used in Domain Task 1 are a) a Domain Ordered Item Book, or DOIB, b) Domain Item Maps, and 3) the Domain Task 1 form. The DOIB contains the items in a panelist's pool in order of difficulty, within teacher domain. Teacher domains are presented in the DOIB in the order they are represented by columns from left to right on the Domain Item Map. This is in order of their difficulty within score domain, with score domains ordered by difficulty from left to right on the Domain Item Map.

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. The form for a Group (A or B) listed only the items in the Group's pool. Items were identified on the form by their handle. Polytomously-scored items were listed only once, and were identified by the highest score possible on the item (the last score point). Items were listed in order of their difficulty with the order of Polytomously-scored items determined by the scale value of their highest score point.



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			
	M66			
N3) Place Value and Notation	D6			
	M49			
	M52			
	M84			

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

Panelists responded to the question, "I see how this item is like other items in its domain" for each item in their pool that was classified into a teacher domain. 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.

Items were considered in the order they appeared on the form. Items were ordered by difficulty within Teacher Domain within Subscale. Teacher Domains were ordered by difficulty within Score Domain and Score Domains were ordered by their percent correct curves, or overall difficulty. Before considering the items within a given Teacher Domain, panelists read the narrative of the Teacher Domain definition and looked at the sample items (see Figure 1 for an example of a domain definition).

Materials for Domain Task 1 included a Domain Ordered Item Book (DOIB). The DOIB contained the teacher domain definitions and items in the Group's pool in the same order they appeared on the Domain Task 1 form. For items in the Group's pool, the DOIB contained a copy of the first page of the item's corresponding page in the OIB (for multiple choice and dichotomously-scored constructed response items) or the CROIB (for polytomously-scored items), plus the scoring rubric (for constructed response items).

TEACHER DOMAINS: Data Analysis **Name**________________________

Scale	Common Data Displays		Elementary Probability and Sampling		Central Tendency	Advanced Data Displays	Abstract Reasoning	
Above			P36_4		M118		M117	M119
151								
148								
145								
142								
139								
136					P35_4			
133						M116	P34_2	
130								
127							P31_2	P34_1
124								
121						D19		
118							D18	
115								
112								
109			M107					
106			P20_2					
103					P35_3			
100			M100					
97					M97			
94								
91			P36_3			P13_2		
88					D12		P31_1	
85			P9_4	P12_2				
82			M75		P35_2	M80		
79			D10					
76					M67	P13_1		
73			M63					
70								
67			P20_1	P9_3				
64			M53	M54				
61	P2_4	M46						
58			P9_2					
55								
52	M36		P36_2					
49	M27	M29	P9_1	P12_1				
46	M25							
43			P36_1					
40	D3							
37								
34					P35_1			
31								
28								
25								
22	M12							
19	P2_3	D1						
16								
13	M8							
10								
7								
4								
1								
Below	P2_1	P2_2						
Border Adv.:	96 %		85 %		76 %		47 %	
Border Prof.:	88 %		63 %		44 %		16 %	
Border Basic:	70 %		35 %		21 %		6 %	

Figure 14. Domain Item Map for Data Analysis and Probability Subscale.

Domain Item Maps were also used in the domain tasks of Round 2. Panelists were given one Domain Item Map for each subscale. Figure 14 shows the Domain Item Map for the Data Analysis and Probability subscale. Panelists observed the trend of increasing difficulty in the teacher and score domains as one goes from left to right in the Domain Item Map. Facilitators also drew panelists' attention 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.

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 were the same percent correct scores shown in the Percent Correct Table and highlighted on the Domain Score Charts.] 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 is 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. 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 in turn only a sample of the items that could be included in the domain. Therefore, the reported, expected percent correct score on a domain is itself an unreliable indication of student performance on the domain. The reliability of a performance index generally depends on the number of items used to obtain it, and is lowest for a single item.

The meaning of the 0.67 response probability criterion and of distance on the item map was enhanced for panelists by drawing their attention to the following:

- when items tended to lie below a cut score, the expected percent correct score on the items was above 67%

- when items tended to lie above a cut score, the expected percent correct score on the items was below 67%
- when items tended to be distributed equally above and below a cut score, the expected percent correct score on the items was about 67%

When panelists finished reviewing items belonging to teacher domains within a given subscale, they were shown a plot of expected percent correct curves for the subscale. Figure 10 shows the plot that was presented for the Data Analysis and Probability subscale. The plots were used to reinforce the idea that the ALDs represent a range of achievement and that panelists' must decide where the lower borderline of the achievement level should be. Panelist could see that the expected percent correct scores increase within an achievement level and that 'typical' performance within the level is usually quite different from performance at the lower borderline.

Panelists were prepared for Domain Task 1 by having performed the KSA review in Round 1. The KSA review taught panelists to see similarities, as well as differences, among items. The KSAs identified for an item might have been included in the domain title or narrative, or have seemed to be required by the sample items for a domain. Panelists may have noted the same KSAs for items classified into the same domain.

Domain Task 2: Evaluating the Domain Scores In Domain Task 2, panelists make 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 15 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.

Panelists could conceivably answer the Domain Task 2 question on the basis of whether they thought the domain score should be higher or lower than 67%. Scores of 67% were circled in the Domain Score Chart. Domain scores greater than or equal to 67% were highlighted in the Percent Correct Table. A horizontal line at 67% was marked on domain percent correct plots (see Figure 10, for example).

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					
	A5. Creating and Recognizing Expressions	A--3	37%			
	A6. Advanced Functions and Concepts					

Figure 15. Domain Task 2 Form for Proficient Achievement Level.

Panelists were encouraged to think more generally, however. 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 except that it did not require the panelists to state precisely what was acceptable, only to indicate whether an acceptable score was higher, lower, or about equal to the domain score associated with the Round 1 median.

Panelists' Domain Task 2 judgments were similar to their Round 1 bookmark placement judgments. As in Round 1, panelists used the ALDs to make their judgments. In Round 1, panelists made connections between item KSAs and the ALDs. In Round 2, panelists made connections between domain KSAs and the ALDs. In Round 1, panelists judged whether a 0.67 probability of getting an item correct was "good enough" for the lower boundary of an achievement level. In Round 2, panelists judged whether a given percent correct score on a domain was good enough for the lower boundary of an achievement level.

Instructions for Round 2 Cut Score Recommendations Panelists used the Domain Score Chart to choose a scale value for their Round 2 cut score recommendations. Instructions for this choice 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, one would probably want to recommend a cut score close to the median. If all of the checks were in the "higher" column, one would probably want to select a cut score higher than the Round 1 median.

Most instruction time concerned the case where judgments about appropriate domain scores do not agree with the patterns found in the Domain Score Chart. This was illustrated by an example of a form on which there were checks in both the "higher" and "lower" columns. Panelists were told they should use their own judgment to balance the many competing factors that exist in such cases. They were told to look to the ALDs for guidance as to which domains were most important, and to think about the percent correct scores that they felt were appropriate for these domains.

Some instructions panelists were given about deciding the relative importance of domains were based on technical considerations. Panelists were advised to give less importance to domains represented by smaller numbers of items, other things being equal, because domain scores derived from fewer items are less reliable. For similar reasons, panelists were told to give less importance to domains when the expected score is very high or very low and to focus on scores near 67%, or where the expected domain score changes most with change in the cut score.

Panelists were told also told that their Round 1 bookmark placement could be a factor in their Round 2 cut score recommendation. They had circled the scale value derived from their Round 1 bookmark placements on the Domain Score Chart. If the domain scores associated with their Round 1 cut score recommendation 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 derived from their Round 1 bookmark placement.

In making their Round 2 cut score recommendations, panelists were instructed to work independently. Beginning with Proficient, then Basic, then Advanced, panelists chose a scale value and recorded the scale value on their recommendation form. Panelists were instructed to circle the scale value they chose for their Round 2 cut score recommendation on their Domain Score Chart and to circle the map-interval containing the scale value on their Primary Item Map.

Round 3

Feedback At the beginning of Round 3, panelists were given a new Primary Item Map, a new Percent Correct Table, new Domain Score Charts, and their OIB. The new Primary Item Map was stapled on top of the maps they had used in the previous rounds, including their Round

1 Primary Item Map and their Domain Item Maps. The form panelists' used to record their Round 2 cut score recommendation was returned to them.

- *Numerical values*. 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.
- *Primary Item Map*. Panelists were instructed to draw horizontal lines across their 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 12 only this time to show the location of the Round 2 median, the highest and lowest recommended cut scores from Round 2, and 67% expected scores within the high/low range. Panelists circled their Round 2 cut score recommendations on the chart.
- *The OIB*. 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.

Whole-Group Discussion: Putting It All Together The whole group discussion was guided by a presentation during which questions were addressed to the whole group. The presentation was designed to increase understanding of both item-level information (the OIB) and domain-level information (the DSC) as related to the concept of borderline performance.

- The concept of borderline performance was reinforced by showing how percent correct curves increase across an achievement level. Panelists were asked if they were comfortable with the difference between borderline and typical performance within an achievement level;
- The idea that even very low domain scores, such as 20%, could represent some degree of knowledge, skill, and ability in a domain was illustrated with percent correct curves showing expected performance lower than 20% at the lowest end of the achievement scale.
- Panelists were reminded that they should not place too much importance on where their cut score lay with respect to a single item. Their work with domains reminded them that a skill worthy of consideration is broader than a single item, and that the difficulty of one item does not represent the difficulty of a broader skill.
- Panelists were invited to consider more broadly the spatial relationship between items and their cut scores on the item map. They were invited to think about "how far" on the item map their cut score lay with respect to an item and how related items were distributed on the map with regard to their cut score.

Rater Group Discussion: Sharing Perspectives Most of the time in Round 3 was spent on a “Rater Group Discussion.” Within each group, tables were pulled together and panelists took turns sharing the following: 1) how they chose their Round 1 bookmark placement, 2) how they choose their Round 2 cut scores, and 3) what information they were thinking of using to choose their Round 3 cut scores. The discussion lasted about 90 minutes, with each group discussion being attended to by a facilitator. Facilitators kept the discussion on track, focused on the

Achievement Level Descriptions, and encouraged all panelists to participate. The discussion began with the Proficient level, then moved to Basic, and finished with Advanced.

For the rater group discussion, panelists had available all of the key materials they had used to recommend cut scores in Rounds 1 and 2. These included the Achievement Level Descriptions, Ordered Item Books, Primary Item Map, Domain Item Maps, Domain Descriptions, Domain Score Chart, and Percent Correct Table (based on Round 2 median cut score).

Round 3 Cut Score Recommendations For recommending Round 3 cut scores, panelists were instructed to work independently, study the feedback from Round 2, reflect on the discussion, choose a scale value for a cut score, and record the cut score on the form provided. In considering cut scores, panelists were instructed to look at items in the OIB with scale values less than or equal to the cut score they were considering and think about whether a borderline student should have mastery of those items. They were also instructed to locate the scale value/cut score on their domain score chart and think about whether the domain scores associated with the cut score indicated acceptable borderline performance. They were also asked to consider which domain scores should be 67% or higher for the borderline student.

Panelists recorded their cut score recommendation on their Domain Score Chart, Ordered Item Booklet, Primary Item Map, and on the Cut Score Recommendation Form. For recording their cut score recommendation in the Ordered Item Book, they were given a chart that showed the OIB page number of the last item whose scale value was less than or equal to their recommended cut score.

Round 4

Feedback Feedback after Round 3 was presented using the same materials and formats that were used to present feedback after Round 2. Panelists were given a new Primary Item Map, Domain Score Chart, and Percent Correct Table. 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.

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 included. The consequences data were based on the Round 3 median cut scores. The feedback was presented in the form of a bar graph and pie chart. Panelists were also instructed to write the percentages of students in each achievement level and below basic in the left margin of their Primary Item Map.

The consequences data were discussed prior to panelists' making their Round 4 cut score recommendations. As a lead-in to the discussion, panelists were told that student performance is estimated from tests like the ones they took, which were given under similar conditions. Panelists were told that the sample was nationally representative, that student performance was influenced by student motivation and by the amount of time available. But regardless of what students can do, it’s what students should be able to do, according to the Achievement Level Descriptions that “rules the day.” The discussion was largely left open to panelists, but a number of questions were suggested for discussion: Were they surprised by the percentages? Were their expectations influenced by their own experience? What allowance did they feel should be made for motivation or for timed conditions of the test? What justification was there for considering student performance data when setting criterion-referenced standards?

Round 4 Cut Score Recommendations The purpose of Round 4 cut score recommendations was to allow panelists to adjust their cut score recommendations based on feedback after Round 3, including the consequences data. 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 recommendation. Panelists recorded their cut score recommendations as they did in Round 3.

Post-Rounds Activities

Feedback Feedback after Round 4 was given in the usual fashion except that panelist's individually-recommended cut scores were not indicated in the feedback materials. Panelists had already marked the location of their Round 4 cut score recommendation in materials that they had from Round 3, and the new materials would not be used for another round of cut score recommendations. A new Primary Item Map, Domain Score Chart, and Percent Correct Table were distributed. The feedback included consequences data based on the Round 4 medians. Panelists recorded the percent in each achievement level, and the percent below basic, in the margins of their item maps.

Panelists were told that the Round 4 medians would be reported to NAGB as one of the key outcomes of the ALS meeting. It was very important that panelists understood what students at the cut scores “can do,” which is the purpose of the feedback, and that they should evaluate the cut scores based on the match between the criterion-referenced feedback, the Achievement Level Descriptions, and their concept of borderline performance.

Consequences Questionnaire A consequences questionnaire was given to panelists to assess their reactions to the cut scores after viewing the consequences data. For each level, panelists could endorse the Round 4 cut score or recommend a different cut score.

Exemplar Item Ratings The use of exemplar items are specific to NAEP. Activities related to the selection of exemplar items are not essential to the Mapmark method and are therefore not described here.

General Issues and Procedures in Mapmark

In designing the Mapmark method, positions were taken on certain issues in standard setting. Some issues have been broached in the introduction of this paper. The following text identifies some remaining issues and explains how the Mapmark process was designed with regard to these issues.

The RP criterion This is regarded as a critical issue because different choices of the RP criterion can lead to different cut scores. A panelists' recommended cut score in Round 1 is the scale value of the item that receives the bookmark. If panelists do not take the RP criterion into account when placing their bookmarks, higher RP criteria will produce higher cut scores, other things being equal. Considerations regarding choice of the RP criterion and empirical results concerning this issue are presented in another paper in this session (Williams & Schulz, this session).

Procedurally, the Mapmark process dealt with the RP criterion issue in two ways. First, panelists were made fully aware of the RP criterion through orientation and training. This is expected to help panelists take the RP criterion into account when placing their bookmarks. Second, panelists use domain scores to select scale values for their cut score recommendations in subsequent rounds.   Since domain score feedback is independent of the RP criterion, its use should mitigate the effect of the RP criterion initially on bookmark placements in Round 1. Evidence consistent with this expectation from Mapmark field trials using different RP criteria is presented in the Williams and Schulz paper (this session).

Developing a Concept of Borderline Performance In Angoff-based standard setting procedures, it is recommended that panelists develop a consensus about what lower borderline students should be able to do before they begin rating items in Round 1 (Loomis & Hanick, 2000). A clear concept of lower borderline performance seems advisable because panelists must project this concept onto each and every item. A considerable amount of time can be spent in the process of developing consensus on a detailed, borderline description.

In Mapmark, panelists independently develop and use their concept of what students at the lower borderline of an achievement level should be able to do in the process of placing their Round 1 bookmarks. It is possible for panelists to develop their concept of borderline in the *process* of placing their bookmarks because the OIB, along with the extensive KSA review they performed earlier, provides them with a hierarchy of KSAs that they can apply to the achievement level descriptions and to the general concept of lower borderline performance that they are given—performance that "just qualifies" a student to be in the achievement level. The concept of borderline performance is subsequently discussed and developed further over successive rounds with reference to bookmarks and domain scores associated with the median (across all panelists) cut score, and panelists' individual recommended cut scores.

Independence Among Panelists For NAGB, the Mapmark process was implemented in a way that encouraged panelists to learn from the perspective and experience of other panelists, but also to maintain their own perspective and independent judgment. This approach is consistent with a theory of decision making described in the book, *The Wisdom of Crowds* (Surowiecki, 2004). The following points, in terms applicable to standard setting, are derived from *The Wisdom of Crowds* theory and were emphasized to Mapmark panelists in the course of their orientation and training:

- No single panelist can have all of the experience and perspective needed to set cut scores.
- No panelist can absorb, much less perfectly weigh all of the information presented to panelists for their cut score judgments.
- Rather, the *group*, which is all of the panelists taken together, has all the experience and perspective needed to set cut scores.
- All of the information relevant to setting cut scores will be weighed appropriately if panelists represent their own background and experience faithfully and exercise independent judgment in their cut score recommendations.

Mapmark panelists were also told that, in order for the collective wisdom of the group to manifest itself in the process…

- panelists are expected to share their perspective, but should not pressure others to make the same judgments or select the same cut scores, and
- panelists are expected to learn from the perspectives and experiences of other panelists, but also to faithfully represent their own perspective and experience. They should not subordinate their judgment to another panelist. Specifically,
- panelists should not allow themselves to be affected by the actual bookmark placements or cut score recommendations of other panelists.

Questions were placed on the process evaluation questionnaires to reinforce this perspective and to evaluate whether it was accepted by panelists.

Considerations of cut score reliability also favor an emphasis on independent judgment. The expected value of the mean of panelists' independent judgments is the same across different samples of panelists, other things being equal. Lack of independence means that the expected value of the mean is not the same across groups or occasions, but rather, tends towards the value of the most influential panelist within the group or occasion.

Criterion-Referenced versus Norm-Referenced Feedback In keeping with the value placed on independent judgment and with the criterion-referenced nature of performance standards, the feedback given to panelists after each round maximizes criterion-referenced meaning and minimizes "norm-referenced" meaning of panelists' individual cut scores.

Information that allows a panelist to see or estimate the number of panelists who recommended a cut score more or less extreme than they did, such as might be gleaned from a histogram of the distribution of cut scores across panelists is strictly norm-referenced information and is not provided in Mapmark. Only information about the median, highest, and lowest cut scores from the previous round is provided, and the information provided about these cut scores is criterion-referenced through the Domain Score Chart and the OIB. The criterion-referenced meaning of the median is used as a common point of reference for general discussions of what borderline students should be able to do. The criterion-referenced meaning of the high and low cut scores shows panelists the range of performance that was considered appropriate for borderline performance in the previous round. This range provides focus for their current round of criterion-referenced judgment.

PROCESS EVALUATIONS

The validity of standard setting outcomes depends on what is called "procedural validity." Evidence of procedural validity was gathered through six process evaluation questionnaires administered to panelists over the course of the meeting. The responses summarized in this section were collected on Likert scales. Some questions date back to the standard setting process that ACT used in 1992 to set achievement levels for the NAEP mathematics assessment. Others were added to address specific issues in the Mapmark procedure. On the five-point Likert scales used (1 to 5), averages above 3.5 have 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.

Only results bearing most directly on the Mapmark process itself will be presented in this paper. Results having to do with more general issues such as the organization and clarity of presentations, the skills of the facilitators, the quality of orientation materials, and so forth were generally good, and no doubt explain to some degree results that are more specific to the procedural validity of Mapmark. For example, if instructions in Mapmark procedures had been disorganized or lacking in clarity, panelists ratings of understanding of related concepts would be low. The reader may therefore assume that ratings on non-specific factors were commensurate with the results described here in terms of the overall quality of the process.

Many tables in this section contain a column that shows the questionnaire number (1 to 6) and sequence number for locating the question. This information will not be useful to readers of this report.

General Evaluation

The Mapmark ALS process compared well with methods ACT used in past standard setting work for NAGB. Table 2 shows the mean ratings of Mapmark and previous ALS methods on the key process evaluation questions. Both of the previous ALS methods represented in this table were modified-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 the Mapmark method generally compared well with the averages for the other two methods. It should be noted that on the scale for amount of time allocated for tasks, 3 was an optimum, 1 indicated too little time and 5 indicated too much.

Table 2.
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

In addition, most panelists said they would be willing to sign a statement recommending the use of the achievement levels resulting from the standard setting procedure. Possible responses to this question were "definitely" (coded 4), "probably" (coded 3), "probably not" (coded 2) and "definitely not" (coded 1). Of the 29 panelists who completed the last process evaluation questionnaire, nineteen responded "definitely", 9 responded "probably", and only one responded "probably not". This rate of endorsement (97% favorable) compares well with previous standard setting processes that ACT has conducted for NAGB.

Understanding of Concepts, Tasks, Feedback

Panelists' understanding of concepts and tasks in Mapmark was generally good. In Table 3, it can be seen that panelists understood the concepts associated with using their item maps, OIB and domain scores. In Table 3, it can be seen that Panelists understood how to choose their bookmarks in Round 1 and how to choose scale values for their cut score recommendations in subsequent rounds.

Table 3
Understanding of Concepts

I understand/understood …
(5=Totally Agree; 3=Somewhat Agree; 1=Totally Disagree)

Round	Question Location	Activity	Average Rating
Pre	1-7	the purpose of the NAEP achievement level setting meeting	4.35
Pre	1-10	the difference between criterion-referenced and norm-referenced standards	4.63
1	2-3	the score levels of polytomous items	4.10
1	2-6	how to use my item map and ordered item booklet	4.42
2	3-7	the concept of domain scores	4.30
2	3-10	how to use the domain item maps	4.19
2	3-11	how to use the domain ordered item booklet	4.52
2	3-12	how to use the domain score chart	4.39
Post	6-22	the purpose of this meeting	4.80

Table 4
Understanding of Tasks

My understanding/level of understanding of…
(5=Totally Adequate; 3=Somewhat Adequate; 1=Totally Inadequate)

Round	Question Location	Activity	Average Rating
1	1-24	our tasks in the KSA review	4.03
1	2-30	how to use the ALDs to choose my bookmarks	4.13
2	3-23	how to choose cut scores for Round 2	4.30
3	4-19	how I was to choose cut scores for Round 3	4.42
4	5-18	how I was to choose cut scores for Round 4	4.53

Panelists' had good understanding of the feedback they were given. As shown in Table 4, average ratings of understanding of general types of feedback such as the numerical values of the cut score (Round ___ median cut scores), rater location feedback, and domain score feedback were well above 4.0 after Round 1 and continued to increase with each round in most cases. Understanding the difference between borderline performance and typical performance was not a form of feedback, but was essential for understanding the feedback because feedback pertained to borderline performance.

Table 5
Understanding of Feedback

I understand/understood …
(5=Totally Agree; 3=Somewhat Agree; 1=Totally Disagree)

	Round			
Information/Concept	1	2	3	4
The Round __ median cut scores	4.58	4.68	4.70	4.73
What students at the Round __ median cut scores can do	4.45	4.45	4.57	4.67
The Rater location feedback	4.68	4.68	4.72	---
The domain score feedback	4.55	4.52	4.67	4.70
The difference between borderline performance and typical performance	4.52	4.58	4.47	---
The consequences data	---	---	4.70	4.50

Developing a Concept of Borderline Performance

As shown in Table 5, panelists were comfortable using the concept of borderline performance to place their bookmarks in Round 1. By Round 2, their concepts of borderline performance were well-formed and continued to become better formed over subsequent rounds.

The pattern of responses in Table 6 is similar to patterns seen in previous standard setting work for NAGB, where the question about how "well formed" panelists' concept of borderline performance was at the time of item ratings was asked in every round. Round 1 averages were near 3.5 and averages for subsequent Rounds were above 4.0.

Table 6
Development of Borderline Concept

I was comfortable using the concept of performance at the lower borderline of _____
(5=Very Well Formed; 3=Moderately Formed; 1=Not Well Formed)

	Round			
Level	1	2	3	4
Basic	3.87	---	---	---
Proficient	3.81	---	---	---
Advanced	3.84	---	---	---

At the time I provided the/my Round __ bookmark placements/cut score recommendations my concept of the lower borderline performance at the ___ level was…
(5=Very Well Formed; 3=Moderately Formed; 1=Not Well Formed)

	Round			
Level	1	2	3	4
Basic	---	4.35	4.37	4.59
Proficient	---	4.39	4.39	4.53
Advanced	---	4.29	4.47	4.50

In addition to the data in Tables 5 and 6, the responses of panelists to the question concerning the difference between borderline performance and typical performance, summarized by Round in Table 5, should be noted. We attribute the clear understanding indicated by averages near 4.5 in part to the illustration of achievement level boundaries by vertical lines on domain score plots such as in Figure 10. Illustrations of how performance changes over the range of an achievement level focuses panelist's attention on the concept of borderline performance.

Table 7 shows that the perceived consistency between the ALDs and panelists' cut score recommendations increased over rounds. This is what one would expect from the patterns of understanding and concept formation evident in previous tables of this section.

Table 7
Consistency of Cut Score Recommendations with ALDs

I believe my Round ___ bookmark placements/cut score
Recommendations are consistent with the ALDs
(5=Totally Agree; 3=Somewhat Agree; 1=Totally Disagree)

Round	Question Location	Mean
1	2-27	3.94
2	3-20	4.13
3	4-17	4.48
4	5-16	4.63

Comfort and Confidence

As shown in Table 8, panelists were comfortable with key features of the Mapmark process including the value of the response probability criterion (0.67) and its meaning (mastery).   In Round 2 (Questionnaire #3), panelists had acceptable levels of confidence in deciding whether domain scores should be higher or lower at the borderline (3.84) and in choosing a scale value rather than a bookmark placement to recommend a cut score (3.90). These are good average ratings considering that Panelists invested relatively more time in item-level tasks and judgments in Round 1, and were performing their domain-level judgments for the first time in Round 2. Panelists' confidence in their cut score recommendations increased steadily from Round 1 (3.28) to Round 4 (4.43). These levels of confidence, and the trend of increasing confidence over rounds, are typical of other methods and achievement level setting meetings ACT has conducted for NAGB. Confidence in Round 1 judgments is typically lower than 3.5 because panelists have not received any feedback about their judgments.

Table 8
Comfort and Confidence

I think I will be/I was comfortable …
(5=Totally agree; 3=Somewhat Agree; 1=Totally Disagree)

Round	Question Location	Activity	Average Rating
1	1-17	Using a 2/3 or 0.67 probability to interpret the location of an item on my map	4.23
1	2-7	Working through the ordered item booklet on my own	4.39
1	2-33	Using a 0.67 probability to define mastery in placing my bookmarks	4.00
2	3-8	Thinking about whether an item was like other items in its domain (Domain Task 1)	4.39
2	3-26	Choosing scale values instead of placing bookmarks to recommend cut scores	3.90

The most accurate description of my level of confidence in …
(5=Totally Confident; 3=Somewhat Confident; 1=Not at All Confident)

Round	Question Location	Activity	Average Rating
2	3-9	deciding whether domain scores should be higher or lower	3.84
4	5-8	using the consequences data to recommend cut scores	4.30

Usefulness/Helpfulness of Materials and Information

Results in Table 9 show that panelists found the KSA activities generally to be useful. The three KSA activities asked about in this regard involved some level of group work, as opposed to KSA Activity 3, which was the independent OIB review. The bottom panel of Table 8 shows that the information and materials in the Mapmark process were generally perceived to be helpful. Average ratings for all materials and information specific to the Mapmark process were above 4.0 and were higher than the average rating for the helpfulness of consequences data (the percent of students in achievement levels), at 4.07. This may be regarded as a positive outcome since the consequences data are purely normative information. Average ratings of helpfulness of item maps and domain score feedback were good. The OIB was perceived to be most useful, with an average rating of 4.76.

Table 9
Usefulness/Helpfulness of Activities/Information

The ____ was
(5=Very Useful; 3=Somewhat Useful; 1=Not at All Useful)

Question Location	Activity	Average Rating
1-25	Whole group work on common constructed response items (KSA Activity 1)	4.23
2-2	Table group review of the remaining constructed response items (KSA Activity 2)	4.37
2-12	Table discussion of the ordered item booklet (KSA Activity 4)	4.37

During the ALS process, I found the __________
(5=Very Helpful; 3 = Somewhat Helpful; 1 = Not at all Helpful)

Question Location	Information/materials	Average Rating
6-31	The achievement level descriptions	4.38
6-32	The ordered item booklet	4.76
6-33	The primary item map	4.24
6-34	The domain-ordered item maps	4.24
6-35	The rater location data	4.46
6-36	The domain score feedback	4.21
6-37	The consequences data	4.07

The relatively high average rating for helpfulness of the rater location data, 4.46, suggests that panelists did not need to know more about the location of their cut scores relative to that of other panelists other than knowing the median, highest, and lowest cut scores from the previous round, as well as their own cut scores.

Independence of Judgment and Perspective

Process evaluation results indicated that the general instructions panelists were given with regard to maintaining their perspective and independent judgment were effective. As shown in Table 10, 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").

Table 10
Perceived Influences/Pressure on Cut Score Recommendations

I felt pressure to recommend bookmarks/cut scores that were close to those recommended by other panelists
(5=Totally Agree; 3 = Somewhat Agree; 1 = Totally Disagree)

Round	Question Location	Mean
1	2-32	1.37
2	3-25	1.71
3	4-21	1.43
4	5-20	1.63

Domain Coherence

Table 11 shows results from Domain Task 1. In this task, panelists indicated 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 is 93%. By panelist type the percentage is 96% for teachers, 91% for non-teacher educators, and 89% for general public representatives. By individual panelist, the percentage ranges from 62% (a general public representative) to 100% (for two teachers). These percentages indicate that the domains were generally coherent and that the task was a reasonable task for panelists to perform. One would expect the percentage of 'yes' responses to be higher among teachers than non-teachers, and lowest for the general public representatives since these types have the most and least experience related to the task, such as thinking about what mathematics skills may be involved in solving a test item.

Table 11
Percentage of "Yes" Responses to Domain Task 1 by Panelist and Panelist Type

"I see how this item is like other items in its domain"
106 and 109 Items for Groups A and B Respectively

		Panelist	Panelist	Percentage
Group	Table	ID	Type	"Yes"
A	1	A1201	GP	76%
		A1202	NT	84%
		A1203	TR	96%
		A1204	TR	94%
		A1205	TR	89%
	2	A1206	GP	92%
		A1207	NT	97%
		A1208	TR	98%
		A1209	TR	97%
		A1210	TR	94%
	3	A1211	GP	98%
		A1212	GP	97%
		A1213	TR	98%
		A1214	TR	95%
		A1215	TR	96%
B	4	B1216	GP	90%
		B1217	NT	94%
		B1218	TR	96%
		B1219	TR	98%
		B1220	TR	100%
	5	B1221	GP	99%
		B1222	GP	62%
		B1223	NT	96%
		B1224	TR	96%
		B1225	TR	93%
	6	B1226	GP	96%
		B1227	GP	92%
		B1228	NT	81%
		B1229	TR	96%
		B1230	TR	100%
		B1231	TR	96%
			Average:	93%

T (Teachers):	96%
NT (Nonteacher Educators):	91%
GP (General Public):	89%

Relationship Between Domain Task 2 and Subsequent Change in Cut Scores

The relative frequency of checks in the higher/OK/lower categories of Domain Task 2 was related to the difference between Round 1 cut scores and cut scores from later rounds. This relationship is shown in Table 12. The percentage of checks by category was averaged across domains and panelists. At all three levels, the majority of checks were in the "OK" category and the difference between the percentage of checks in the lower versus higher categories was small (9 points or less). It therefore seems reasonable that cut scores did not change very much from Round 1.

At the Advanced level, where there was no change in the cut score over rounds, the percentage of checks in the "OK" category was largest (70%) and the difference between the percentage of checks in the highest versus the lowest category was smallest (3 points).

At the Basic and Proficient levels, where Round 2 through Round 4 cut scores were higher than the Round 1 cut scores, there were more checks in the "higher" than in the lower category (25% versus 19% for Basic; 27% versus 18% for Proficient).

Table 12
Relationship Between Domain Task 2 and Subsequent Movement in Cut Scores

Achievement Level	Domain Task 2 Categories	Round 1 Cut	Percentage of Checks in Category	Round 2 Cut	Round 3 Cut	Round 4 Cut
BASIC	Higher OK Lower	x	25 56 19	x+1	x+2	x+2
PROFICIENT	Higher OK Lower	y	27 54 18	y+2	y+1	y+1
ADVANCED	Higher OK Lower	z	12 70 15	z	z	z

Reactions to Consequences Data

In the Round 3 whole group discussion of consequences data—the percent of students at or above each of the achievement levels—panelists generally voiced surprise and disappointment that the percentages were not higher, but did not feel that the cut scores should be lowered. It can be seen from Table 12 that the median cut score did not change from Round 3 to Round 4. This result, along with comments voiced during the whole group discussion, indicates that panelists were strongly committed to the criterion-referenced meaning of their cut score recommendations.

As shown in Table 13, a large majority of panelists endorsed the Round 4 cut scores after viewing the consequences data once again. Of those who chose to recommend a different cut score, the majority recommended lower cut scores, as one would expect if some panelists had higher expectations of students than were borne out by the data. The number of panelists recommending lower cut scores increased with the achievement level. At Basic, equal numbers recommended higher versus lower cut scores. At Advanced, seven out of eight recommended a lower cut score.

Table 13
Cut Score Endorsements/Recommendations after Seeing Round 4 Consequences Data

Achievement Level	Lower	Number Endorsing Round 4 Cut Score	Higher
Basic	4	23	4
Proficient	5	23	3
Advanced	7	23	1

DISCUSSION AND CONCLUSIONS

The Mapmark method makes full use of item response theory and the latest developments in domain score theory and technology. Item response theory is used to order items in the Ordered Item Book and to arrange items on item maps by a response probability criterion. Items are organized into columns on the item maps corresponding to areas of knowledge, skills, and abilities called domains. Item response theory was used to estimate domain scores conditional on student achievement scale values in the Mapmark process.

The Mapmark method is not necessarily "easier" or less complex than other methods, but the Mapmark tasks build understanding that is essential for setting performance standards. Panelists initially invest many hours understanding the progression of student achievement in the OIB and on item maps. Then they invest more time understanding growth in student achievement as an increase in percentage correct scores on domains. These tasks are complex, but the process evaluation results indicate that they are meaningful and not too difficult. They help Mapmark panelists understand how student achievement increases as a sequential mastery of knowledge, skills, and abilities. This understanding is essential for setting performance standards.

The Mapmark component of the standard setting process conducted for NAGB contributed positively to the overall procedural validity of the process. Results from the process evaluation questionnaires show that panelists understood the concepts and tasks specific to the Mapmark method, were confident in their cut score recommendations and believed that the process was effective and allowed them to use their best judgment.

A high percentage of the panelists said they would definitely or probably sign a statement endorsing the cut scores resulting from the process. A high percentage also endorsed the Round

4 cut scores after viewing the consequences data. These results suggests that the cut scores and the achievement level percentages associated with them may be more generally perceived as reasonable.

Results from the ALS meeting also added to results from previous studies ACT conducted in this project which showed that domain scores are a reasonable and useful addition to a standard setting process. Panelists understood the domain score information they were given, were able to evaluate it relative to the achievement level descriptions, and to use it to recommend cut scores. The scale values they recommended in subsequent Rounds were logically related to their evaluation of the domains scores (higher/OK/lower) associated with the Round 1 cut scores. The usefulness of the domains may be related to the domains' coherence, as indicated by the high percentage of "yes" responses in Domain Task 1 ("I see how this item fits with other items in its domain;" yes/no/not sure).

Although cut scores did not change much over rounds in the ALS meeting, the overall pattern of change in this and other studies conducted in the project, suggests that domain score feedback does influence cut score recommendations. In field trials and in a Grade 8 study, cut scores changed upwards or downwards from Round 1, depending on the RP criterion. When a 0.67 RP criterion was used (Field Trial 1 and Grade 8 study) cut scores dropped by 3 to 10 points. When a 0.5 RP criterion was used, cut scores increased by 5 to 13 points. These results suggest that domain scores have a moderating influence on the effect of the RP criterion, as expected. The changes were not great enough to produce the same or even comparable cut scores across studies, but seemed large enough to mitigate differences of approximately 0.1 or less in the RP criterion. In a separate study of the Mapmark procedure using Grade 12 data and a 0.67 RP criterion (Pilot study) Round 1 cut scores started out lower for Basic and Proficient, but then increased over rounds by a few points and ended up very close to where the Round 1 cut scores were in the ALS meeting. It therefore seems reasonable to suppose that cut scores did not change very much across rounds in the ALS meeting because the domain score feedback associated with Round 1 cut scores was truly satisfactory to panelists.

Questions for the future are 1) whether clearly defined sequences of related domains covering a wide range of difficulty can be developed in other subject areas, and 2) if not, whether domains will be as useful in standard setting or for explaining achievement levels to the general public. Mathematics is generally regarded as the most hierarchical of subjects. It may be more difficult to define domains with similar characteristics in content areas such as Reading. If domains tend to be similar in difficulty, they may be less useful for defining achievement levels. One would not be able to describe each achievement level in terms of at least one domain that has been mastered and at least one domain that has not been mastered, with regard to a fixed percent correct criterion for mastery. However, percent correct score feedback may still be useful to panelists if they understand the domains well and are able to project their concept of the borderline of each achievement level into a percentage correct score on the domain. The organization of items into columns representing similar areas of content could still serve the purpose of alerting panelists to the unreliability of inferences based on single items and percentage correct scores on the domains could still provide a more reliable basis for inference.

In the long run, it would be most advantageous to incorporate the goals that guided domain development in this project into the framework development process. The domain development component of the project was focused on producing domains with specific characteristics and goals in mind. Items could be written to represent content areas or skills that have an expected order of difficulty based on instructional timing or theories of learning. The incorporation of such content areas and skills, along with expectations of difficulty-order, into the test plan and item development process would serve many purposes well. One goal of domain development in the present project was to be able to provide reliable, criterion-referenced descriptions of what growth in student achievement means, and what NAEP achievement levels mean, to educators, policy makers and the general public alike.

REFERENCES

Council of Chief State School Officers (2001). *State Student Assessment Programs Annual Survey*. Data Volume II. Washington D.C. Author.

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.

Engelhard, G. Jr., & 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*. Ablex Publishing Corporation. Stamford, CT.

Forsyth, R. A. (1991). Do NAEP scales yield valid criterion-referenced interpretations? *Educational Measurement: Issues and Practice*, *10*(3), 3-9.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics*, *23*, 35-56.

Kolstad, A., Cohen, J., Baldi, S., Chan, T., DeFur, E., & Angeles, J. (May, 1998). *The response probability convention used in reporting data from IRT assessment scales: Should NCES adopt a standard?* Report prepared for the National Center for Education Statistics, Washington, DC.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). Standard Setting: A bookmark approach. In D. R. Green (Chair), *IRT-based standard-setting procedures using behaviorial anchoring.* Symposium conducted at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Loomis, S. C. & Hanick, P.L. (2000). *Developing achievement levels for the 1998 NAEP in civics: Final report*. Iowa City, IA: ACT.

Masters, G. N., Adams, R. & Lokan, J. (1994). Mapping student achievement. *International Journal of Educational Research*, *21,* 595-609.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: psychological perspectives. In G. J. Cizek (Ed.), *Setting Performance Standards*. Mahwah, NJ: Lawrence Erlbaum Associates.

Schulz, E. M., Lee, W., & 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.

Stone, G. E. (2001). Objective standard setting (or Truth in Advertising). *Journal of Applied Measurement*, *2*(2), 187-201.

Surowiecki, J. (2004). *The Wisdom of Crowds.* New York, NY: Doubleday.

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., & Schulz, E. M. (April, 2005). *An Investigation of Response Probability (RP) Values used in Standard Setting*. Poaper presented at the Annual Meeting of the National Council on Measurement in Education. Montreal, Canada.

Wright, B. D., and Masters, G. N. (1982). *Rating Scale Analysis*. Chicago: MESA Press.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (December, 2000). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. Iowa City, IA: ACT, Inc.