Developing Achievement Levels on the 2005 National Assessment of Educational Progress in Grade Twelve Mathematics

Process Report

Presented by ACT, Inc.
April 29, 2005

REDACTED BY GOVERNING BOARD

Developing Achievement Levels on the 2005 National Assessment of Educational Progress in Grade Twelve Mathematics

Process Report

The work for this report was conducted by ACT, Inc. under contract ED-03-CO-0099 with the National Assessment Governing Board.

Process Report
Table of Contents

APPENDICES

EXECUTIVE SUMMARY

OVERVIEW

This report describes the process and outcomes of a meeting that was held in November 2004 to set achievement levels for the 2005 National Assessment of Educational Progress (NAEP) in Grade 12 Mathematics. The meeting was conducted by ACT, Inc. under contract with the National Assessment Governing Board (NAGB). The contract calls for ACT to conduct achievement level setting (ALS) activities consistent with NAGB policies and to develop recommendations for setting achievement levels. The actual setting of achievement levels is a policy judgment by NAGB, based on contractor recommendations. ACT bases its recommendations for achievement levels on evidence that the ALS process had procedural validity and was reliable, and that the outcomes are likely to be viewed as reasonable. This report is a summary of such evidence.

In addition to describing the ALS meeting process, the report presents the recommended achievement level descriptions, recommended cut scores, and identifies items that may be used to illustrate what students in the achievement levels know and can do (exemplar items). Recommendations for future level setting are also included.

Key issues and conclusions from project activities that preceded the ALS meeting are also summarized in this report. NAGB standard setting contracts generally call for field trials, pilot studies, and other research activities designed to improve the standard setting process and the way standard setting results are reported (Reckase, 2000). For the ALS meeting in this project, ACT developed a new standard setting procedure, Mapmark, through a series of field trials and pilot studies. In a Pilot Study, Mapmark was compared to the Item Rating method that ACT used to set achievement levels for the 1998 NAEP in Civics.

Additional information about the project may be found in other sources. The Technical Report documents technical advice ACT received in the project, data analysis procedures, pilot activities and describes the materials used throughout the process. A report submitted to the National Center for Education Statistics (Schulz, 2003) and a paper by Schulz, Lee, and Mullen (2005) describe the general process of developing content domains like those used in the operational ALS meeting.

BACKGROUND

The National Assessment Governing Board (NAGB) has been setting achievement levels for grades and subject areas in the National Assessment of Educational Progress (NAEP) since 1992. Achievement levels have been set for the National Assessments in Reading, Writing, Mathematics, Science, History, Geography, and Civics. As currently specified by NAGB policy, there are two stages to the NAEP ALS process. In Stage 1, grade-specific and subject-specific achievement level descriptions (ALDs) are developed from general policy definitions for three achievement levels—Basic, Proficient, and Advanced.

The ALDs represent what students in the achievement levels should know and be able to do. In Stage 2, the ALDs are translated into cut scores. Stage 1 occurs before the ALS meeting. Stage 2 is the ALS meeting.

Achievement levels have become the most publicly visible aspect of the Nation's Report Card. Achievement level percentages—the percent of students in each achievement level and the percent at-or-above each achievement level show how students are performing relative to what students should know and be able to do. Trends in achievement level percentages have become a major resource to educators and policy makers assessing the nation's progress towards its educational goals.

The setting of achievement levels for the 2005 NAEP in Grade 12 Mathematics is unique in that it is concerned with developing new achievement levels that were set at an earlier point in time. The framework for the 2005 National Assessment in mathematics is significantly different at Grade 12 from the previous framework. NAGB policy is to update the achievement levels as needed, typically when assessment frameworks are updated. Because achievement scores on tests constructed from different frameworks are fundamentally not comparable, NAGB decided to report results for the 2005 NAEP in Grade 12 Mathematics on a new scale—one that does not support comparisons of achievement scores to previous assessments. Unfortunately, this does not prevent comparisons of the achievement level percentages from occurring.

For the current project, ACT proposed to develop a new standard setting method, Mapmark, and to compare Mapmark to the method ACT used to set achievement levels for the 1998 NAEP in Civics. The 1998 method is called "Item Rating" in this project and is based on a modified-Angoff method. Mapmark is based on the bookmark procedure (Mitzel, Lewis, Green, & Patz, 2001). Bookmark was introduced in 1996 (Lewis, et al., 1996), and has since become the most widely-used standard setting method in state assessments (CCSSO, 2001). ACT's proposal to NAGB recognized that the bookmark method contained some very attractive features for setting achievement levels, but predicted that it could be improved with the use of item maps (Masters, Adams, & Loken, 1994) and domain-score feedback (Schulz, Lee, & Mullen, 2005).

ACT conducted two field trials (Field Trials 1 and 2) and a Grade 8 Study for the purpose of developing the Mapmark method. Mapmark was then compared to the Item Rating method in a Pilot Study. ACT presented the results of these activities, and its recommendations, to the NAGB Committee on Standards, Design, and Methodology (COSDAM). COSDAM chose the Mapmark method for the operational ALS meeting. The following points were noted by COSDAM.

- The change in framework presents a natural opportunity to consider a new achievement level setting method because new achievement level percentages do not have to agree with previously reported achievement level percentages.

- Results from the methods (Item Rating and Mapmark) differed somewhat from each other and both sets had one or more notable difference from previous achievement level percentages.

- Both methods (Item Rating and Mapmark) have procedural validity and the achievement level percentages produced by either method are likely to be viewed as reasonable.

- Mapmark is more likely to be understood by educators because it has basic similarities to the bookmark procedure, which has become the most widely used standard setting method in state assessments.

- ACT gave assurances that Mapmark could be conducted in four days without compromising the integrity of the process. COSDAM felt that this flexibility was important for recruiting panelists and maximizing panelists' effort and satisfaction with the process.

Based on the COSDAM decision, ACT implemented the Mapmark method in the operational ALS meeting.

ALS MEETING

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 Panelists

Panelists were selected using the same basic design used in the 1998 NAEP Civics ALS meeting. The design included stratified random sampling with school districts as the basic sampling unit. Three random samples of school districts were drawn for each of three panelist types: Teacher, non-teacher educator, and general public. The sample of school districts for non-teacher educators was supplemented with post-secondary institutions, state departments of education, and other specific institutions or positions. School districts were contacted for nominators—persons who could nominate qualified panelists of the given type. A total of 1,385 potential nominators were contacted and 167 persons were nominated. Thirty-one panelists participated in the ALS meeting. The percentages 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 represented 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.

The ALS Meeting Process

As proposed by ACT and eventually implemented in the operational Grade 12 Achievement Level Setting meeting, the Mapmark method used a bookmark procedure (Mitzel, Lewis, Patz & Green, 2001) in Round 1, and provided domain score feedback in Round 2 and subsequent rounds. Domains are areas of content more general than a single item, but more specific than the test as a whole. The domains used in the ALS meeting were defined during the course of the project using methods described by Schulz, Lee, and Mullen (2005).

In a bookmark method, panelists recommend cut scores by placing bookmarks that divide items in an ordered item booklet (OIB) into two groups—those that they feel a borderline student should have mastery of and those that are too difficult for this expectation. Mastery is defined as having a certain probability of answering the items correctly. A response probability (RP) of 0.67 was used in the Mapmark procedure for the ALS meeting. Figure 1 illustrates a bookmark placement.

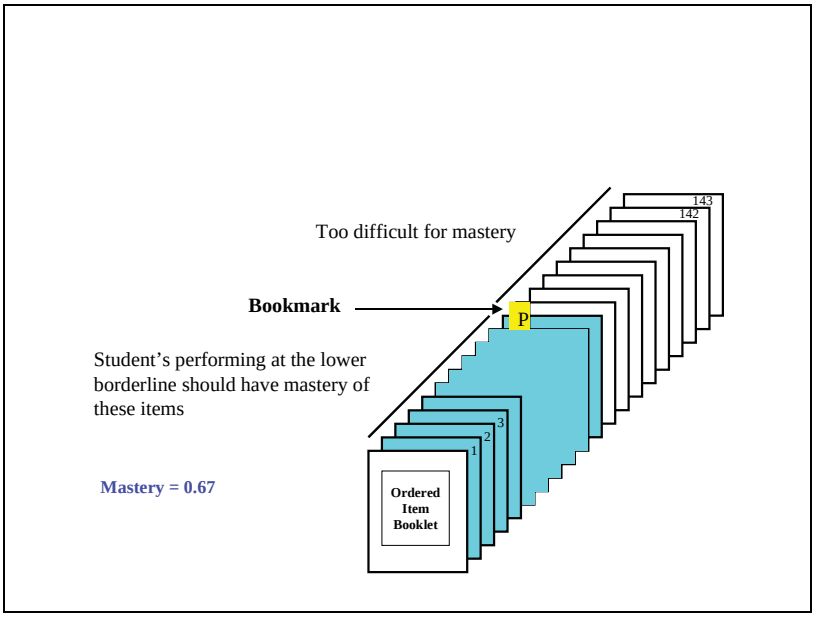
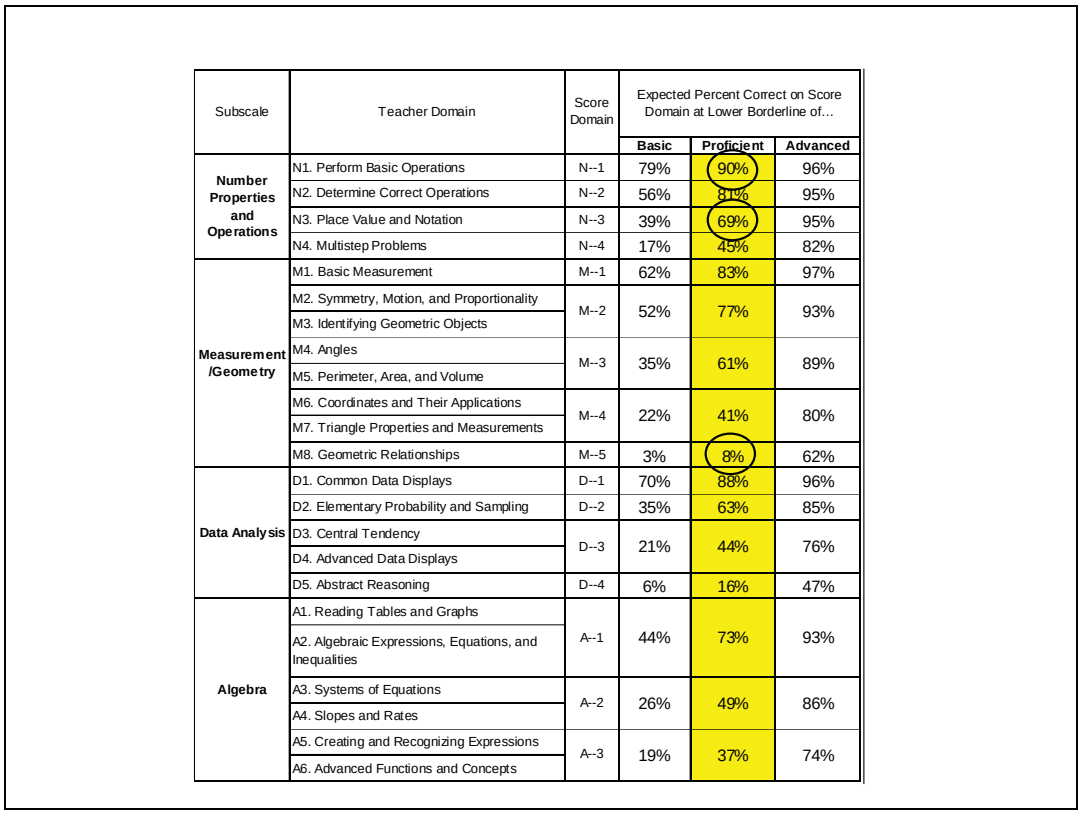


Figure 1. Illustration of a bookmark placement.

In Round 2 and subsequent rounds of Mapmark, panelists recommended cut scores directly by selecting scale values. The selection of scale values was facilitated by domain score feedback. A percent correct table showed the domain scores expected of students at the cut score. Figure 2 shows a slide that was used to introduce the percent correct table to panelists. The table highlights expected domain percent correct scores at the

lower borderline of Proficient. High and low domain scores are circled to call attention to the fact that panelists will be asked to judge whether these scores are too low, OK, or too high for the borderline of each achievement level.



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 2. Slide presenting the Percent Correct Table to panelists.

The "Teacher domains," identified by their titles in Figure 2, were defined by ACT earlier in the project. The teacher domains had, or were intended to have, the following characteristics:

1) Well Defined. Each domain should be well represented by a *domain definition* consisting of a title, a brief narrative description, and up to three sample items (if available). The title and narrative should represent in relatively jargon-free language that can be understood by teachers, non-teacher educators, and the general public alike, the knowledge, skills, and abilities required by items in the domain. Sample items should be drawn from released items on the NAEP website.

2) Coherent. Teachers should be able to reliably and independently classify items into the domains by content, using only the domain definitions. Standard setting panelists should be able to see and understand how items fit, or belong, in their domain.

3) Variable in Difficulty. The domains within each subscale should differ in difficulty and collectively cover a wide range.

It can be seen that the teacher domains were occasionally combined into a smaller number of "score domains" represented by domain scores. Some teacher domains were large enough or distinct enough in difficulty to stand alone as score domains. Domain scores were based on items from the 2005 assessment that were classified into domains defined earlier in the project.

Two domain tasks (Domain Tasks 1 and 2) were designed to familiarize panelists with the domains and to help panelists decide what the domain scores at the lower borderline of an achievement level should be. The selection of a scale value for a cut score was facilitated by a domain score chart showing the domain scores associated with every scale value within a wide range centered on the cut score from the previous round.

Item maps (Masters, Adams, & Loken, 1994) were used in every round of Mapmark. Figure 3 illustrates the essential features of an item map. Test items were arranged vertically on a scale that represents both item difficulty and student achievement. Cut scores were represented on the same scale. Items were located on the map by the same RP criterion used to order the items in the OIB. In the example below, student "Y" has a 0.67 chance of answering Item 6 correctly and a greater than 0.67 chance of answering Item 5 correctly.

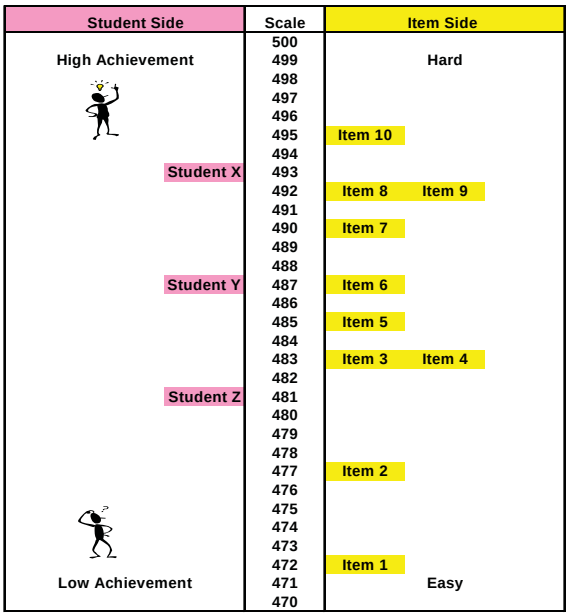


Figure 3. Item map showing spatial array of items by relative difficulty on a scale of student achievement.

In Mapmark, items were also organized into columns representing different content. There were two types of item maps. On the Primary Item Map, columns corresponded to subscales of the assessment. On Domain Item Maps, columns corresponded to teacher

domains. Figure 4 shows a simplified version of a Domain Item Map illustrating the location of a cut score. Students at the cut score are expected to have mastery of items (0.67 or higher probability of answering correctly) below the cut score.

Figure 4. Simplified Domain Item Map illustrating cut score location.

Evaluations of the ALS Meeting Process

Procedural validity of the ALS process was evaluated through process evaluation questionnaires given to panelists at the conclusion of each round and day. Many of the questions had been used in the Pilot Study and in previous ALS meetings. A detailed summary of responses is contained in the full report. However, the data in Table 1 are representative of the fact that the Mapmark process was well implemented. Average responses from the 1992 and the 1998 ALS processes are shown for comparison. The Mapmark process was viewed as positively as the previous ALS processes.

The average response to the *amount of time* question in Table 1 shows that panelists generally felt that the amount of time they had to perform their tasks was adequate. A response of '3' to this question is most favorable. A very large majority of responses to this question were '3' and the average was close to 3.0. Responses to this question and to many similar questions concerning the allocation of time for specific tasks helped confirm ACT's claim that Mapmark could be performed in four days, rather than five days as had been traditionally used for the Item Rating method, without compromising the integrity of the process.

Table 1: Summary 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… (5=Far too long; 3=About right; 1=Far too short)	Mapmark ALS	3.03
	1998 Civics	3.21
	1992 Math	3.12

On the other key summary process evaluation questions in Table 1, where 1 is least favorable and 5 is most favorable, the average response has historically been 4.0 or higher. This was the case with the Mapmark ALS process. On ratings of effectiveness, confidence, clarity of instructions, panelists' understanding of their tasks, and providing panelists the opportunity to use their best judgment, the Mapmark ALS process performed well in relation to previous ALS processes.

The ALS process was also evaluated on the basis of the following criteria:

- reasonable cut scores;
- reasonable variability of cut scores across panelists;
- absence of extreme reactions to consequences data (the percent of students at or above each achievement level);
- adequate number of exemplar items for each achievement level; and
- consistency of results with previous studies.

Evaluations of the ALS process on these criteria were positive. Details are provided in the full report and in other sections of this executive summary.

ALS PROCESS OUTCOMES

The ALS process consists of all activities leading up to the setting of achievement levels by NAGB. In setting the achievement levels, NAGB adopts three major outcomes of the

ALS process: achievement level descriptions (ALDs), cut scores, and exemplar items. Exemplar items are used to illustrate what students in each achievement level know and can do.

Achievement Level Descriptions

The development of achievement level descriptions in this project conformed to the two-stage process described above in that they were developed before the ALS meeting. In the past, the contractor (e.g., ACT) has developed the achievement level descriptions. In the current project, the ALDs were developed by NAGB. They were used in the Pilot Study as well as in the ALS meeting. The achievement level descriptions are contained in Appendix A of the full Process Report.

ACT endorsed the achievement level descriptions used in the ALS meeting. In the Pilot Study, as well as the ALS meeting, panelists reported that they understood the achievement level descriptions and found them useful for setting the cut scores. In a special task performed on the last day of the Pilot Study, panelists reported seeing items related to virtually every statement in the achievement level descriptions. Based on this and other results of the Pilot Study, ACT did not see the need to recommend any changes or modifications to the achievement level descriptions for the ALS meeting.

In the ALS meeting itself, the following process evaluation results were obtained concerning the ALDs:

- On a scale of 1 to 5, with 5 being very helpful, the average rating given to the achievement level descriptions for setting cut scores was 4.38.

- On a scale of 1 to 5, with 5 being totally agree, the average response to the statement, "I believe my bookmark placements/cut score recommendations are consistent with the ALDs" was 3.94, 4.13, 4.48, and 4.63, respectively, for Rounds 1 through 4.

In addition to these quantitative results, panelists' remarks to facilitators and observers concerning the ALDs indicated that they generally found the ALDs to be very good summaries considering their brevity and the complexity of the subject matter they represent.

Cut Scores

ACT recommended that NAGB adopt the median cut scores from Round 4 of the ALS meeting. The cut scores used to provide feedback to panelists after each round of the Mapmark process were the median cut scores across all panelists for each achievement level. This recommendation was based partly on the conclusion of ACT's TACSS that the ALS process had procedural validity and produced reliable results. It is also based on the conclusion that the achievement level descriptions associated with the cut scores are likely to be considered reasonable. Also, Round 4 cut scores are based on all of the information that ACT recommends be considered by panelists in adopting cut scores, including student performance data.

Figure 5 shows Round 4 achievement level percentages from the Pilot Study and the ALS meeting. Those from the ALS meeting are intermediate in every respect with the two sets of Pilot Study results. For example, the ALS percent below Basic, 37.4%, is intermediate with the Pilot Item Rating percent (38.5%) and the Pilot Mapmark percent (35.3%). Since COSDAM noted that both sets of Pilot Study results are likely to be considered reasonable, those from the ALS meeting, being intermediate with both sets, are also likely to be considered reasonable. It should be noted that the achievement level percentages in Figure 5 are based on 2004 field test results, and may differ from those reported in the 2005 assessment.

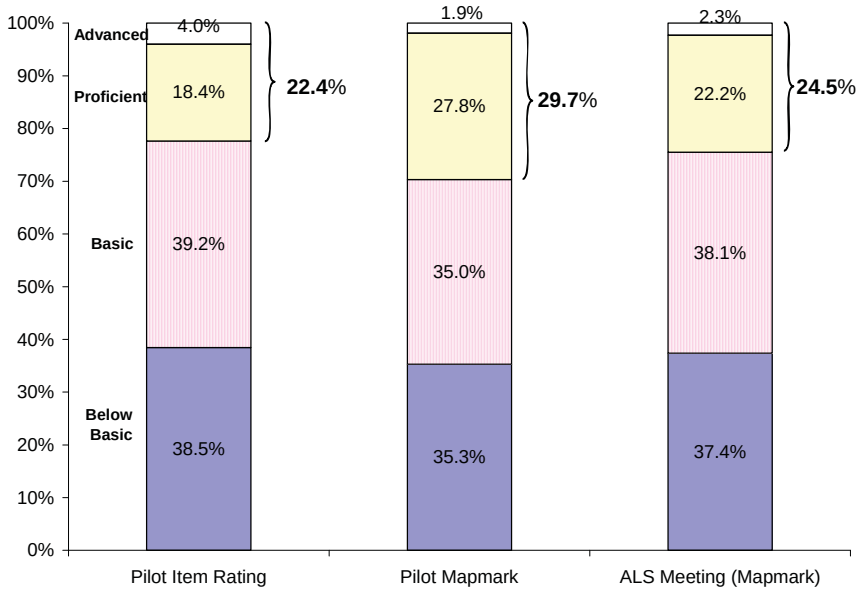


Figure 5. Round 4 achievement level percentages from Pilot Study and ALS meeting.

The numerical values of the cut scores are not meaningful by themselves and, in order to maintain confidentiality, correspond only indirectly to the achievement scale that will be used to report results of the 2005 assessment. The cut scores acquire meaning through information such as student performance data reported in Figure 5, and their reliability. The reliability of the cut scores was established through statistical analyses of their variability across panelists within and across the Pilot Study and ALS meetings.

Exemplar Items

Following Round 4 of the ALS meeting, panelists provided input on the suitability of selected items for illustrating what students in the achievement levels, as defined by Round 4 cut scores, know and can do. The statistical criteria ACT used to associate items with achievement levels for the rating task used the response probability (RP) criterion panelists had used to define mastery. This criterion associated an adequate number of items with each achievement level. Of the 68 score points available, 17 were associated with the Basic level, 31 with Proficient, and 17 with Advanced. Three score points were

"too difficult" to be associated with any level. Panelists individually rated the items as "OK," "definitely use," and "do not use," based on the match between item content and the achievement level description.

ACT did not apply a fixed criterion to panelists' ratings in order to identify items that either should or should not be used as exemplars. Rather ACT recommends that panelists' ratings be used along with other information, such as item type (e.g., multiple choice or constructed response), item content, and other statistical criteria, such as discrimination, to select exemplar items.

In the full report, ACT demonstrates that adequate numbers of potential exemplar items or score points would be available for each achievement level even after applying a reasonable, fixed criterion to panelists' ratings. Items that survived the criterion include polytomously-scored constructed response items whose score levels, or points, are associated with different levels of achievement.

KEY ISSUES AND CONCLUSIONS REGARDING THE MAPMARK METHOD

The ALS method developed in this project supplements a relatively new standard setting procedure, bookmark, with the latest developments in item mapping and domain score theory and technology. The following is a summary, by topic, of key issues, conclusions, and recommendations that are presented in the full report to guide future standard setting activities using Mapmark or a similar method. Please refer to the corresponding sections in the full report if more detail is needed to understand the summaries presented here.

The Response Probability Criterion

The response probability (RP) criterion used in item mapping and bookmark-based standard setting procedures should be treated as a policy decision by the policy making body responsible for setting performance standards. The RP criterion determines the basic task panelists perform and different cut scores are likely to result from the use of different RP criteria. Human factors, convention, and the reasonableness of results, as well as statistical criteria, should be considered in selecting an RP criterion. The 0.67 RP criterion was a good choice from all perspectives in this project and ACT recommends that it be one of the choices considered in future standard setting activities.

The Use of Domains

Domains were developed in this project for the purpose of helping panelists understand student performance on the test and make reliable inferences about student achievement as an increasing progression of knowledge, skills, and abilities. They were designed to be understood by teachers, non-teacher educators, and the general public. They proved useful for this purpose and were successfully integrated into the standard setting process. Panelists were able to understand the domains and found them useful in understanding the achievement levels. They had no difficulty using domain score feedback to select scale values for their cut score recommendations.

ACT recommends that NAGB explore the use of the domains developed in this project for reporting student achievement on the 2005 NAEP in Grade 12 Mathematics. More generally, ACT recommends that NAGB explore whether domains having properties like those developed in this project—coherent and covering a wide range of difficulty—can be developed in other subject areas. In the long run, it would be most advantageous to incorporate the goals that guided domain development in this project into the framework and test development process. In the short term, ACT recommends that domains based directly on the framework or specially-developed for standard setting be incorporated into the standard setting process.

Identification of Item Knowledge, Skills, and Abilities (KSAs)

ACT and its technical advisory committee felt that the identification of knowledge, skills, and abilities (KSAs) required by assessment items in the context of the ordered item booklet (OIB) was the key, most useful feature of the Mapmark process. This feature was adopted without modification, except for the concurrent use of an item map, from the traditional bookmark method. The "KSA review" helped panelists understand student achievement on the assessment as a progression of increasing knowledge, skills, and abilities. This understanding was essential in translating the achievement level descriptions, which are themselves descriptions of progression in student achievement, into cut scores on the assessment.

By ordering items according the RP criterion and student performance data in the KSA review, the role of probability judgment in panelists' task is minimized and panelists are free to concentrate more on test content, on what higher levels of performance on the test mean, and on mapping the achievement level descriptions to actual levels of student performance. ACT recommends that student performance data continue to be used in this way to guide panelists in their content-related judgments and to minimize the role of probability estimation in their tasks.

The Use of Item Maps

Item maps proved to be a valuable addition to the bookmark components of Mapmark and were essential in the components of Mapmark that used domains. Generally, an item map provides a simple, comprehensive, visual layout of the standard setting problem on which panelists can keep track of process details and the overall results of the process. Items are located by difficulty on a scale that represents both student achievement and item difficulty. Panelists were able to keep track of meaningful differences between items, of where cut scores lay with respect to items, and of the magnitude of difference between achievement level boundaries, and between a given boundary or cut score and any given item. In Mapmark, item maps had the added feature of items organized into columns representing areas of similar content. Panelists understood their item maps and found them useful in their tasks. ACT recommends that item maps be used in any future method of standard setting.

The Concept of Borderline Performance

ACT found that panelists have no difficulty developing a concept of borderline performance independently in the process of placing their Round 1 bookmarks. The KSA

review prepares them to make this judgment. The concept of borderline performance was discussed in great detail over subsequent rounds with reference to criterion-referenced feedback about the median cut score. At the same time, plots and item maps facilitated discussions of how borderline performance differed from "typical" performance or "high" performance within achievement levels. Panelists' concept of borderline performance became well-formed over the course of the meeting and they reported high levels of understanding regarding the distinction between borderline performance and typical performance within the level. ACT recommends that panelists be allowed to provide their own concept of borderline performance in the process of placing their Round 1 bookmarks in future standard settings based on a bookmark procedure.

Independence Among Panelists

ACT recommends that bookmark-based standard setting processes be implemented in a way that encourages panelists to learn from the perspective and experience of other panelists, but to maintain their own perspective and independent judgment. Specifically, they should not be asked to develop a consensus on the cut score at the table, group, or whole group levels over rounds. This recommendation is based on considerations regarding the reliability of cut scores and on a theory of decision making presented in the book, *The Wisdom of Crowds* (Surowiecki, 2004). This approach prevents dominant panelists from skewing the results of the standard setting process to their particular point of view and recognizes that pressure to conform can reduce the contribution that difference in perspective and experience adds to a process such as standard setting. With the emphasis on independence in Mapmark, cut scores among panelists had about the same level of variability, convergence over rounds, and reliability that ACT has seen in past standard setting work for NAGB.

Identifying a Range of Uncertainty

In the instructions to panelists for placing their bookmarks, it is important to tell panelists to identify a range of items for possible bookmark placement. Panelists are told to go through the OIB in the easy-to-hard direction to find the place where items should be divided into two groups—those that the borderline student should be able to master (defined as having at least a 0.67 probability of correctly answering the item) and those that are too difficult for this expectation. ACT observed that some panelists have a tendency to divide the items at the first item they come to that seems too difficult for the borderline student. Given the judgment error and other sources of error in the process, this could result in the standard being set lower than the panelist intended. ACT instructed panelists to "go beyond the first item that seems too difficult" to make sure there are no later items that may belong in the mastery category. This amounts to identifying a range of items about which the panelist may not be sure should or should not be mastered, taking this range into account in their decision.

ACT found that panelists responded well to this idea. Some panelists used it to effectively apply the RP criterion in their task. They associated the range of uncertainty with a group of three or more items, at least 2/3 of which they felt the borderline student should be able to answer correctly.

RECOMMENDATIONS

ACT's principal recommendations concern the three principal outcomes of the Achievement Level Setting Process—Achievement Level Descriptions, Cut Scores, and Exemplar Items—and the use of the domains developed in the project.

- ACT endorses the Achievement Level Descriptions.
- ACT recommends the cut scores from Round 4 of the ALS meeting.
- ACT recommends that NAGB use the lists of items and panelists' ratings from the ALS meeting in the process of selecting exemplar items.
- ACT recommends that NAGB explore the use of domains in describing achievement levels for Grade 12 mathematics

The basis for these recommendations is provided in the full report.

CONCLUSIONS

The achievement level setting process had procedural validity and produced results that are reliable and likely to be useful and considered reasonable by parents, educators, policy makers, and the general public.

Developing Achievement Levels for the 2005 NAEP in Grade Twelve Mathematics: Process Report

INTRODUCTION

Background on NAEP Achievement Level Setting Activities

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. Achievement levels have been set for NAEP assessments in Reading, Writing, Mathematics, Science, History, Geography, and Civics. Achievement level percentages—the percent of students at-or-above each achievement level—have become the principal means by which educational policymakers assess the nation's progress in meeting its educational goals.

There are three components of NAEP achievement levels: Achievement level descriptions (ALDs), cut scores, and exemplar items. Achievement level descriptions are brief descriptions specific to the subjects and grades assessed in NAEP (4th, 8th, and 12th) of what students should know and be able to do in each level—Basic, Proficient, and Advanced. Cut scores are numerical representations of the lower borderline of each level. Exemplar items are matched with achievement levels in order to illustrate the kinds of tasks, knowledge, and skills required for performance at each level.

As currently specified by NAGB policy, there are 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. Stage 2 has traditionally been performed in an achievement level setting meeting (ALS meeting) by a panel of teachers, non-teacher educators, and representatives of the general public. The targeted percentages of these panelist "types" are, respectively, 55%, 15%, and 30%. This is in keeping with NAGB policy that the development of achievement levels shall be a widely inclusive activity.

Other details of the NAEP achievement level setting process are specified in contracts for achievement level setting activities. NAGB policy states that the Board will "ordinarily engage the services of a contractor who will prepare recommendations for the Board's consideration on the levels, the descriptions, and the exemplar exercises." NAGB typically contracts for both stages of the process—the development of the ALDs and the translation of the ALDs into cut scores. The contractor ordinarily obtains input from the achievement

level setting panel on the suitability of items as exemplars. NAGB contracts call for field trials, pilot studies, and other research activities designed to improve the standard setting process and the way that standard setting results are reported.

Ultimately, the setting of achievement levels is an exercise of policy judgment by NAGB. Key criteria in NAGB's policy judgment are the validity and reliability of the achievement level setting process and the apparent reasonableness of results. The Final Reports specified in NAGB contracts for achievement level setting activities are the principal means of documenting these criteria for specialists in the field as well as for the general public.

Background on the Current Project

Recent changes in the mathematics framework for the National Assessment of Educational Progress (NAEP) led NAGB to issue a procurement for setting new mathematics achievement levels for Grade 12. The new framework includes topics from more advanced courses up to and including Algebra II and Pre-Calculus. The Grade 12 assessment will contain more items on these, more advanced topics, and fewer items on pre-Algebra and earlier content. A related change at Grade 12 only (not Grades 4 or 8) is that items representing two of the five content areas in the framework, Measurement and Geometry, are combined into a single subscale in the psychometric scaling of test results. The scale for reporting 2005 NAEP results for Grade 12 mathematics will be a weighted average of four, rather than five unidimensional subscales: 1) Number Properties and Operations, 2) Measurement and Geometry, 3) Data Analysis and Probability, and 4) Algebra and Functions.

Item statistics and student distribution data for all achievement level setting activities in this project are based on a field test of the 2005 NAEP (in Grade 12 mathematics) that was conducted in the spring of 2004. This field test is referred to throughout this report as the "2004 field test," and the 2005 NAEP in Grade 12 mathematics will be referred to as the 2005 assessment. The schedule of activities in this project was based in part on the fact that item statistics and student distribution data from the 2004 field test were not available until June 2004. The field test was administered to a large, nationally representative sample of approximately 10,000 students, thus providing a reasonable, but not error-free, prediction of the achievement level percentages that would be reported for the 2005 assessment if NAGB adopted the cut scores recommended by ACT.

It is important to understand that differences between past achievement level percentages and those resulting from the cut scores set in this project are fundamentally uninterpretable. Scores on tests constructed from different frameworks are not comparable. NAGB decided to report results for the 2005 NAEP in Grade 12 mathematics on a different metric—one that clearly prevents comparisons of future (2005) to past test scores. Unfortunately, this does not prevent comparisons between old and new achievement level percentages from occurring. One must keep in mind, however, that differences in the percent at-or-above an achievement level in such a comparison cannot be attributed to any one factor such as change in student achievement, change in the ALD, or change in the method used to translate the ALD into a cut score. Rather, all of these factors, and more,

combine in unknown and unknowable ways to produce the observed difference in the achievement level percentage.

This report provides a detailed description of the method and outcomes of a meeting that was held in November 2004 to set achievement levels for the 2005 National Assessment of Educational Progress (NAEP) in Grade 12 Mathematics. It also summarizes project activities that preceded the ALS meeting. Project activities included the development of a new standard setting method, Mapmark, through a series of special studies, and the development of content domains for use in the Mapmark method. Mapmark uses the bookmark procedure (Mitzel, Lewis, Green & Patz, 2001) in Round 1, and provides domain score feedback in Round 2 and subsequent rounds. Item maps are used in every round of Mapmark. An item map shows the test items arranged on a linear continuum that represent both item difficulty and student achievement.

In its proposal to NAGB, ACT provided reasons for developing the Mapmark procedure rather than relying solely on the Angoff-based procedure ACT used in the past to set standards for NAGB (Reckase, 2000). The bookmark method was introduced as recently as 1996 (Lewis, Mitzel, & Green, 1996). Since then, it has become the most widely-used standard setting method in state assessments. ACT believed that the bookmark method contains some very attractive features for setting standards, but that it could be improved with the use of item maps (Masters, Adams, & Loken, 1994) and domain-score feedback (Schulz, Lee, & Mullen, 2005). ACT had conducted extensive research on these issues through previous standard setting contracts with NAGB (Reckase, 2000), through other NAEP-related projects (Schulz, Lee, & Mullen), and in support of its own assessment programs (Schulz, Kolen, & Nicewander, 1999).

ACT consulted with its Technical Advisory Committee on Standard Setting (TACSS) in all aspects of the project. The TACSS is a five-member group that collectively represents expertise in standard setting, mathematics education, and experience with the NAEP. (See Appendix B for a list of the TACSS members.) The TACSS met six times over the course of the project and provided input on key components of the project including the design of the Mapmark method, the design of special studies, the conduct of the ALS meeting, data analysis procedures, and the formulation of conclusions and recommendations presented to NAGB.

CONTRACT ACTIVITIES PRIOR TO THE ALS MEETING

Contract activities prior to the ALS meeting fall into two general categories: 1) domain development and 2) achievement level setting studies. These activities are briefly described in the following sections.

Domain Development Activities

ACT proposed to develop for use in the Mapmark method, the kinds of domains that would be most useful for describing to educators and noneducators alike, in a clear and reliable fashion what it is that students at a given level of achievement can or cannot do, and what growth in achievement means (Schulz, Lee, & Mullen, 2005). This meant using methods

similar to those used by Schulz, et al., (2005) to define, within each subscale of the assessment, "teacher domains." The teacher domains were to have the following features:

1) Well Defined. Each domain should be well represented by a *domain definition* consisting of a title, a brief narrative description, and up to three sample items (if available). The title and narrative should represent in relatively jargon-free language that can be understood by teachers, non-teacher educators and the general public alike, the knowledge, skills, and abilities required by items in the domain. Sample items should be drawn from released items on the NAEP website.

2) Coherent. Teachers should be able to reliably and independently classify items into the domains by content, using only the domain definitions. Standard setting panelists should be able to see and understand how items fit, or belong, in their domain.

3) Variable in Difficulty—the domains within each subscale should differ in difficulty and collectively cover a wide range.

The teacher domains were to be combined, if necessary, into no fewer than three and no more than five "Score Domains." The Score Domains would have the following features:

1) Reliable Separation. In terms of difficulty, the score domains within a subscale should cover a wide range—comparable to that of Basic, Proficient, and Advanced achievement level cut scores—and be as evenly-spaced as possible. Each score domain should be large enough in terms of the number of items it represents and/or distinct enough in difficulty from other score domains so one could reasonably expect the score domains to have the same relative and absolute difficulty if another random sample of items for the domains were drawn from the NAEP item pool.

2) Coherent. The combination of teacher domains into score domains for the above two purposes should not be based solely or even primarily on putting teacher domains of similar difficulty together, but should also make sense in some fashion to teachers and curriculum experts.

The reason for defining no fewer than three and no more than five score domains was so standard setting panelists would have enough detail (at least three), but not too much detail (more than five) for their domain-related work.

Table 2 summarizes steps in the development of the Grade 12 domains. Activities conducted before June 15 used all items in the Grade 12 NAEP pool deemed by the test development contractor to be relevant to the new framework, both secure and publicly-released items. Special consideration was given to items in the 2005 assessment, since the domain scores presented to Mapmark panelists in the Grade 12 ALS meeting would be

based solely on items in the 2005 assessment. The statistics necessary to compute domain scores were not available for all items in the 2005 assessment, however, until June 2004.

Table 2: Development of Grade 12 Content Domains

Event/Process	Date/deadline (2004)	Purpose/Product
Domain Development Meeting	January 8-12	First Draft of Teacher Domains
Refine	By February 21	Second Draft of Teacher Domains
Item Classification Study	February 28	Evaluate Domain Coherence
Refine	By June 15	Teacher Domains Item Classifications Score Domains

Table 3 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. 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 6 shows percent correct curves for teacher domains and score domains representing the Measurement and Geometry subscale. The curves are based on the items for the 2005 assessment. It can be seen that the score domains are fewer in number and are more evenly spaced over the range of the Mapmark scale. It can also be seen that score domains were not created simply by combining teacher domains that were closest together in difficulty. Rather, the combination was based on perceived similarities of content and/or curriculum, as judged by experts on the domain composition team.

Figure 7 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 3: Titles of Teacher Domains and the Correspondence between Teacher and Score Domains by Subscale of the 2005 Assessment

Number Properties and Operations

Teacher Domain	Title	Score Domain
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	

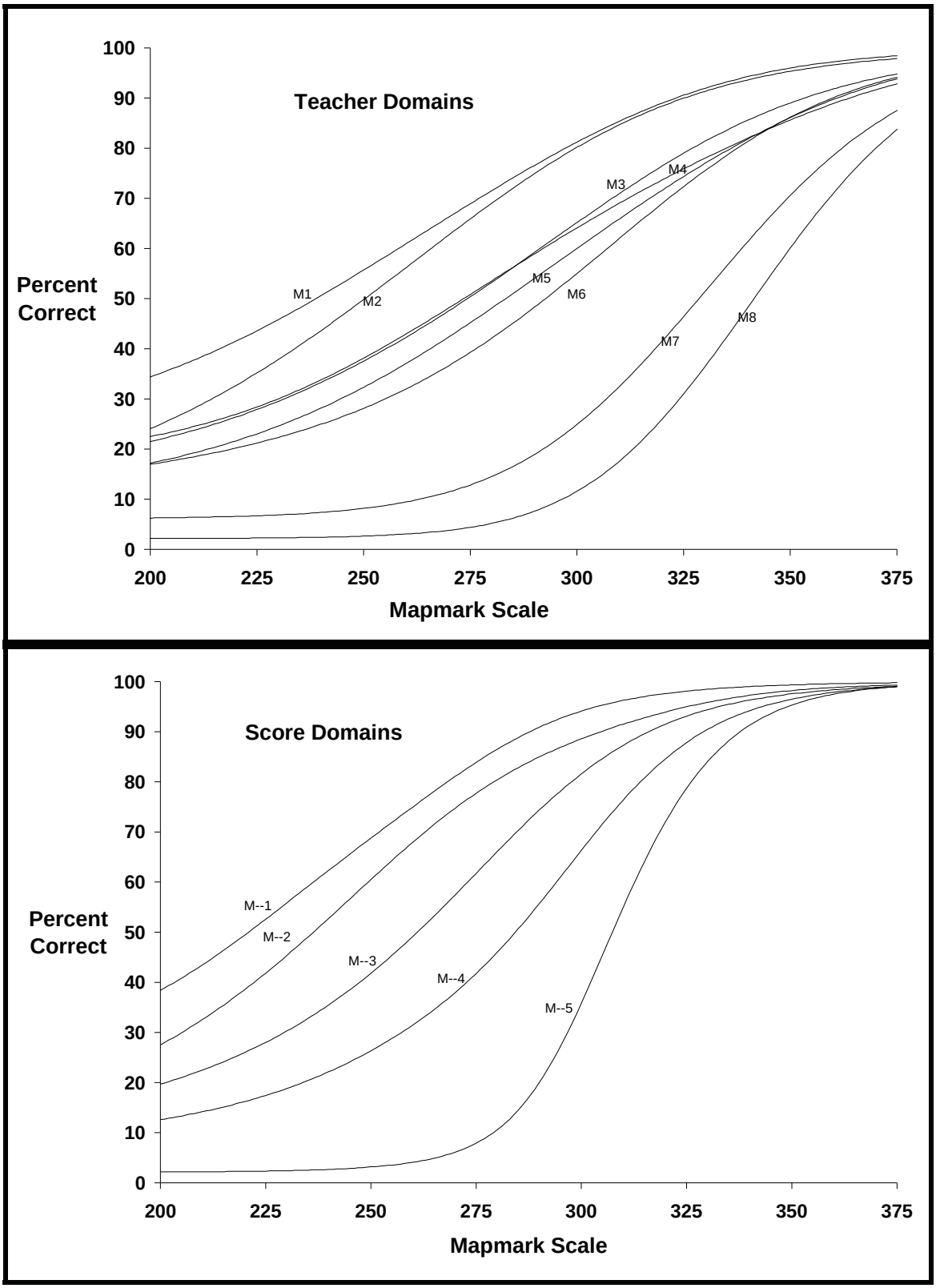


Figure 6. Percent correct curves for Teacher and Score Domains in the subscale representing Measurement and Geometry Content Areas.

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.

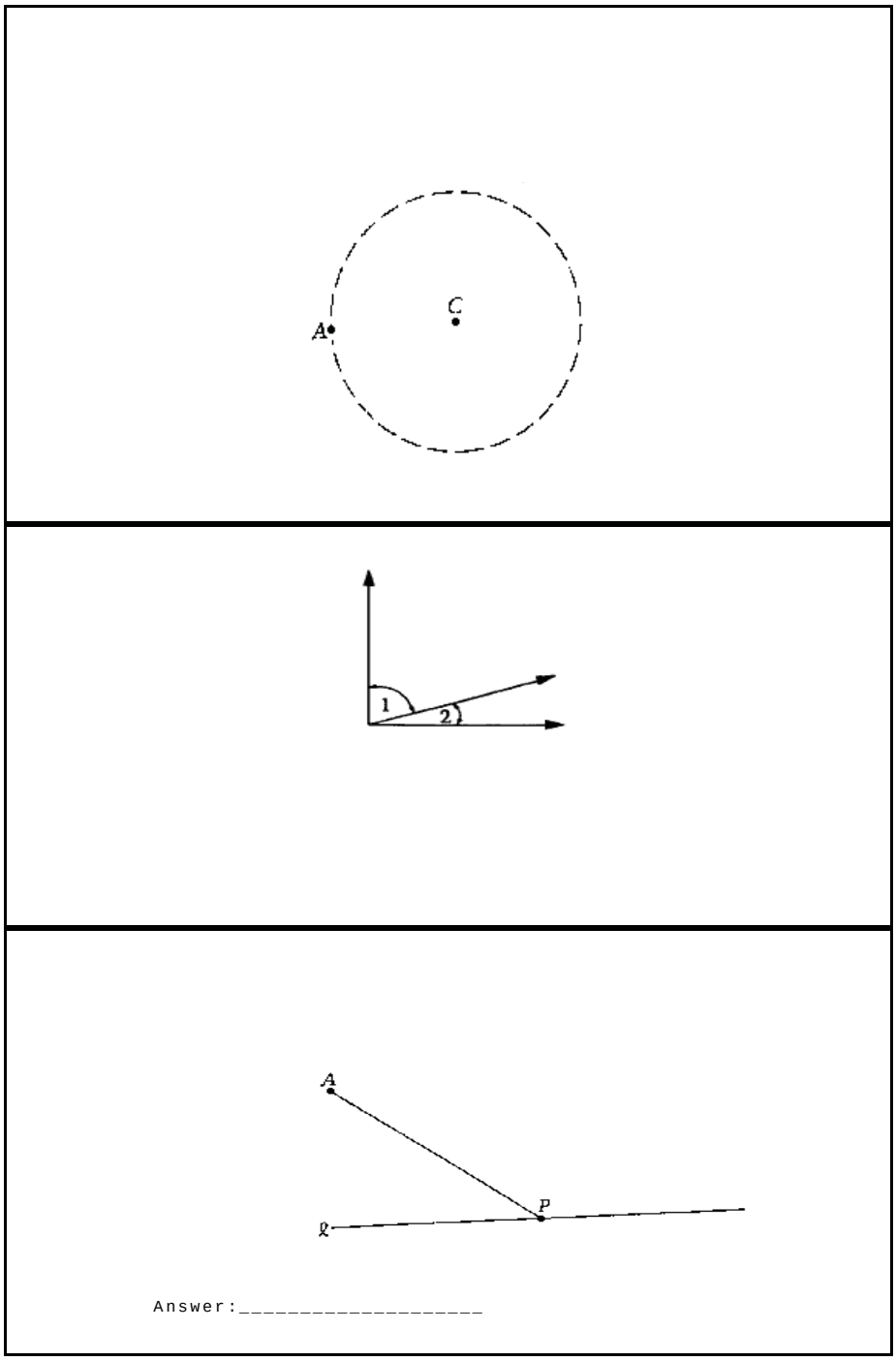


Figure 7. Domain definition for Teacher Domain M4.

Achievement Level Setting Studies

A number of achievement level setting studies took place for the purpose of developing and evaluating Mapmark. Table 4 shows the timing and purpose of the studies. The field trials and Grade 8 Study used items from the 2003 Grade 8 mathematics assessment. Items from this assessment were classified into domains that had been defined in an earlier study (Schulz, Lee, & Mullen, 2005). This allowed the domain-related features of Mapmark to be developed and evaluated before Grade 12 domains were ready. The Pilot Study, like the ALS meeting itself, used items from the 2005 Grade 12 mathematics assessment. Grade 12 domains for the Pilot Study were created in this project as described in the previous section.

Table 4: Achievement Level Setting Studies

Study	Date (2004)	Items	Purpose
Field Trial 1	January 15-16	50% of Grade 8 items in 2003 NAEP	Mapmark development
Field Trial 2	February 12-13		
Grade 8 Study	March 11-15	All Grade 8 items in 2003 NAEP	Mapmark development and evaluation relative to 1992 ALS method
Pilot Study	July 15-19	All Grade 12 items in 2005 NAEP	Mapmark evaluation relative to 1998 ALS method

Field Trials

The field trials were designed to try out key features of the Mapmark method. Key questions centered on panelists' understanding of their item maps and domain-related materials and on whether they actually used them in evaluating and recommending cut scores. Another important issue in the field trials was the selection of an RP criterion. The choice of RPs for the field trials was influenced by several factors, but ultimately, RPs of 0.67 and 0.5 were chosen for Field Trials 1 and 2, respectively. Besides these major issues, there were many questions about procedure and design of materials to be answered.

Process evaluation results indicated that item maps and domain scores in the Mapmark process were effective and understood by panelists. Panelists generally gave high ratings concerning their understanding and use of domain-related materials and feedback. Panelists also understood and used their item maps. Panelists were comfortable using item-level and domain-level information together to recommend cut scores and were also comfortable selecting a scale value, rather than placing bookmarks, to recommend cut scores after Round 1.

In debriefing sessions, panelists in both field trials indicated that they took the RP criterion into account when placing their bookmarks. However, panelists using the 0.5 RP expressed difficulty deciding how to place a bookmark with respect to content that students should know and be able to do according to the achievement level descriptions. The 0.5

RP represented only an "even chance" of getting a bookmarked item correct. Some panelists placed their bookmark on easier items in order to adjust for the lower RP. They should have placed their bookmark on harder items. Panelists using the 0.67 RP expressed no similar difficulty or confusion. They were comfortable associating a 0.67 probability with "mastery" of an item's content.

Cut scores resulting from the field trials are shown in Table 5. Both RP criteria yielded cut scores that were lower than those set in 1992 using an Angoff-based method. Cut scores set using the 0.5 RP were lower by 35 points, 21 points, and 10 points, respectively, for Basic, Proficient, and Advanced. Cut scores set using the 0.67 RP were lower by 10 points, 8 points, and 1 point, respectively.

Table 5: Cut Scores from Special Studies Using Grade 8 Assessment Data

Source	Achievement Level		
	Basic	Proficient	Advanced
Current (Est. 1992)	262	299	333
Field Trial 1	252 (-10)	291 (-8)	332 (-1)
Field Trial 2	227 (-35)	270 (-21)	323 (-10)
Grade 8 Study	251 (-11)	289 (-9)	331 (-2)

Note: Values in parentheses are the difference between current and special study cut scores.

The consequences of the cut scores are shown by the achievement level percentages in Figure 8. It can be seen that the percentages of students at or above achievement levels are generally higher using Mapmark Field Trial cut scores than the cut scores set in 1992 using an Angoff-based method. Given no change in the assessment framework or achievement scale since 1992 this result is suggestive of a method effect on cut scores, however, other factors could also explain the difference as described in the next section. In any case, it is conceivable that differences from the 1992 cut scores could have been reduced, at least for Basic and Proficient levels, by selecting a higher RP. However, it did not seem wise to choose an RP criterion solely for this reason. The achievement level percentages from the 0.67 RP (Field Trial 1) seemed quite reasonable when considered on their own. In consultation with TACSS and NAGB staff, ACT decided to use a 0.67 RP for the Grade 8 Study.

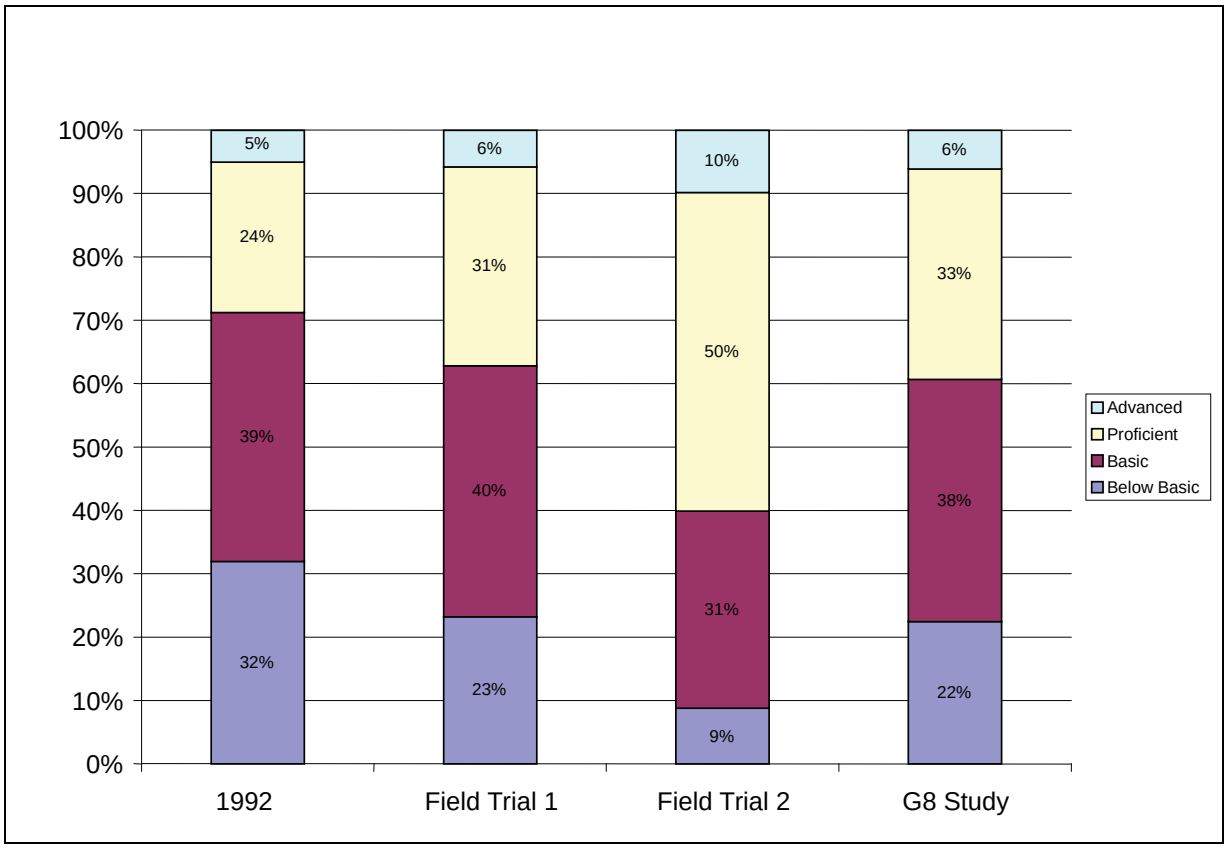


Figure 8. 2003 NAEP achievement level percentages by study.

The Grade 8 Study

The purpose of the Grade 8 Study was twofold: 1) it allowed further refinement and testing of Mapmark procedures and materials, and 2) it provided a comparison of a complete Mapmark process to an ALS process used previously by ACT to set cut scores for NAGB—the 1992 ALS process. Though some further refinements of the Mapmark process were to be expected, the process was essentially complete in that it included all of the essential elements ACT used in the past, such as national recruitment of panelists, advance mailing of a briefing booklet, and a five-day schedule that included full orientation to the test and the process, administration of the test to panelists, and presentation of consequences data (achievement level percentages) to panelists.

Like the field trial panelists, Grade 8 Study panelists gave positive feedback on materials and procedures related to item maps, domains, and other elements of Mapmark that were tried for the first time in the field trials. On an "agreement" scale of 1 to 5 with 5 being "totally agree," panelists reported that they understood what "mastery" means (4.50) and that they were comfortable using a 0.67 probability to place their bookmarks (4.70).

As seen in Table 5, cut scores from the Grade 8 Study were lower than the 1992 cut scores. The lower cut scores placed a higher percent of students at-or-above each achievement level than the 1992 cut scores (Figure 8). While this difference is suggestive of a method effect, the difference in cut scores could also be explained by factors not strictly related to the methods. Two, of many, possible confounding factors were 1) change in the educational climate for standard setting since 1992, and 2) change in NAGB policy for setting cut scores.

Regarding change in educational climate, field trial panelists had repeatedly mentioned the federal "No Child Left Behind" (NCLB) legislation which was passed after 1992. To counter a possible effect, facilitators and presenters explained that the legislation had no direct connection to NAEP standards and stressed the criterion-referenced nature of their task and the importance of the ALDs. On a scale of 1 to 5, with 1 representing "totally disagree," panelists tended to disagree with the statement, "NCLB will probably influence my judgments in this standard setting meeting" (1.55). Still this and other changes in the educational climate could have influenced panelists using any method to adopt a more liberal concept of "lower borderline performance" for their tasks.

Regarding changes in NAGB policy, it was noted that the ALDs used in the Grade 8 Study were developed in the 1992 process during the ALS meeting. It is not clear what cut scores would have resulted from a modified-Angoff process that began with the ALDs, as specified in NAGB policy adopted since 1992.

A higher RP-criterion could conceivably yield Basic and Proficient cut scores closer to those set in 1992, but it did not seem wise to select an RP criterion on this basis alone. Achievement level percentages from the Grade 8 Study alone might be considered quite reasonable by the mathematics education community. In consultation with TACSS and NAGB staff, ACT chose to use a 0.67 RP for Mapmark in the Pilot Study.

The Pilot Study

In the Pilot Study, two fully-operational standard setting methods, Mapmark and Item Rating, were compared using data from the 2004 field test (of the 2005 NAEP Grade 12 mathematics assessment). The Item Rating method was fundamentally similar to the method ACT used to set achievement levels for the 1998 NAEP in Civics. Essential elements of the 1998 ALS procedure are described in the Final Report (Loomis & Hanick, 2000). The essential elements of the Pilot Study Mapmark procedure and differences from the ALS meeting are described in the ALS meeting process section of this report.

Table 6 shows average ratings on summary process questions by method. ACT and its TACSS judged the process of both methods to be acceptable, but noted that there was room for improvement in the clarity of instructions in the Mapmark method. Average responses to summary questions on clarity of instructions and panelists' understanding of their tasks were statistically lower in the Mapmark method compared to Item Rating. Improvement in instructions would be expected to lead to improvement in panelists' understanding of their tasks. The need for improvement was traced to bookmark placement instructions in Round

Table 6: Average Responses to Summary Process Questions by Method

Key Summary Questions			Item Rating	Mapmark
Instructions during each round			4.30	3.81
Absolutely Clear 5	Somewhat Clear 3	Not at All Clear 1		
Understanding of the tasks during each round			4.40	3.86
Totally Adequate 5	Somewhat Adequate 3	Totally Inadequate 1		
Confidence in my ratings/judgments			4.20	4.19
Totally Confident 5	Somewhat Confident 3	Not at All Confident 1		
Effectiveness of this ALS process			4.00	4.00
Highly Effective 5	Somewhat Effective 3	Not at All Effective 1		
Opportunity to use my best judgment			4.35	4.19
To a Great Extent 5	Somewhat 3	Not at All 1		
ALS achievement levels are defensible			4.30	4.00
To a Great Extent 5	Somewhat 3	Not at All 1		

1 and the introduction to domains and domain task instructions in Round 2. Based on comparisons to the Grade 8 Study, where ratings of instructions were generally high, ACT was able to pinpoint differences that accounted for the lower ratings in the Pilot Study. ACT assured NAGB that ratings on instructions and understanding could be improved if Mapmark were selected for the ALS meeting.

Statistical analyses of cut scores showed that both methods had acceptable reliability. Both item pool effects and table effects were modest. Estimates of the standard error of cut scores ranged from 2 to 4 points for both methods. These estimates are conservative, since they do not take into account all factors that vary across independent replications of the same method. Differences between pilot study cut scores and cut scores obtained in corresponding operational ALS meetings across several years and subject areas (civics, writing, science, history, and geography) of ACT/NAGB standard setting averaged 5.4 points.

Given these reliabilities, cut scores from the two methods were different, but could not be regarded as very different, from each other. On a 300-point scale, ranging from 100 to 400, the Basic cut scores from the two methods differed by just three points (239 for Mapmark vs. 242 for Item Rating), the Proficient cut scores differed by seven points (270 for Mapmark vs. 277 for Item Rating), and the Advanced cut scores differed by nine points (316 for Mapmark vs. 307 for Item Rating). A majority of panelists in both methods endorsed the cut scores after viewing the achievement level percentages. Table 7 shows the corresponding achievement level percentages, along with the achievement level percentages last reported for the Grade 12 assessment in the 2000 NAEP.

Table 7: Grade 12 Achievement Level Percentages by Year and Method

Year and Method	Below Basic	Basic	Proficient	Advanced
2000 NAEP*	35%	48%	14%	2%
2004 Pilot, Item Rating	41%	40%	16%	4%
2004 Pilot, Mapmark	35%	35%	28%	2%

*The 2000 NAEP involved a different framework, assessment, achievement level descriptions, and scale from the 2004 data in this table.

ACT presented the results of the Pilot Study to the NAGB Committee on Standards, Design, and Methodology (COSDAM). ACT and its TACSS indicated a slight preference for the Mapmark method. COSDAM chose the Mapmark method for the operational ALS meeting. The following points were noted by COSDAM.

- Results from the methods (Item Rating and Mapmark) differed somewhat from each other and both sets had one or more notable differences from previous achievement level percentages.
- The change in framework presents a natural opportunity to consider a new achievement level setting method because new achievement level percentages do not have to agree with previously reported achievement level percentages.
- Both methods (Item Rating and Mapmark) are valid and the achievement level percentages produced by either method are likely to be viewed as reasonable.
- Mapmark is more likely to be understood by states because it has basic similarities to the bookmark procedure, which has become the most widely used standard setting method in state assessments.
- ACT gave assurances that Mapmark could be conducted in four days without compromising the integrity of the process. COSDAM felt that this flexibility was important for recruiting panelists and maximizing panelists' effort and satisfaction with the process.

Based on the COSDAM decision, ACT implemented the Mapmark method in the operational ALS meeting.

THE ACHIEVEMENT LEVEL SETTING PROCESS

Achievement level setting in NAEP refers to the overall process through which the three outcomes of 1) achievement level descriptions (ALDs), 2) cut scores, and 3) exemplar items are obtained. The Achievement Level Setting (ALS) meeting is just one part of the

process. Activities leading up to the ALS meeting include the development of the ALDs, recruitment of panelists, and mailing of advance materials.

Developing the Achievement Level Descriptions

NAGB staff coordinated the development of the ALDs in this project. The ALDs were developed in advance of the Pilot Study and ALS meeting. They are shown in Appendix A. The development, preliminary approval, and evaluation of the ALDs included, but was not limited to, the following steps:

- An ALD panel consisting of six experts who were members of the task force that developed the 2005 framework for NAEP mathematics met at NAGB headquarters in Washington, DC on February 23, 2004. They produced a draft consisting of a preamble and, for each achievement level, a bulleted list of knowledge, skills, and abilities organized by subscale of the assessment.

- The draft was approved by NAGB at their meeting March 4-6, 2004.

- The ALD panel subsequently developed a narrative paragraph for each achievement level. The paragraphs summarized the corresponding bulleted lists.

- The ALDs were used and evaluated in the Pilot Study. Results of this evaluation supported the use of the ALDs, as written, in the ALS meeting.

ALS Panelist Selection

ACT implemented the same basic design for selecting panelists to set achievement levels that ACT had used for the 1998 ALS process. Primary requirements were that the panel be broadly representative, and that 70% be educators and 30% noneducators. Moreover, classroom teachers should comprise 55% of the group. In addition to these primary requirements, both demographic characteristics and group size were key considerations in the selection of panelists.

The process of selecting panelists had several steps. The following summary highlights the main features of each step in the process of selecting panelists to set achievement levels.

Selection of School Districts

School districts served as the basic sampling unit for the panelist selection process. Principles of sampling were used for drawing stratified random samples of school districts from a national database. ACT drew samples that were proportional to the regional share of districts. The regional proportions were as follows:

Northeast	21%
South	23%
Midwest	37%
West	19%

The samples of districts were drawn to include at least 15% with enrollments of 25,000 or more students, and 15% with at least 25% of the population below the poverty level. A total of 577 public districts and 110 private schools were sampled. Please see Table 8 for the distribution of districts sampled. The total number of districts selected and the proportion in each nominator type were based on previous experience with response rates from nominators in other subjects.

Table 8: Distribution of Districts Sampled

Nominator Type	Public Districts	Private Districts	Total
Teacher	318	98	416
Nonteacher Educator	36	12	48
General Public	223	--	223
Total	577 (84%)	110 (16%)	687

Identification of Nominators

ALS nominators were identified by drawing three separate samples of districts without replacement. One sample of public school districts was drawn from which nominators of teacher panelists were identified, a second for nominators of nonteacher educators, and a third sample for nominators of general public representatives. Nominators of private school teachers were identified from a sample of private schools drawn separately. A total of 1,385 nominators were contacted. Please see Table 9 for the distribution of nominators. Nominators were persons holding a specific title or position, such as the following.

Table 9: Distribution of Nominators Contacted

Nominator Type	Public Districts	Private Districts	State	College/ Universities	Employers	Total
Teacher	444	49	45	--	--	538
Nonteacher	36	12	24	25	--	97
General Public	548	--	--	--	202	750
Total	1028 (74%)	61 (4%)	69 (5%)	25 (2%)	202 (15%)	1385

Nominators of teachers were:

- district superintendents
- leaders of teacher organizations
- state curriculum directors
- principals or heads of private schools

Nominators of nonteacher educators were:

- non-classroom educators (e.g., principals, district social studies curriculum coordinators)
- state assessment directors

- deans of colleges and universities (two-year and four-year; public and private)

Nominators of members of the general public were:

- education committee chairpersons of the local Chambers of Commerce
- mayors
- school board presidents
- employers of persons in a math-related position or with a math-related background

Nominators could submit up to four candidates whom they judged to be well qualified to serve as standard setting panelists. They were encouraged to nominate members of minority groups.

Selection of Panelists

Nominees represented a specific role (teacher, nonteacher educator, or member of the general public). A total of 167 candidates were nominated to serve as potential panelists.

A computerized algorithm was developed to select panelists from the pool of nominees. Nominees were rated according to their qualifications based on information provided on the nomination form (e.g., years of experience, professional honors and awards, degrees earned). Nominees with the highest ratings had the highest probability of being selected, other factors being equal. The selection program was designed to yield panels with:

- 55% of the members representing grade-level classroom teachers
- 15% of the members representing nonteacher educators
- 30% of the members representing the general public
- 30% of the members from diverse minority racial/ethnic groups
- up to 50% of the members male
- 25% of the members representing each of the four NAEP regions

Thirty panelists were required for the panel. Forty-six persons were selected from the nominee pool and contacted about serving as an ALS panelist. Some of the persons who were selected were unable to serve at the scheduled time. A total of 31 panelists participated in the ALS study representing 23 states. A list of the panelists who participated in the ALS is presented in Appendix C.

Advance Materials

Before the ALS meeting, all panelists were mailed materials that contained important background information on setting achievement levels. These advance materials were distributed across two separate mailings. The first mailing was sent on October 7, 2004. The cover letter for the first mailing contained instructions on how to make airline reservations and provided a brief description about what panelists could expect. Enclosures were:

- *2005 Mathematics Framework*;
- 2005 NAEP Mathematics Preliminary Achievement Level Descriptions; and
- Preliminary Agenda.

The second mailing was sent on October 28, 2004. The cover letter contained detailed instructions related to travel arrangements and accommodations. Enclosures were:

- Briefing Booklet;
- Confidentiality Agreement;
- Request for Press Release Form;
- Request for Taxpayer I.D. Number and Certification; and
- Hotel diagram and directions to the meeting.

A Briefing Booklet was first used by ACT for the 1994 ALS process. It includes a complete description of each step in the process including the purpose for each step, and defines key terms used in the standard setting meeting.

In addition to the materials that were mailed in advanced, ACT distributed various brochures about NAEP and NAGB to the panelists during the registration phase of the meeting.

The ALS Meeting

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 to 6:00 PM, except the last day, which adjourned at 12:30 PM. The agenda is shown in Appendix D.

Design Factors

Prior to the meeting, panelists were assigned to two groups of about 15 persons each: Group A and Group B. Each group rated a different, but overlapping set of items as explained in the next section. 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 and tables as equal as possible with respect to panelist type, gender, region, and race/ethnicity

Item Pool Division

All items in the 2005 assessment pool were used in the ALS meeting. There were a total of 180 items representing 237 score points. Items were in two basic formats: multiple choice and constructed response. Three types of items were identified for panelists in the following terms: 1) multiple choice, 2) dichotomously-scored (constructed response), and 3) polytomously-scored (constructed response). The numbers of items by type were 119, 24, and 37, respectively.

The item pool was divided into equivalent, but overlapping, pools for Groups A and B. Equivalence was monitored with regard to: 1) mean and variation of item difficulty, 2) representation of subscales, 3) number of items of each type, and 4) number of score points (steps) of polytomously-scored items. Item difficulty statistics were based on the 2004 field test.

The equivalence criteria were met by assigning blocks of items to the pools. Blocks are equivalent sets of 17 to 21 items created for purposes of test form construction. The 2005 assessment consists of ten blocks—3 through 12. Six blocks were assigned to each pool. The pools had two blocks in common. The "common blocks," blocks 3 and 4, were scheduled to be released to the public after the assessment. [Block 12 was also scheduled for release, but the item pools could not be balanced by assigning blocks if there were three common blocks.]

After the initial assignment by blocks, a few items were transferred from one group to another so that each pool would contain at least two items and at least three score points within each teacher domain. This reassignment did not change the equivalence of the pools in other respects, as can be seen in Table 10. This table summarizes the item pool for each group with regard to the key characteristics listed above. Detailed information about the item pools by block is presented in Appendix E.

Table 10: Item Difficulty Statistics and Number of Items by Subscale and Type Within Group

Group	Total Items	Subscale[1]				Item Type[2]			Item Difficulty (Scale values at RP[3] of 0.67)				
		1	2	3	4	MC	DI	Poly	N[4]	Mean	SD	Min	Max
A	107	19	37	23	28	70	15	22	142	278	37	141	401
B	109	19	34	21	35	73	13	23	143	279	40	157	401

[1] 1=Number Properties and Operations, 2=Measurement and Geometry, 3=Data Analysis and Probability, 4=Algebra and Functions

[2] MC = Multiple choice; DI = Dichotomously scored constructed response; Poly = Polytomously scored constructed response

[3] RP = Response Probability (of getting the item correct or earning the score point or higher)

[4] N = Number of score points greater than zero (total score if a student took all items and performed perfectly).

Facilitation, Observers, and Room Setup

The NAEP ALS Project Director served as the primary facilitator for the meeting. A member of the ALD panel, Mary Jo Messenger, served as the primary content facilitator. The bookmark component of the Mapmark process was facilitated by a bookmark process facilitator, Howard Mitzel, and a bookmark content facilitator, Jason Schwartz. Dr. Mitzel was a co-developer of bookmark and President of Pacific Metrics, a project subcontractor. Mr. Schwartz, Director of Test Development at Pacific Metrics, has an advanced degree in Mathematics and has extensive experience serving as a process and content facilitator for bookmark processes. All facilitators had participated in the Pilot Study and were experienced in the procedures performed in the ALS meeting.

Because the meeting involved only one grade and group of panelists, all sessions were held in the same room. Panelists were seated at a total of six tables (five panelists at each of

five tables and six panelists at one table). The entrance to the room was the "back," at which observers were seated. (See Figure 9.)

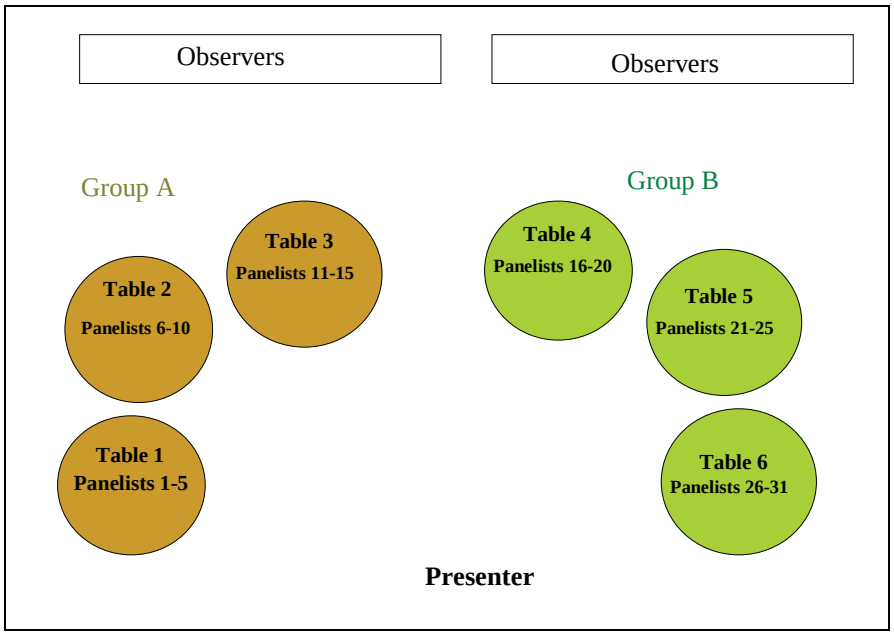


Figure 9. Room and table setup.

A total of ten observers were present at various times: two NAGB staff, two members of the NAGB Committee on Standards, Design, and Methodology (COSDAM), two external evaluators, three members of ACT's TACSS, and one member of ACT's internal Technical Advisory Team (TAT).

General Orientation

In a brief welcome and introduction session, panelists were introduced to key operational staff, process and content facilitators, and to the observers. The role of observers was explained and panelists were asked to limit their interactions with observers.

Following the welcome and introduction, panelists received a presentation on NAEP and NAGB by a NAGB staff member. This session covered the history, organizational structure, procedures, and key policies of the National Assessment of Educational Progress (NAEP) as well as the purpose of setting achievement levels. Information about the NAEP mathematics assessment was also presented.

Panelists were then given a general introduction to achievement level setting. This introduction explained how the ALS meeting fits into the overall NAEP achievement level setting process and described some basic concepts and procedures involved in organizing and conducting the meeting. Topics included how panelists were selected, the meaning of criterion-referenced standard setting, and general themes that account for the various presentations, tasks, and exercises in the ALS meeting. The themes were:

- understanding the context of achievement level setting,

- understanding the assessment and student performance,
- understanding the achievement levels,
- understanding the tasks performed in recommending cut scores,
- performing the tasks to recommend cut scores, and
- evaluating the process.

Taking a Form of the NAEP

In the final session of the morning, panelists took a form of the NAEP selected for their group. Item blocks in the form administered to Group A were not in the Group A item pool, and likewise for Group B. The panelists took the test under standard test-administration conditions for the NAEP. After completing the test, panelists reviewed their own responses using scoring guides. Panelists were told that their test would not be scored or used in any other way during the meeting.

Orientation to the ALS Method and Materials

Method-specific aspects of the ALS meeting began after lunch on Day 1 with an orientation to the Mapmark method. The purpose of this orientation was to give panelists a general overview of the process and to introduce them to key concepts and materials they would be using. Figure 10 shows a slide of a simplified item map. This slide was used to explain the general principal of an item map as spatially representative of a journey. In this case, the map represents the journey from low to high achievement. Points along the journey are represented by "markers," i.e., test items.

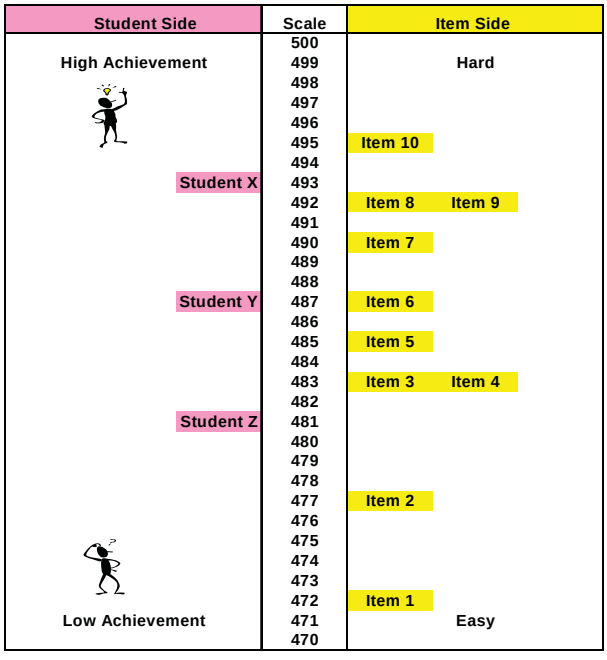


Figure 10. A simplified item map as spatially representative of a journey from low to high achievement.

Figure 11 shows a slide that was used to explain the role of the response probability criterion (RP-criterion) in determining the location of item "markers" on the map. This slide explains the location of Item 5 in Figure 10. This early-introduction to the RP criterion is important in helping panelists to understand this criterion and to take it into account in their bookmark placements.

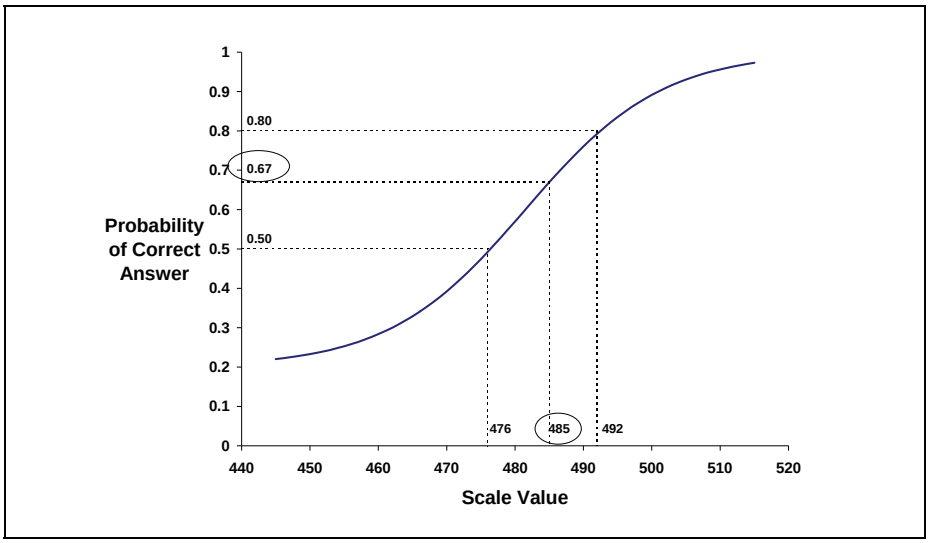


Figure 11. Explaining the relationship of the RP-criterion to an item's scale value.

Domains were introduced to panelists as areas of content more specific than that of the test as a whole, but broader than a single item. As such, two types of domains were identified for panelists to be aware of in the Mapmark process: Domains corresponding to the framework (subscales) and specially defined domains (teacher domains).

It was explained to panelists that both types of domains would be represented by columns on their item maps, but that only teacher domains would be used directly, in the form of domain score feedback, to recommend cut scores, and that domain-score use would begin in Round 2.

Figure 12 shows the Primary Item Map, on which columns correspond to subscales of the assessment. A Domain Item Map, in which columns corresponded to Teacher Domains, is illustrated later in this report (see Figure 26). The meaning of colors and other information in the Primary Item Map will be explained below.

Scale	Number Properties and Operations	Measurement/Geometry	Data Analysis	Algebra
Above			M117 M118 M119 P36_4	D24 P37_4
354				
351				
348				
345				
342				P37_3
339			P35_4	
336			M116 P34_2	P32_4 P33_3
333				D23
330			P31_2 P34_1	
327		P29_2 D22 P30_2		P33_2 D21
324		M115 P28_2 D20	D19	
321	P27_4	P26_2 M114 M113	D18	
318	P27_3 P24_2	P25_2		M112
315	P27_2	M110 P30_1 D17		P23_2 M111 P33_1
312	P27_1		M107	D16 M108 P22_3 P37_2 M109
309	M106	P21_4	P20_2	M105
306		M102 M103	P35_3	M101 M104 P19_2
303		P29_1 P28_1 P18_2 D15	M100	M99
300		M96 P18_1 P25_1 P17_2	M97	P32_3 M98
297	M95 P24_1	P16_2		M93 P23_1 M94
294	M86 D14	M88 M87 P14_2 M92	P13_2 P36_3	M89 M90 M91 P15_2
291	D11 M84		D12 P31_1	M85 D13 P22_2
288		P10_3 P11_2 M82	P9_4 P12_2	P8_3 P19_1 P15_1 M81 M83
285	M78	M74	P35_2 M75 M80	M76 M79 M77
282	P7_2	M70 P21_3 M73 M72	D10	M68 M69 M71
279		P5_2 P11_1 P17_1	M67 P13_1	P6_2
276	D9 M66 P4_2	M65	M63	M64 P32_2
273	P4_1 M62	M58 M60		M57 M59 M61 P22_1
270	M55		P20_1 P9_3	M56
267	M49 P3_2 M52	P26_1 P5_1	M53 M54	M51 M50
264	P7_1	D7 M45 M44 M47	P2_4 M46	D8 M48
261		M41 M42	P9_2	M43
258		P10_2 P21_2		M39 M40
255	M33 M37 M38 D6	M32 M34	P36_2 M36	M31 M35 D5
252	M30	M26 M28 P14_1	M27 P9_1 M29 P12_1	P8_2 P6_1
249			M25	
246	P1_2 M22 P3_1	M23 M24	P36_1	
243		M21 D4	D3	
240	M20			
237		M19	P35_1	M18
234		M17		P37_1
231		M15 M16 P10_1		
228		D2		P32_1
225		M13 M14	M12	
222		M11	P2_3 D1	
219		P21_1		
216	M10	M7 M9	M8	
213	M5 M6	P16_1		P8_1
210				
207				
204	M4			
Below	P1_1 M2	M3	P2_1 P2_2	M1

Figure 12. Primary Item Map.

Panelists were told that their Ordered Item Book (OIB) contained all of the items with which they would be working in order of their difficulty, beginning with the easiest. Figure 13 was used to illustrate this concept.

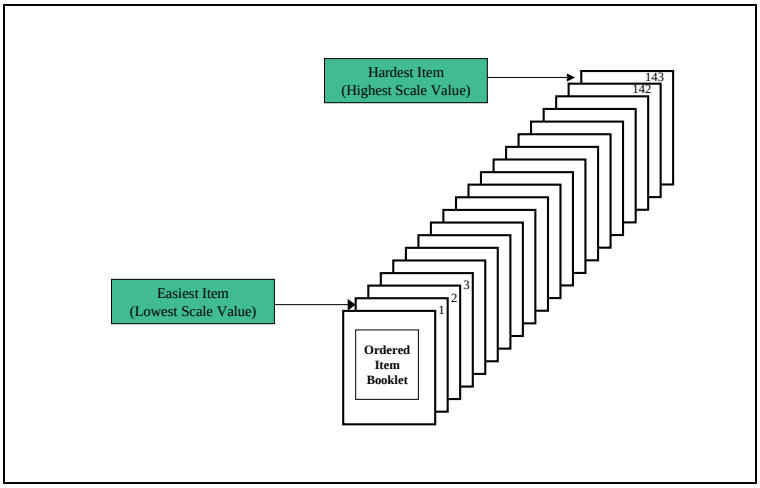


Figure 13. Illustration of how items are ordered by difficulty in the Ordered Item Book (OIB).

Panelists were told that they would be performing a bookmark task in Round 1, but that they would use domain scores to recommend cut scores in Round 2 and in subsequent rounds. It is important that panelists not feel "surprised" in Round 2 by the domain score feedback and the new tasks they are asked to perform in recommending cut scores. Slides used to illustrate the tasks that panelists would perform are presented in later sections. (See Figures 19, 23, and 24.)

In explaining how panelists would be prepared to make their cut score recommendations, the purpose of presentations and tasks such as the general orientation, taking the examination, bookmark placements, rater location feedback, domain score feedback, and so fourth, is explained in terms of the process themes discussed in the general orientation.

Panelists were told about the different types of items in NAEP and shown how to interpret information in their materials. It was important for panelists to understand how to interpret information in their materials before they were instructed in the tasks that require the materials so their attention would be focused more on understanding the tasks than on understanding the materials. The presentation used specific items as examples and referred panelists to pages in their OIB and items on their Primary Item Map.

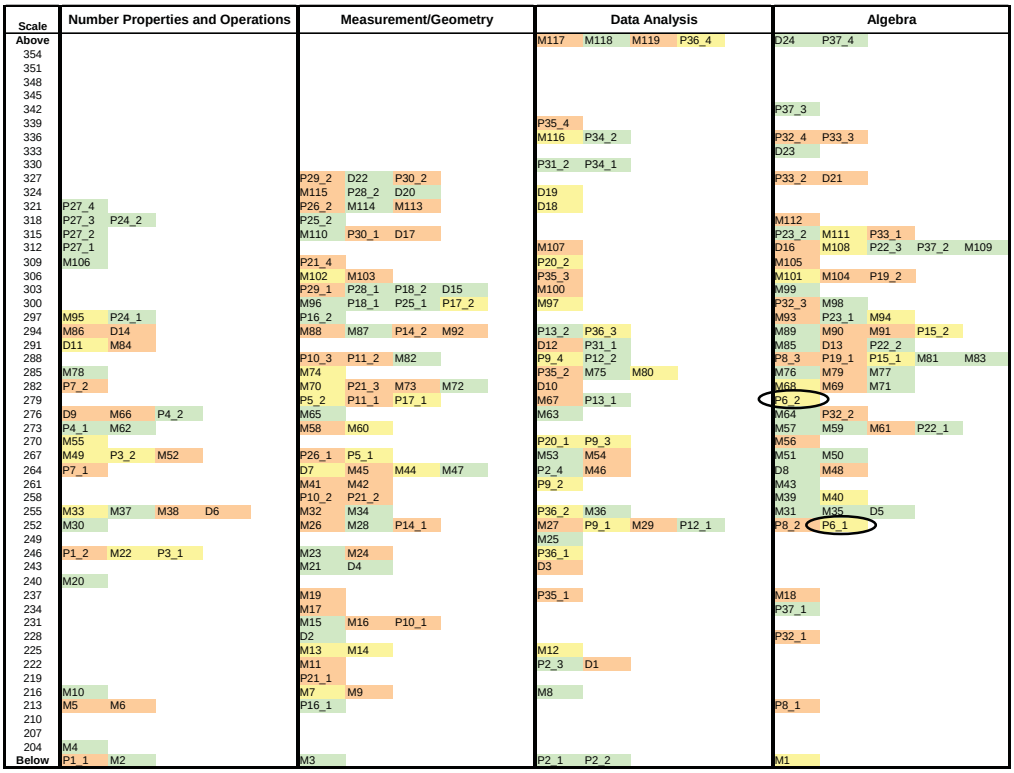


Figure 14. Primary Item Map on which score levels for polytomously-scored item P6 (P6_1 and P6_2) are marked by circles.

Figure 14 shows sections of an actual Primary Item Map that was used to illustrate key information about materials. Items were represented on item maps by a handle consisting of a character followed by a number. The character indicated item type (P=polytomously-scored, D=dichotomously-scored constructed response, and M=multiple choice). The

number indicated the easiness rank of the item (1=easiest within item type). Handles for polytomously-scored items include an underscore '_' followed by the score level. Polytomously-scored items were ordered by the difficulty of their last score level.

Circles on the map in Figure 14 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.

Figure 15 shows the location of the score points of item P6 in the Group A and Group B OIB and indicates the information contained in the OIB for each step. Score points of polytomously-scored items were treated as separate "items" in the OIB, just as they were on the item map. In the Group A OIB, the first score point of item P6 was located on page 33 and the second score point was located on page 71. There were actually at least two pages for each score point of a constructed response item in the OIB—one showing the item and one showing the scoring rubric—but the page numbers in the OIB increase only when the item or score level changes.

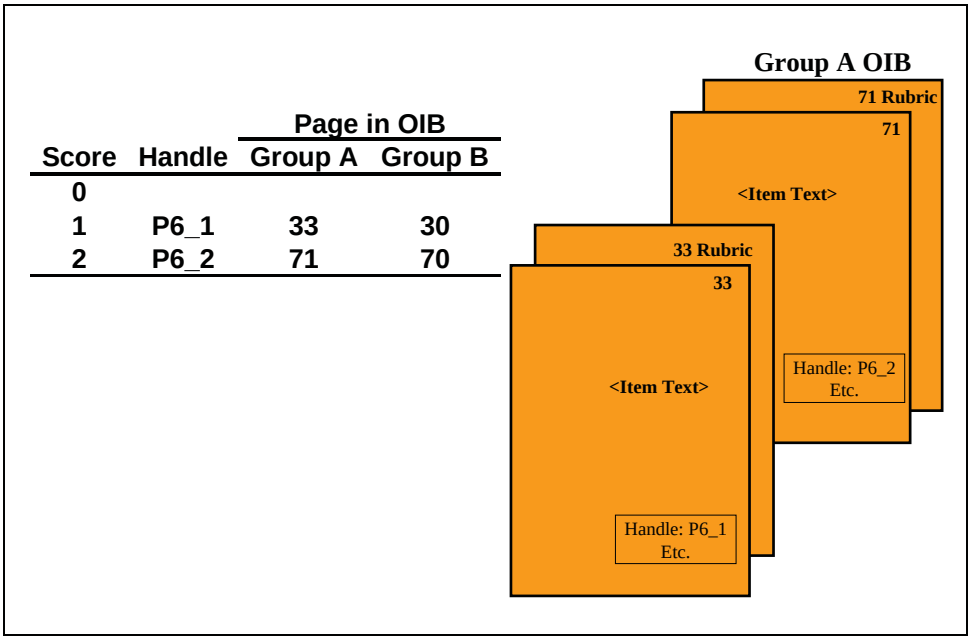


Figure 15. Information showing location and materials for Item P6 in OIB.

On the OIB page that contained the item's text, there was a framed box, as shown in Figure 15. The box contained the item's or score-point's:

- handle,
- scale value (the scale value at which a student has a 0.67 probability of earning the score point or correctly answering the item),
- map value (the midpoint of the interval containing the item on the item map),
- subscale classification in the 2005 framework, and
- complexity classification in the 2005 framework.

The information-box was brought to panelists' attention and the information was explained.

Besides item types, other aspects of the NAEP design explained to panelists included how the test items were organized into blocks, which blocks were assigned to which group, which ancillary materials were needed for each block, and how to tell if calculator use was permitted for an item. Ancillary test materials include shapes, spinners, rulers, protractors, and calculators. The ancillaries needed for students to answer items were available to the panelists at all times.

Understanding the Assessment and Student Achievement

To set cut scores on an assessment, one must have a good understanding of the assessment and of student achievement on the assessment—the knowledge, skills, and abilities (KSAs) the assessment requires students to demonstrate in order to earn successively higher scores on the test.

The first step in helping panelists acquire this understanding was a presentation on the test framework. Panelists had been instructed to read the Framework prior to the meeting. To reinforce this learning, the framework presentation provided a clear, comprehensive account of the content and organization of the mathematics framework. The framework presentation lasted forty-five minutes and was given by the primary content facilitator.

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.

> 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.
>
> 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.

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.

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 included the Constructed Response Ordered Item Book (CROIB) and a Note-template. The CROIB contained all the polytomously-scored items in a group's 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 16 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.

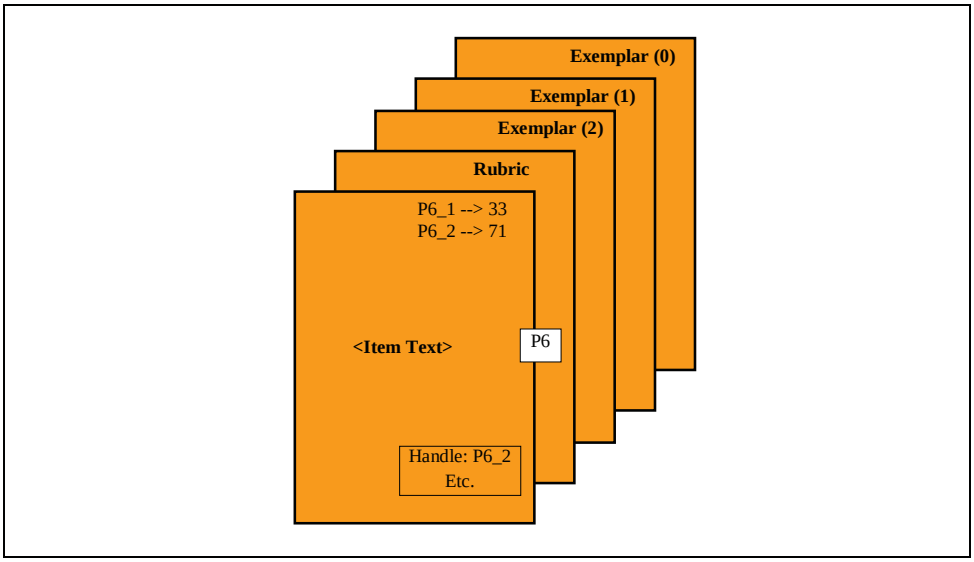


Figure 16. Slide illustrating contents of the Constructed Response OIB.

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 11"x17" pages with outlines for accommodating six stickies per page.

Within each stickie-outline was an item handle and OIB page number identifying the stickie 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 17 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 saw score points of polytomously-scored items relative to the difficulty of all other items in their pool in KSA Activity 3, they could 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.

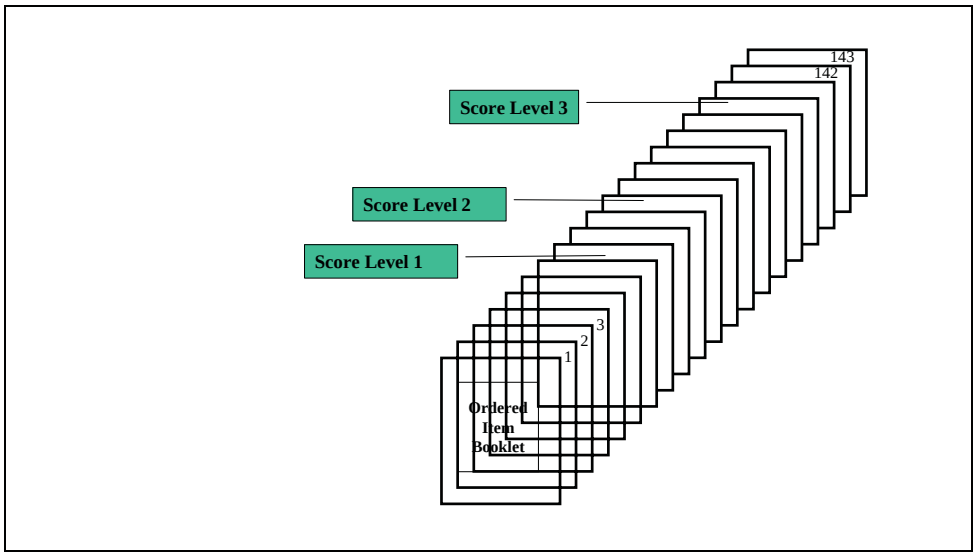


Figure 17. Score levels of a polytomously-scored item are treated as separate items and appear at different places in the OIB.

Panelists checked items off on their Primary Item Map as they progressed through the OIB. Figure 18 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.

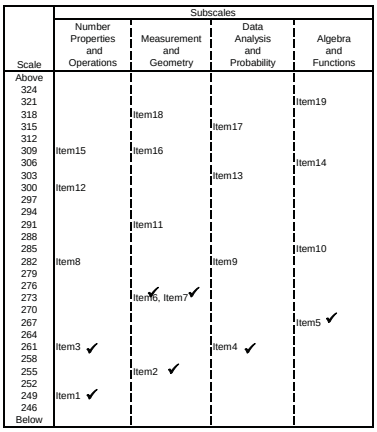


Figure 18. Simplified item map illustrating results of item check-off procedure as a panelist progresses through OIB up through Item 7 in KSA Activity 3.

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 was 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 prepared panelists to understand the continuum of increasing knowledge, skills, and abilities represented by the achievement level descriptions—Basic, Proficient, and Advanced.

Understanding the Achievement Level Descriptions

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 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 in their Round 1 bookmark placements.

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, and

- The panelist's job is to decide what the lower *borderline* of that range should be.

Panelists were told to think of the lower borderline in terms of a student who was "just qualified" to be in the achievement level and to decide for themselves what "just qualified" meant 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 was initially described to panelists as a process of going through the OIB, beginning with the easiest item, until they came to an item that they judged to be too difficult for mastery by the borderline student. Mastery was defined as having at least a 0.67 probability of answering the item correctly. The bookmark was placed on the item immediately preceding the "too difficult" item. Figure 19 illustrates a bookmark placement.

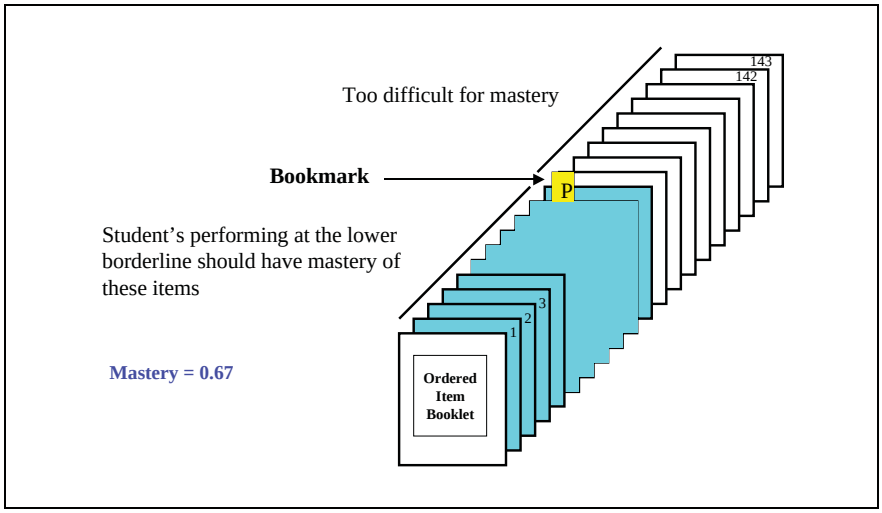


Figure 19. Bookmark placement task simplified.

Once panelists had this basic idea, the instructor explained to panelists that it was possible for them to be unsure of where to place their bookmarks because: 1) they may not have felt there was a noticeable or meaningful difference between adjacent items in terms of difficulty, and 2) they may have felt that a few items in the OIB were out of order with their own expectations of relative difficulty.

The initial description of the process was then supplemented with the instruction to go further, beyond the first item they judge to be too difficult, to see if there were any later items that they felt the borderline student should have mastery of. This instruction was represented to panelists visually by showing a "range of uncertainty" in a slide-depiction of the OIB. All items below this range were "sure mastery" items. All items above this range

were “sure non-mastery” items. Figure 20 shows a slide that was used to illustrate this concept for panelists.

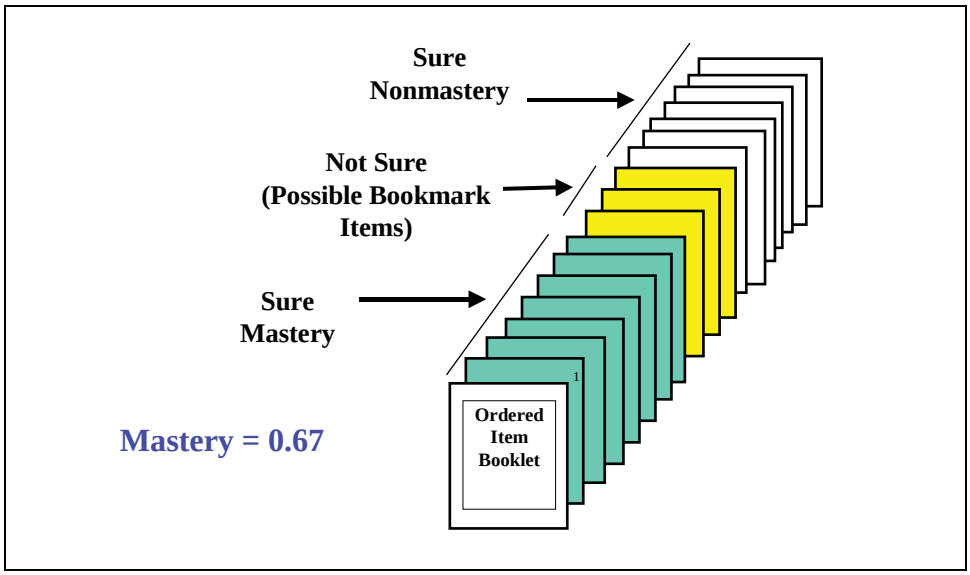


Figure 20. Slide illustrating range of uncertainty in bookmark placements.

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.

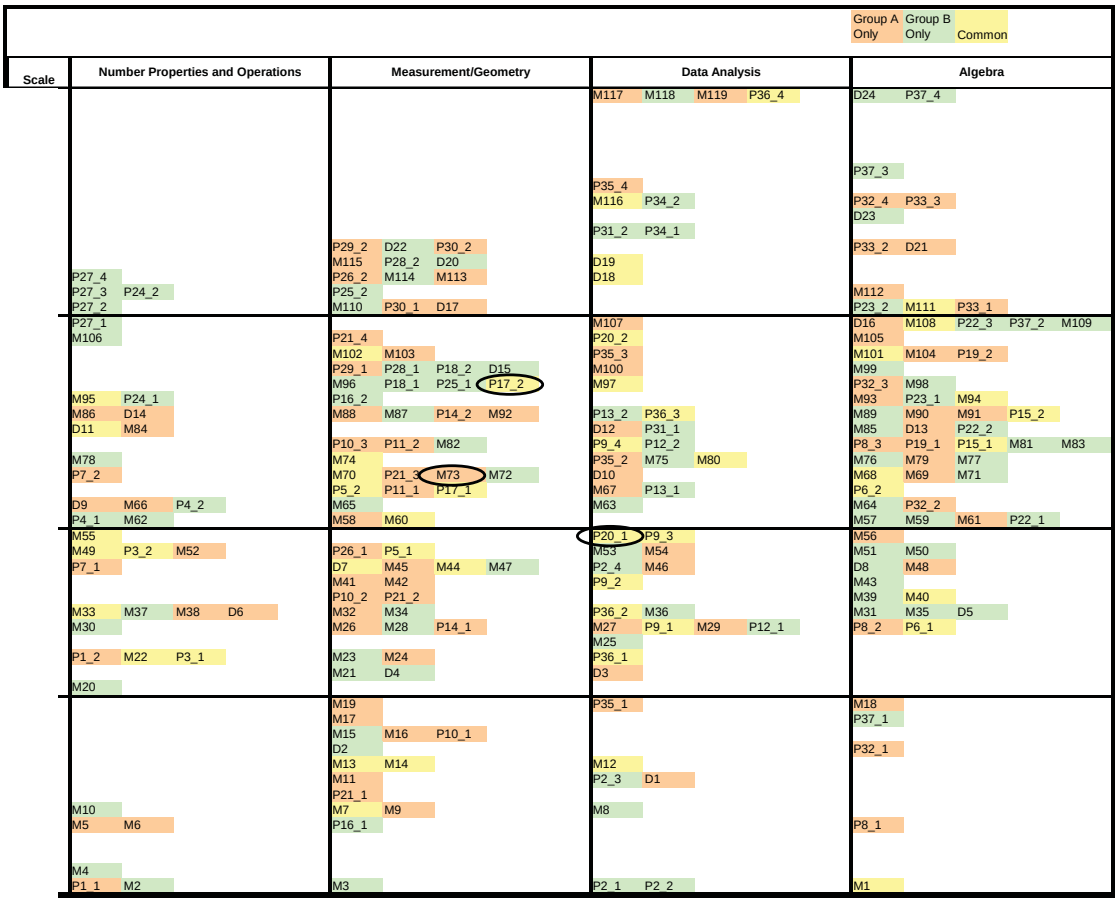
Panelists recorded the page number of their bookmark placements on a special form designated for this purpose and circled the handle of their bookmarked item on their Primary Item Map. Page numbers were entered into an interactive computer program that returned the scale value of the item on the bookmarked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

Feedback After Round 1

Feedback after Round 1 consisted of: 1) median cut scores, 2) high and low cut scores, 3) rater-location, and 4) 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 21 shows how the median cut scores and a panelist’s 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.



21. Primary Item Map showing Round 1 median cut scores (horizontal lines) and the location of Panelist X'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. (See Domain Development Activities section of this report.)

Expected percent correct curves based on subscales were shown to illustrate that the subscales were not as widely separated in difficulty as desired for purposes of defining and differentiating achievement levels.

Expected percent correct curves based on score domains defined within each subscale were shown to illustrate that teacher and score domains had the desired attributes and to show panelists where the expected domain scores in the Percent Correct Table (PCT) came from. Figure 22 shows the curves for score domains in the Data Analysis subscale. Vertical lines in this plot represent the Round 1 cut scores. The dashed, horizontal line represents a 67% criterion for mastery of the domains.

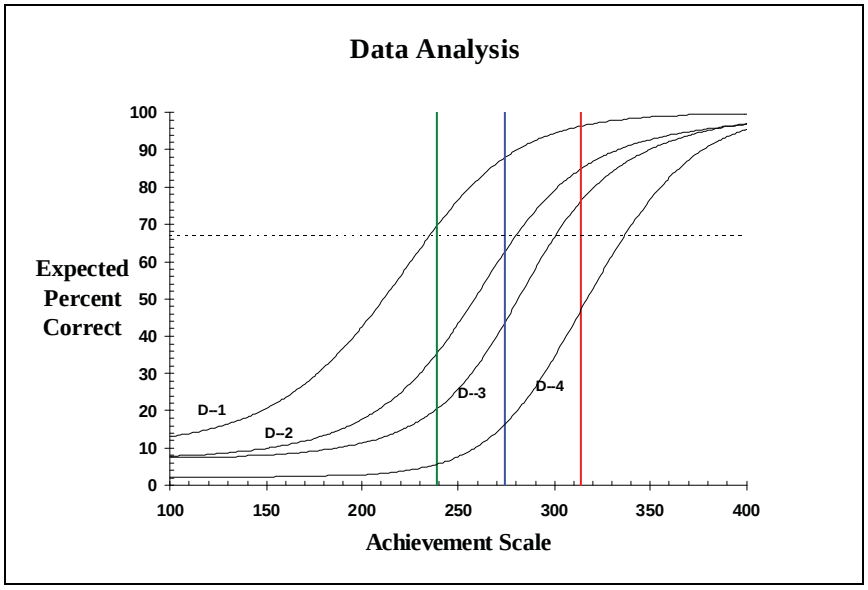


Figure 22. Percent correct curves for score domains in Data Analysis subscale, with vertical lines showing location of Round 1 cut scores and a horizontal line representing a 67% criterion for mastery.

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 23. 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.

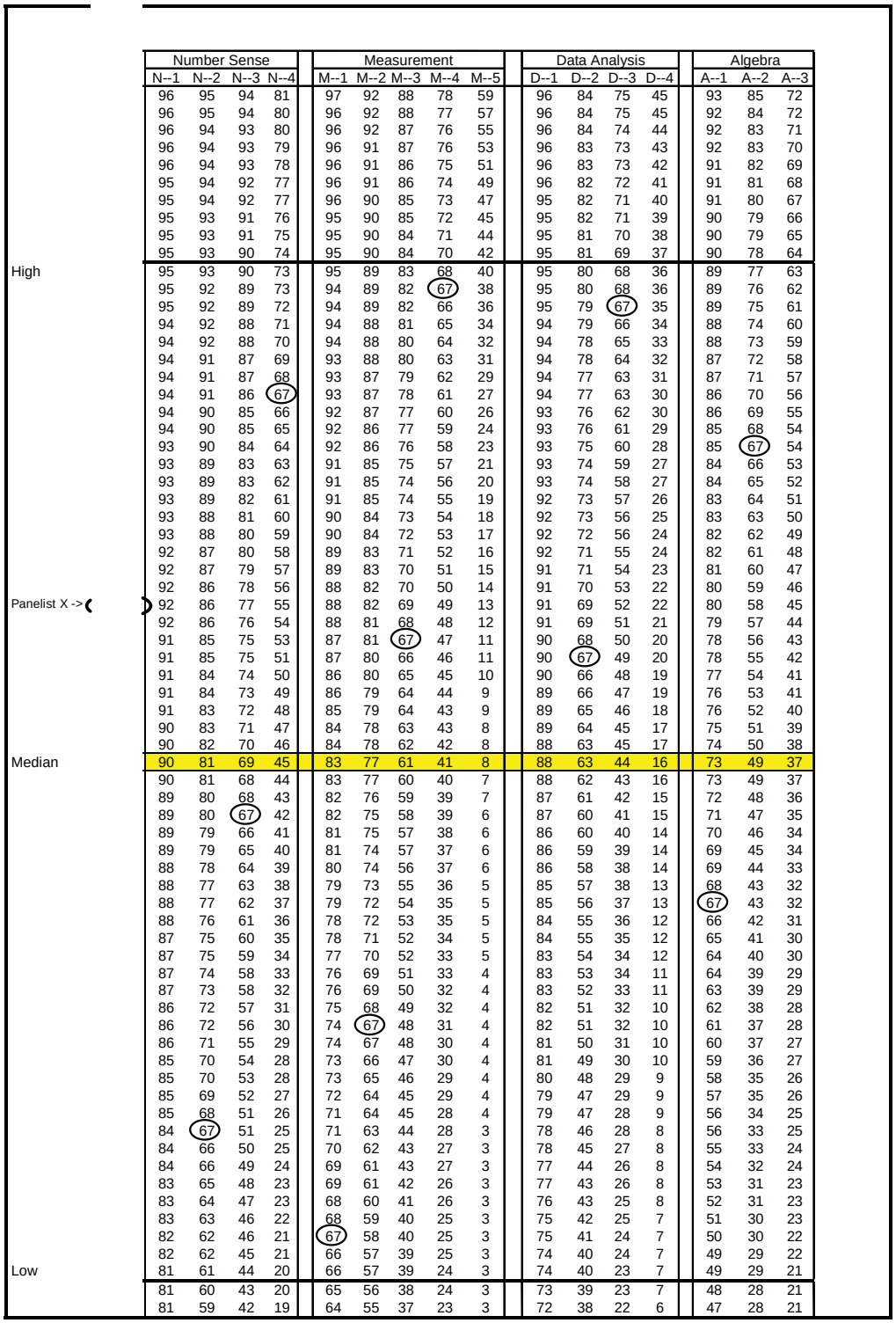
The highest, lowest, and closest-to-67% domain scores for the Proficient cut score in the PCT were circled (see Figure 23) to draw panelists' attention to the fact that in one of their Domain Tasks, they would be asked to make the "higher/OK/lower" judgment for each domain score in the table.

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

After panelists were aware 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 the 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 24 shows the DSC for the Proficient Achievement Level with the location of Panelist X 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.



	Number Sense				Measurement					Data Analysis				Algebra		
	N--1	N--2	N--3	N--4	M--1	M--2	M--3	M--4	M--5	D--1	D--2	D--3	D--4	A--1	A--2	A--3
	96	95	94	81	97	92	88	78	59	96	84	75	45	93	85	72
	96	95	94	80	96	92	88	77	57	96	84	75	45	92	84	72
	96	94	93	80	96	92	87	76	55	96	84	74	44	92	83	71
	96	94	93	79	96	91	87	76	53	96	83	73	43	92	83	70
	96	94	93	78	96	91	86	75	51	96	83	73	42	91	82	69
	95	94	92	77	96	91	86	74	49	96	82	72	41	91	81	68
	95	94	92	77	96	90	85	73	47	95	82	71	40	91	80	67
	95	93	91	76	95	90	85	72	45	95	82	71	39	90	79	66
	95	93	91	75	95	90	84	71	44	95	81	70	38	90	79	65
	95	93	90	74	95	90	84	70	42	95	81	69	37	90	78	64
High	95	93	90	73	95	89	83	68	40	95	80	68	36	89	77	63
	95	92	89	73	94	89	82	67	38	95	80	68	36	89	76	62
	95	92	89	72	94	89	82	66	36	95	79	67	35	89	75	61
	94	92	88	71	94	88	81	65	34	94	79	66	34	88	74	60
	94	92	88	70	94	88	80	64	32	94	78	65	33	88	73	59
	94	91	87	69	93	88	80	63	31	94	78	64	32	87	72	58
	94	91	87	68	93	87	79	62	29	94	77	63	31	87	71	57
	94	91	86	67	93	87	78	61	27	94	77	63	30	86	70	56
	94	90	85	66	92	87	77	60	26	93	76	62	30	86	69	55
	94	90	85	65	92	86	77	59	24	93	76	61	29	85	68	54
	93	90	84	64	92	86	76	58	23	93	75	60	28	85	67	54
	93	89	83	63	91	85	75	57	21	93	74	59	27	84	66	53
	93	89	83	62	91	85	74	56	20	93	74	58	27	84	65	52
	93	89	82	61	91	85	74	55	19	92	73	57	26	83	64	51
	93	88	81	60	90	84	73	54	18	92	73	56	25	83	63	50
	93	88	80	59	90	84	72	53	17	92	72	56	24	82	62	49
	92	87	80	58	89	83	71	52	16	92	71	55	24	82	61	48
	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	92	86	78	56	88	82	70	50	14	91	70	53	22	80	59	46
Panelist X ->	92	86	77	55	88	82	69	49	13	91	69	52	22	80	58	45
	92	86	76	54	88	81	68	48	12	91	69	51	21	79	57	44
	91	85	75	53	87	81	67	47	11	90	68	50	20	78	56	43
	91	85	75	51	87	80	66	46	11	90	67	49	20	78	55	42
	91	84	74	50	86	80	65	45	10	90	66	48	19	77	54	41
	91	84	73	49	86	79	64	44	9	89	66	47	19	76	53	41
	91	83	72	48	85	79	64	43	9	89	65	46	18	76	52	40
	90	83	71	47	84	78	63	43	8	89	64	45	17	75	51	39
	90	82	70	46	84	78	62	42	8	88	63	45	17	74	50	38
Median	90	81	69	45	83	77	61	41	8	88	63	44	16	73	49	37
	90	81	68	44	83	77	60	40	7	88	62	43	16	73	49	37
	89	80	68	43	82	76	59	39	7	87	61	42	15	72	48	36
	89	80	67	42	82	75	58	39	6	87	60	41	15	71	47	35
	89	79	66	41	81	75	57	38	6	86	60	40	14	70	46	34
	89	79	65	40	81	74	57	37	6	86	59	39	14	69	45	34
	88	78	64	39	80	74	56	37	6	86	58	38	14	69	44	33
	88	77	63	38	79	73	55	36	5	85	57	38	13	68	43	32
	88	77	62	37	79	72	54	35	5	85	56	37	13	67	43	32
	88	76	61	36	78	72	53	35	5	84	55	36	12	66	42	31
	87	75	60	35	78	71	52	34	5	84	55	35	12	65	41	30
	87	75	59	34	77	70	52	33	5	83	54	34	12	64	40	30
	87	74	58	33	76	69	51	33	4	83	53	34	11	64	39	29
	87	73	58	32	76	69	50	32	4	83	52	33	11	63	39	29
	86	72	57	31	75	68	49	32	4	82	51	32	10	62	38	28
	86	72	56	30	74	67	48	31	4	82	51	32	10	61	37	28
	86	71	55	29	74	67	48	30	4	81	50	31	10	60	37	27
	85	70	54	28	73	66	47	30	4	81	49	30	10	59	36	27
	85	70	53	28	73	65	46	29	4	80	48	29	9	58	35	26
	85	69	52	27	72	64	45	29	4	79	47	29	9	57	35	26
	85	68	51	26	71	64	45	28	4	79	47	28	9	56	34	25
	84	67	51	25	71	63	44	28	3	78	46	28	8	56	33	25
	84	66	50	25	70	62	43	27	3	78	45	27	8	55	33	24
	84	66	49	24	69	61	43	27	3	77	44	26	8	54	32	24
	83	65	48	23	69	61	42	26	3	77	43	26	8	53	31	23
	83	64	47	23	68	60	41	26	3	76	43	25	8	52	31	23
	83	63	46	22	68	59	40	25	3	75	42	25	7	51	30	23
	82	62	46	21	67	58	40	25	3	75	41	24	7	50	30	22
	82	62	45	21	66	57	39	25	3	74	40	24	7	49	29	22
Low	81	61	44	20	66	57	39	24	3	74	40	23	7	49	29	21
	81	60	43	20	65	56	38	24	3	73	39	23	7	48	28	21
	81	59	42	19	64	55	37	23	3	72	38	22	6	47	28	21

Figure 24. Domain Score Chart showing Round 1 results and location of Panelist X for the 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. The corresponding scale values had been written by staff at the conclusion of Round 1, and the form was returned to panelists at the beginning of the Round 1 feedback.

By encircling 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 benefits help panelists understand their essential task of recommending cut scores.

The principal materials used in Domain Task 1 were: 1) a Domain Ordered Item Book, or DOIB, 2) Domain Item Maps, and 3) the Domain Task 1 form. The DOIB contained the items in a panelist's pool in order of difficulty, within teacher domain. Teacher domains were presented in the DOIB in the order they were represented by columns from left to right on the Domain Item Map. This was in order of their difficulty within score domain, with score domains ordered by difficulty from left to right on the Domain Item Map.

Figure 25 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 25. 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 7 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).

Domain Item Maps (DIMs) were also used in the domain tasks of Round 2. Panelists were given one DIM for each subscale. Figure 26 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 DIM. 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.

Figure 26. Domain Item Map for Data Analysis and Probability subscale.

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 DIM 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 were 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 were 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 itself was 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. The lower panel of Figure 6 shows the plot for Measurement/Geometry. Figure 22 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 made judgments about whether the domain scores associated with the Round 1 median cut score should higher, lower, or were OK as a standard of lower borderline performance for a given achievement level. Figure 27 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 22, 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 27. Domain Task 2 form for the 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 Round 1 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. Checks in both the "higher" and "lower" columns of the Domain Task 2 form were a simple example. 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, based on likely differences in reliability. For similar reasons, panelists were told to give less importance to domains with very high or very low scores and to focus on scores in the steep part of the percent correct curve (near 67%).

Panelists were 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 were not comfortable with their understanding of the domain scores, their Round 2 cut score recommendation could be the scale value derived from their Round 1 bookmark placement, or close to it.

In making 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.

Feedback After Round 2

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 in drawing 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.

- Domain Score Chart. The DSC was marked as shown in Figure 24 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.

- Ordered Item Book. 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 in 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 scores).

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 to 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 recommendations on their Domain Score Chart, Ordered Item Booklet, Primary Item Map, and on the Cut Score Recommendation form. For recording their cut score recommendations 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.

Feedback after Round 3

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 were the percent of students in each achievement level and the percent at or above each achievement level. The percent of students below basic was also included. The consequences data were based on the Round 3 median cut scores. Figure 28 shows the consequences data that were given to panelists in Round 4. The data were presented in this format. 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.

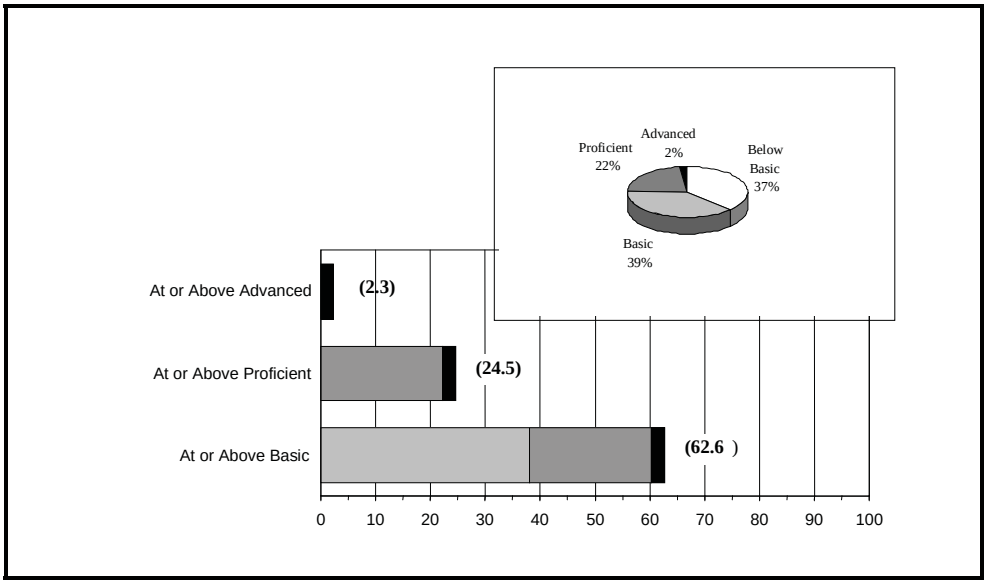


Figure 28. Consequences data presented to panelists in Round 4.

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 was 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.

Feedback After Round 4

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 recommendations 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. This was presented in the format shown in Figure 28. 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

The purpose of the consequences questionnaire was to provide NAGB with information about panelists' reactions to the final consequences data. Using the consequences feedback they were given, panelists wrote down the percent at or above each achievement level on their consequences questionnaire and then proceeded to answer questions about their reaction to this information. The questionnaire asked panelists if they would want to make changes to any of the cut scores after learning the consequences of their cut scores. Panelists could recommend a different cut score to represent each achievement level for any or all three cut scores. A copy of the consequences questionnaire is included in Appendix F.

Ratings of Exemplar Items

The purpose of the exemplar item rating task was to provide NAGB with information concerning the suitability of items for illustrating what students in the achievement levels know and can do. Potential exemplars were drawn from blocks of the assessment that were selected for eventual release to the public. These were Blocks 3, 4, and 12. The panelists had spent many hours working with the Achievement Level Descriptions, translating their meaning into cut scores. They were in a good position to provide NAGB with this input.

Items were selected as potential exemplars for an achievement level if a student at the midpoint (Basic and Proficient) or median (Advanced) of the level had at least a 0.67 probability of correctly answering the item (or reaching the given score level of a partial credit item) and the item or score point had not met this criterion for a lower level. The division of items by a 0.67 probability conditional on a scale value was consistent with the order of items on panelists' item maps and in their Ordered Item Book. The 0.67 probability value produced reasonable-sized pools of items for potential use as exemplars.

The conditioning of the 0.67 probability on the midpoint of an achievement level, as opposed to the median or average conditional probability for students in the level, is an entirely criterion-referenced criterion. The probability had to be conditioned on the median at the Advanced level because there is no upper bound to this level. The median was computed from the field test student distribution.

Figure 29 shows the Exemplar Item Rating form panelists were given for rating items associated with the Basic achievement level. The form listed the items in the order they appeared in the Ordered Item Book, identified the items by handle and the OIB page number where they could be found, and showed the probability of correct response or of attaining the partial credit score level indicated in the item handle or higher. Since Block 12 was not in the Group B item pool, Group B was given a special handout for these items and the page number of Block 12 items in this handout was indicated on the Exemplar Item Rating form.

Item	OIB Page #		Probability of Success	Rating as Exemplar			If Do Not Use, please explain:
	Group A	Group B		Very Good	OK	Do Not Use	
M1	2	2	0.83				
M7	6	8	0.94				
M9	7	H-1	0.93				
P21_1	8	H-2	0.94				
M11	9	H-3	0.92				
M12	11	12	0.83				
M13	12	13	0.89				
M14	13	14	0.95				
M17	17	H-4	0.86				
M22	23	21	0.80				
P3_1	24	22	0.80				
P36_1	25	24	0.75				
P9_1	29	26	0.76				
P6_1	33	30	0.71				
P36_2	35	31	0.70				
M33	36	33	0.70				
P21_2	40	H-5	0.69				

Figure 29. Exemplar Item Rating form for the Basic achievement level.

Panelists were instructed to discuss each potential exemplar item with their table group, yet provide independent ratings on the basis of whether the knowledge, skills, or abilities required by the item seemed appropriately matched to the achievement level. They were instructed to consult their Achievement Level Descriptions in this task.

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. Most responses were collected on Likert scales, but several responses were narratives that addressed specific aspects of the process. 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.

The process evaluation questionnaires are presented in their entirety in Appendix G. Along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response. Panelists' responses to open-ended questions are also presented.

General Evaluation

The Mapmark ALS process compared well with methods ACT used in past standard setting work for NAGB. Key evaluation questions on the last process evaluation questionnaire addressed panelists' confidence in their cut score recommendations, their perceptions of the effectiveness of the process, whether the process afforded them the opportunity to use their best judgment, clarity of instructions, understanding of tasks, and the amount of time allocated for tasks.

Table 11 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 11: 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.

Amount of Time Allocated for Tasks

The adequacy of time allocated for tasks was a more important issue than usual in this ALS process because ACT had assured NAGB that Mapmark could be conducted in four days without compromising the integrity of the process. In the Pilot Study and the Grade 8 Study, Mapmark was conducted in five days. Detail concerning the amount of time allocated for tasks is presented in Table 12.

In this table and most others in this section, the questions summarized by their average rating may be located in Appendix G by the questionnaire number (1 to 6) and their sequence number within the questionnaire. One can refer to the appendix to see the frequency of responses per rating scale category. In the case of time allocation, it can be seen that responses generally clustered close to 3.0 when the average was close to 3.0.

That is, there were no cases where an average close to 3.0 corresponded to significant divisions among panelists in whether the time was too long or too short.

Detail in Table 12 indicates that time was sufficient for all tasks. There were only two averages less than 3.0 when rounded to the nearest 0.1. The averages were for the whole group KSA review and table discussion of the OIB. Bookmark facilitators noted that it is not uncommon for some panelists to want to spend more time on item KSA tasks no matter how much time is allocated.

Table 12: Amount of Time Allocated for Activities
(5=Far too long; 3=About right; 1=Far too short)

Day	Question Location	Activity	Average Rating
	1-3	The General Orientation to the NAEP Program	3.35
	1-8	The General Introduction to the NAEP achievement level setting process	3.42
1	1-14	The Mapmark method orientation	3.03
	1-19	The Framework presentation	3.20
	1-22	The whole group KSA review	2.61
	2-1	The table group KSA review	2.97
	2-4	The independent OIB review	2.87
2	2-10	The table discussion of the OIB	2.90
	2-16	The ALD presentation	3.35
	2-28	Placing the bookmarks	2.97
	3-5	Work with domains	3.03
3	3-21	Round 2 cut score recommendations	3.39
	4-5	The rater group discussion	3.13
	5-7	Discussing the consequences data	3.20
4	6-4	The Consequences Questionnaire	3.30
	6-17	The Exemplar Item Rating Task	3.13
All	6-11	The amount of time I had to complete the tasks I was to accomplish during each round	3.03

Clarity of Instructions and Presentations

Table 13 shows average ratings on questions pertaining to clarity of instructions. The ratings are generally high. The whole group KSA review was the only task for which instructions received a rating below 4.0 (3.84). This task was one of the most complex and was not actually performed by panelists on their own.

Table 14 shows average ratings pertaining to clarity of presentations on certain topics addressed the first day of the ALS meeting. The presentations were generally clear. On only two items was the average rating for clarity less than 4.0. One involved the overview of the Mapmark method and the other was for the explanation of how an item map was constructed. The explanation of item map construction was part of the Mapmark overview.

Table 13: Clarity of Instructions by Task

The Instructions on (what/how I/we was/were to do in/for the…)
(5=Absolutely clear; 3=Somewhat clear; 1=Not at all clear)

Round	Question Location	Activity	Average Rating
1	1-23	The whole group KSA review	3.84
1	2-5	The independent OIB review	4.35
1	2-11	The table discussion of the OIB	4.17
1	2-29	Placing the bookmarks	4.45
2	3-6	Domain Score Tasks	4.00
2	3-22	Recommending Round 2 cut scores	4.16
3	4-6	The rater group discussion	4.16
3	4-18	Recommending Round 3 cut scores	4.32
4	5-6	Using the consequences data during Round 4	4.47
4	5-17	Recommending Round 4 cut scores	4.53
Post	6-6	Completing the consequences questionnaire	4.52
All	6-8	What I was to do during each round	4.17

Table 14: Clarity of Topic Presentation

The ____ was
(5=Absolutely clear; 3=Somewhat clear; 1=Not at all clear)

Round	Question Location	Activity	Average Rating
Pre	1-4	Explanation of the development of the NAEP in general	4.26
Pre	1-5	Explanation of the development of the mathematics NAEP	4.37
Pre	1-6	Major organizations involved and the roles of each	4.23
Pre	1-15	Overview of the method to be followed in this meeting	3.79
Pre	1-16	Explanation of how an item map is constructed	3.77
Pre	1-18	Explanation of the information in my Ordered Item Booklet	4.07
1	1-20	Presentation of the Mathematics Framework	4.00

Understanding of Concepts, Tasks, Feedback

Understanding of concepts and tasks depends on the clarity of presentations and instructions, which the previous section shows was generally good. It can be seen in Tables 15 and 16 that panelists had a good understanding of concepts and tasks in the ALS process. In particular, understanding of concepts unique to the Mapmark process, such as the concept of domain scores, how to use item maps, and how to use the domain score chart was high, as indicated by average ratings above 4.0 in Table 15. Understanding of instructions for recommending cut scores and rating exemplar items was high (Table 16).

Table 15: 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 16: 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
Post	6-19	how to perform the Exemplar Item Rating Task	4.34

Panelists' had good understanding of the feedback they were given. As shown in Table 17, 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 17: 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

Understanding the Achievement Level Descriptions and Borderline Performance

Panelists' understanding of the achievement level descriptions was sufficiently high at Round 1 and continued to build over rounds. Understanding was assessed for each achievement level separately by round as shown in Table 18. In past standard setting work for NAGB, average ratings in Round 1 have traditionally been between 3.0 and 4.0, and built over rounds, with averages being above 4.0 by Round 2. This is seen in Table 18. The ALDs did not differ by level in how well they were understood by panelists.

Table 18: Understanding Achievement Level Descriptions (ALDs)
At the time I provided the/my Round __ bookmark placements/cut score recommendations my understanding of the ___ level description was…
(5=Totally Adequate; 3=Somewhat Adequate; 1=Totally Inadequate)

	Round			
Level	1	2	3	4
Basic	3.90	4.35	4.45	4.50
Proficient	3.77	4.39	4.42	4.47
Advanced	3.73	4.29	4.45	4.47

Panelists' formation of the concept of borderline performance showed the same pattern, as shown in Table 19, although the question differed across rounds. The Round 1 question assessed only panelists' comfort with using the concept of borderline performance to make their bookmark placements. For subsequent rounds, panelists were asked, for each level, how well formed their concept of the lower borderline was at the time they made their cut score recommendations.

The pattern of responses in Table 19 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.

Unlike previous standard settings, however, panelists' concept of borderline performance appeared to be equal to their understanding of the achievement level descriptions. This can be seen by comparing the averages across Tables 18 and 19. In previous standard settings using a modified-Angoff-based methodology, average responses to the borderline concept formation were slightly, but consistently lower than average responses to the question concerning understanding of achievement level descriptions (Loomis & Hanick, 2000).

Table 19: 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 18 and 19, the responses of panelists to the question concerning the difference between borderline performance and typical performance, summarized by round in Table 17, 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 22. Illustrations of how performance changes over the range of an achievement level focuses panelist's attention on the concept of borderline performance.

Table 20 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 20: 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 21, 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, 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 21
Comfort and Confidence

I think I will be/I was comfortable …
(5=Totally agree; 3=Somewhat 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 22 show that panelists found the KSA activities 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 22 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.

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.

Table 22
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

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 23, panelists tended to disagree with the statement that they felt pressure to recommend cut scores that were close to those of other panelists. 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 23
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

Domain coherence is the extent to which items within a domain seem to belong together or measure the same thing. Domain coherence was one of the goals in ACT's domain development process.

ACT suggested two ways of assessing domain coherence in this project: 1) agreement among teachers' independent item classifications, and 2) a high percentage of "yes" responses in Domain Task 1. An intermediate set of domains was used to assess agreement among teachers' item classifications. The intermediate set of domains was similar to the final set used in the ALS meeting and the level of agreement in teachers' item classifications was good. But the only assessment of domain coherence based directly on the domains used in the ALS meeting was the percentage of "yes" responses in Domain Task 1 in the Pilot Study and in the ALS meeting itself.

Figure 30 shows the percentage of "yes" responses in Domain Task 1 in the ALS meeting, Pilot Study, the other studies conducted in this project. Results from other studies are presented for purposes of comparison. The scale in Figure 30 goes down to 50% because this is assumed to be a minimum acceptable percentage. At least 50% of responses to the Domain Task 1 question should be "yes." This is roughly equivalent to saying that at least

half of the items classified into a domain should seem to panelists to be like other items in their domain.

The percentage of "yes" responses is 93% in the ALS meeting and 96% in the Pilot Study. These percentages are lower than the percentages for the Grade 8 domains (Grade 8 Study and field trials), but are still reasonably high. Content experts advising ACT in the project predicted that Grade 12 domains would be less distinct and coherent than Grade 8 domains because the sequence of instruction is not as uniform and closely matched to the difficulty of content at the post-secondary level. The percentage of "yes" responses in the ALS meeting is only slightly lower than the percentage in the Pilot Study. This difference, if reliable, may be due to the fact that there was slightly less time allowed for this task in the ALS meeting. There was also evidence from other tasks that ALS panelists were generally more critical in their judgments compared to Pilot Study panelists. Perhaps being part of an operational ALS meeting, as opposed to a Pilot Study, heightens panelists sensitivity to content and other issues in their judgments.

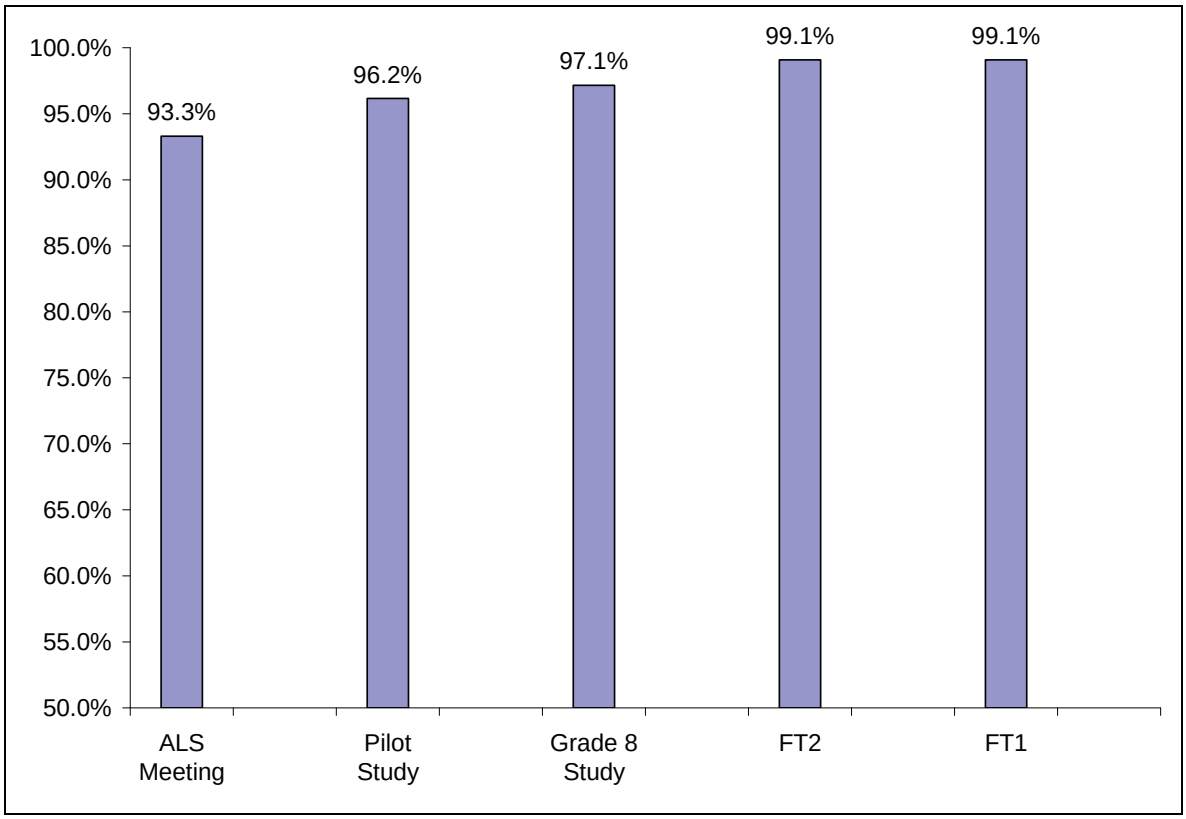


Figure 30. Percentage of "yes" responses in Domain Task 1 in the ALS meeting, Pilot Study, and other studies conducted in the project.

Table 24 presents detail from Domain Task 1 in the ALS meeting. By individual panelist, the percentage of "yes" responses ranges from 62% (a general public representative) to 100% (for two teachers). There were only five panelists who responded "yes" less than 90% of the time. These results indicate that there was no panelist for whom the domains were not coherent.

Table 24
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

Group	Table	Panelist ID	Panelist Type	Percentage "Yes"
A	1		GP	76%
			NT	84%
			TR	96%
			TR	94%
			TR	89%
	2		GP	92%
			NT	97%
			TR	98%
			TR	97%
			TR	94%
	3		GP	98%
			GP	97%
			TR	98%
			TR	95%
			TR	96%
B	4		GP	90%
			NT	94%
			TR	96%
			TR	98%
			TR	100%
	5		GP	99%
			GP	62%
			NT	96%
			TR	96%
			TR	93%
	6		GP	96%
			GP	92%
			NT	81%
			TR	96%
			TR	100%
			TR	96%
			Average:	93%
			T (Teachers):	96%
			NT (Nonteacher Educators):	91%
			GP (General Public):	89%

One would expect the percentage of "yes" responses to be higher among teachers than non-teachers, and lowest for the general public representatives. Teachers have the most experience related to the task, such as thinking about what mathematics skills may be involved in solving a test item, and general public representatives have the least. By panelist type, the percentage of "yes" responses is 96% for teachers, 91% for non-teacher educators, and 89% for general public representatives.

Relationship Between Domain Task 2 and Subsequent Change in Cut Scores

Domain Task 2 was designed with the assumption that it would be useful to break the task of selecting a scale value/cut score based on domain scores into two distinct steps. Domain Task 2 is the first step. In Domain Task 2, panelists considered whether each domain score, conditional on the Round 1 median cut score, should be higher, lower, or is OK according to the achievement level description (ALD) and their concept of what level of performance on the test was minimally acceptable for a student to be in the achievement level.

When selecting a scale value to recommend as a cut score in Round 2 and subsequent rounds, panelists must perform the second step of the process. Since the ALD may specify a pattern of skill different from the progression of knowledge, skills, and abilities evident in the empirical pattern of domain scores, some domains may be checked "higher," while others may be checked "lower." Panelists must therefore internally weight the relative importance of the domains, decide which checks, e.g., "higher" or "lower" to give the most weight, and decide how much higher or lower than the Round 1 median cut score their recommended scale value/cut score should be.

If Domain Task 2 is useful and meaningful to panelists, then the median cut score after Round 1, particularly the median Round 2 cut score, should be logically related to the relative frequency of checks in the higher/OK/lower categories. For example, if a relatively large proportion of checks are in the "higher" category, then the Round 2 median cut score should be higher than the Round 1 median. The larger the proportion in the higher category, the higher the Round 2 median should be. If a large proportion of checks are in the "OK" column, then one would expect little or no movement in the cut score.

If the relationship of Round 2 to Round 4 cut scores to the Round 1 median are not consistent with expectations based on Domain Task 2, it may be that panelists were not confident in their judgments about the domain scores or that they simply decided to give more weight to the item-level information or to student performance data (Round 4) than to the domain score information.

The data in Table 25 show that the relationship of cut scores from later rounds to the Round 1 cut score was generally consistent with the pattern of checks in Domain Task 2. Data from the Pilot Study is included for comparison.

In the ALS meeting, 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 in the ALS meeting did not change very much from Round 1. At the Advanced level, "no change" in the cut score over rounds

is consistent with the fact that the percentage of checks in the "OK" category was large (70%) and the difference between the highest and lowest category percentage was small (12% versus 15%). At the Basic and Proficient levels, the one to two point increase in the cut score over rounds is consistent with the larger percentage of checks in the "higher" compared to the "lower" category (25% versus 19% for Basic; 27% versus 18% for Proficient).

Table 25
Relationship Between Domain Task 2 and Subsequent Movement in Cut Scores

		Pilot Study					ALS				
Achievement Level	Domain Task 2 Categories	Cut Score by Round 1	Percentage of Checks in Category	Cut Score by Round 2	Cut Score by Round 3	Cut Score by Round 4	Cut Score by Round 1	Percentage of Checks in Category	Cut Score by Round 2	Cut Score by Round 3	Cut Score by Round 4
BASIC	Higher		29					25	240	241	241
	OK	243	63	243	243		239	56			
	Lower		8			239		19			
PROFICIENT	Higher		49	273	273	270		27	276	275	275
	OK	266	45				274	54			
	Lower		5					18			
ADVANCED	Higher		27	316	316	316		12			
	OK	312	71				314	70	314	314	314
	Lower		2					15			

A similar, but slightly lower degree of consistency with Domain Task 2 is seen in the Pilot Study. At the Advanced level the percentage of "OK" responses is larger than in the ALS meeting, but the percentage of "higher" responses is much larger than the percentage of "lower" responses. It, therefore, makes sense that the Advanced cut score increased slightly in the Pilot Study. The largest movement in Table 25 is seen in the Proficient cut score from the Pilot Study—an increase of 7 points from Round 1 (266) to Round 2 (272). It makes sense that this was the largest movement because the percentage of "OK" responses was lowest (45%) and the difference between the percentage of "higher" and "lower" responses was highest (27% versus 2%).

The cut score movement least consistent with Domain Task 2 is that of the Basic cut score in the Pilot Study. There was a large difference in the percent of higher versus lower responses (29% versus 8%), but the cut score did not move in Rounds 2 and 3 and moved *lower* in Round 4. Results from the Pilot Study indicated that panelists understood their tasks less well in the Pilot Study, which may explain this inconsistency. The inconsistency seems minor in view of the facts that: 1) most responses (63%) were in the "OK" column (which supports no change in the cut score), and the fact that panelists have been responding primarily to new information (student performance data) when they lowered the cut score in Round 4.

The main reason for presenting the Pilot Study results in Table 25, however, is to show that domain score feedback is meaningful and useful to panelists despite the fact that it produced little change in the cut scores in the ALS meeting. Domain Task 2 results were consistent across the two meetings. Taken together, results from the two meetings indicate

that panelists were satisfied with the domain scores associated with cut scores around 240 to 243 for Basic, 273 to 276 for Proficient, and 314 to 316 for Advanced. Responses to domain scores associated with cut scores outside these ranges were consistent with moving the cut scores to within these ranges, which is what the panelists in both studies did. There was little movement of cut scores in the ALS meeting because the Round 1 cut scores were already close to, or within, these ranges.

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 25 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 26, 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 26
Cut Score Endorsements/Recommendations after Seeing Round 4 Consequences Data

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

OUTCOMES OF THE ACHIEVEMENT LEVEL SETTING PROCESS

There are three components of NAEP achievement levels: Achievement level descriptions (ALDs), cut scores, and exemplar items. The previous sections described the overall ALS process and the ALS meeting, which concern all three components. This section presents ACT's recommendations and information specific to each of the three components.

Achievement Level Descriptions

The Achievement Level Descriptions represent NAGB's attempt to "stipulate what students should know and be able to do at each grade level and content area measured by NAEP" and to "make the NAEP data more understandable to the general user, parents, policymakers, and educators alike." (NAGB, 2003). The Statement of Work for the contract awarded to ACT calls for achievement level descriptions to be "written in a clear, jargon free language suitable for a general public audience," and that the process result in "meaningful descriptions of the levels in terms of the subject area content."

In contract negotiations, NAGB elected to develop the achievement level descriptions. They were developed as described in an earlier section of this report. ACT used the ALDs in the Pilot Study and in the ALS meeting and collected data concerning whether they are likely to serve the purposes stated in the previous paragraph.

An "ALD evaluation task" was included in the Pilot Study to see if there was any need to modify the Achievement Level Descriptions for the ALS meeting. This task was performed only by the Mapmark panelists. The narrative components of the ALDs were broken down into relatively specific statements, usually by sentence. The number of statements was 6, 9, and 8 for Basic, Proficient, and Advanced, respectively. For each task, panelists were asked if they saw test items related to the statement. The forms used and results are presented in Appendix H.

The percentage of "yes" responses to the question, "Did you see test items related to this statement?" was, 99.2%, 98.4%, and 82% for Basic, Proficient, and Advanced, respectively. ACT concludes from these results that the ALDs are well-aligned with the content of the assessment. The lower percentage for Advanced may be due to a difficulty that some panelists may have understanding more abstract and technical statements and identifying items associated with such statements.

On process evaluation questions in both the Pilot Study and in the ALS meeting, panelists reported being satisfied with the ALDs. Tables 27 and 28 summarize, respectively, Pilot Study Mapmark and ALS panelists' responses to questions concerning their satisfaction with the ALDs. Satisfaction with ALDs appears to be lower in the ALS meeting than in the Pilot Study, but is still reasonably high considering satisfaction was assessed relatively early in the ALS process, soon after the ALDs were first introduced (immediately after Round 1). ACT has found that ratings of comfort and satisfaction with materials and information in the ALS process are often below 4 (on a scale of 1 to 5) early in the process, and improve as panelists' exposure to materials and information and/or tasks are repeated.

It should also be noted that panelists in the ALS meeting seemed generally more critical of content-related information presented to them. (See section on Domain Task 1.) It may be that the comparative nature of the Pilot Study, with two groups meeting simultaneously to evaluate different methods, made Pilot Study panelists less critical in general.

The last question in Table 28 was asked at the end of the ALS meeting and was not asked in the Pilot Study. An average response of 4.0 to this question also supports the use of the ALDs for NAGB's purposes.

Table 27
Pilot Study Mapmark Panelists' Responses to Questions about ALDs
Asked Immediately After Round 1 Bookmark Placements

6. The ALDs appear to be reasonably complete and comprehensive statements of what students should know and be able to do at each level of achievement.	Totally Agree		Somewhat Agree		Totally Disagree	
	10	10	1	0	0	4.43
7. My own level of satisfaction with the Basic achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	8	9	4	0	0	4.19
8. My own level of satisfaction with the Proficient achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	8	10	3	0	0	4.24
9. My own level of satisfaction with the Advanced achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	9	10	2	0	0	4.33

Table 28
ALS Panelists' Responses to Questions about ALDs
(Location Identified by Questionnaire and Sequence Number)

Question	**5**	**4**	**3**	**2**	**1**	**Mean Score**
2-17. The ALDs appear to be reasonably complete and comprehensive statements of what students should know and be able to do at each level of achievement.	Totally Agree		Somewhat Agree		Totally Disagree	
	9	11	6	3	2	3.71
2-18. My own level of satisfaction with the Basic achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	10	10	9	2	0	3.90
2-19. My own level of satisfaction with the Proficient achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	10	10	7	3	1	3.81
2-20. My own level of satisfaction with the Advanced achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	9	11	8	1	2	3.77
6-26. I believe that the achievement levels capture meaningful distinctions in mathematics performance as described in the ALDs.	Totally Agree		Somewhat Agree		Totally Disagree	
	9	15	4	1	1	4.00

Note: Questionnaire 2 was administered immediately after Round 1 bookmark placements. Questionnaire 6 was administered at the conclusion of the meeting.

Panelists' ratings of their understanding of the ALDs is presented by level and round in Table 29. Understanding of the ALDs could conceivably be viewed as an evaluation of the process, as opposed to the ALDs specifically. But panelists' understanding of the ALDs also reflects on how well the ALDs themselves can be understood by teachers, educators, and the general public. As shown in Table 29, panelists reported levels of understanding close to 4.0 early in the process, and understanding continued to build over rounds as panelists continued to study and apply the ALDs to their tasks.

Table 29
Understanding of ALDs
At the time I provided the/my Round __ bookmark placements/cut score recommendations my understanding of the ___ level description was…
(5=Totally Adequate; 3 = Somewhat Adequate; 1 = Totally Inadequate)

	Round			
Level	1	2	3	4
Basic	3.90	4.35	4.45	4.50
Proficient	3.77	4.39	4.42	4.47
Advanced	3.73	4.29	4.45	4.47

It should also be noted that the ALDs anchored the process for establishing the cut scores and that panelists found the ALDs to be useful in the process. Responses to one of the questions in Table 22 shows that panelists found the ALDs to be helpful in the process (mean response = 4.38).

ACT endorses the ALDs for use in representing the achievement levels set in this project. This endorsement is based on: 1) the results of an "ALD evaluation" task that was performed by Pilot Study Mapmark panelists, 2) the responses of panelists in the Pilot Study and ALS meeting to process evaluation questions concerning the ALDs, and 3) the fact that the ALDs were used to anchor the process for establishing the cut scores.

Cut Scores

Table 30 shows the cut scores from the ALS meeting for each panelist by round. The cut scores are organized by table and group. Medians for groups and tables are also shown. The values in the row labeled "all" are the whole group medians. The whole group median is the cut score that was reported for each round. ACT recommends the Round 4 medians, highlighted in yellow in Table 30, as the cut scores for the achievement levels (241 for Basic, 275 for Proficient, and 314 for Advanced). These numbers are on the scale used in the ALS meeting.

ACT conducted extensive statistical analysis on the cut scores in order to assess characteristics related to the reliability of the median and the overall quality of the ALS process. Key analyses and conclusions are summarized in the following sections.

Table 30
NAEP Grade 12 ALS Cut Scores by Panelist
And Medians by Group and Table

Group	Table	Panelist ID	Basic				Proficient				Advanced			
			Round				Round				Round			
			1	2	3	4	1	2	3	4	1	2	3	4
A	1		269	240	246	243	283	274	274	274	300	300	300	300
			221	240	232	232	257	270	264	262	308	317	317	315
			237	243	237	237	269	275	275	275	308	318	318	318
			252	236	245	243	280	276	276	273	291	318	316	316
			232	241	246	243	285	279	281	280	314	314	314	314
	2		255	247	247	247	284	280	280	280	324	306	316	316
			257	248	247	252	284	277	280	285	308	310	314	316
			256	240	240	240	290	275	275	275	316	316	316	316
			251	251	246	246	262	281	279	279	326	320	315	314
			257	258	253	255	285	279	279	279	323	321	314	314
	3		257	248	248	248	294	283	283	288	322	319	319	319
			252	246	246	246	263	271	274	274	294	306	302	302
			226	242	246	246	253	268	279	276	294	317	317	312
			227	237	246	246	257	272	265	265	294	309	294	294
			232	236	236	233	257	260	259	259	283	297	293	295
B	4		252	255	255	235	282	288	282	280	314	314	314	314
			205	215	220	220	245	255	260	260	285	290	290	290
			217	220	226	226	270	267	270	270	301	300	301	301
			215	225	226	233	272	272	272	272	314	314	314	314
			233	235	226	226	274	274	274	274	302	309	301	305
	5		233	245	238	240	274	280	272	273	321	316	301	312
			227	234	239	239	265	270	270	273	314	314	314	314
			255	249	248	248	302	301	293	293	337	337	322	322
			226	245	233	235	255	278	266	271	310	317	312	313
			264	243	243	242	297	279	280	279	330	317	317	317
	6		233	233	233	233	285	273	273	273	310	308	310	310
			254	247	246	247	280	280	280	280	308	310	312	310
			239	239	240	246	272	281	277	277	401	345	323	323
			230	230	230	230	257	264	264	255	317	306	306	300
			245	240	241	241	285	276	278	275	321	312	314	314
			239	239	226	226	267	278	278	280	375	336	336	333
Medians	**All**		**239**	**240**	**241**	**241**	**274**	**276**	**275**	**275**	**314**	**314**	**314**	**314**
	Group	**A**	**252**	**242**	**246**	**246**	**280**	**275**	**276**	**275**	**308**	**316**	**315**	**314**
		B	**233**	**239**	**236**	**235**	**273**	**277**	**274**	**274**	**314**	**314**	**313**	**314**
	Table	**1**	**237**	**240**	**245**	**243**	**280**	**275**	**275**	**274**	**308**	**317**	**316**	**315**
		2	**256**	**248**	**247**	**247**	**284**	**279**	**279**	**279**	**323**	**316**	**315**	**316**
		3	**232**	**242**	**246**	**246**	**257**	**271**	**274**	**274**	**294**	**309**	**302**	**302**
		4	**217**	**225**	**226**	**226**	**272**	**272**	**272**	**272**	**302**	**309**	**301**	**305**
		5	**233**	**245**	**239**	**240**	**274**	**279**	**272**	**273**	**321**	**317**	**314**	**314**
		6	**239**	**239**	**237**	**237**	**276**	**277**	**278**	**276**	**319**	**311**	**313**	**312**

Distribution of Cut Scores by Round

The normality of cut scores within rounds and levels was assessed because some of the statistical procedures and summaries one might consider using with the cut scores are for normally distributed data. The median is typically used in bookmark-based methods because the median is less sensitive to outliers than the mean. It is relatively easy for a bookmark or Mapmark panelist to provide an extreme cut score recommendation either out of inexperience or in an attempt to influence the mean. One would, therefore, expect to find outliers in Mapmark data and to find that the data is not normally distributed.

Two outliers in the data are highlighted in blue in Table 30. These outliers were Round 1 cut score recommendations for the Advanced level from two panelists in Table Group 6. It can be seen that these panelists became less extreme after Round 1.

Table 31 presents p-values from the Shaprio-Wilk (1965) test of normality. The test was performed on the cut scores within achievement level and round. The p-values represent the probability of the null hypothesis that data are normally distributed. The least normally distributed data were the Round 1 cut score recommendations for the Advanced level, as one would expect from the outliers highlighted in Table 30.

As the Advanced cut score recommendations from the two outlying panelists became less extreme in subsequent rounds, the Advanced cut score distributions became more normal, but were still not entirely normally distributed. Distributions for other levels and rounds were more normally distributed. Data from the Pilot Study and other studies show similar trends, although outliers may be found at other achievement levels as well, and there may be no outliers at the Advanced level. That is, we cannot conclude from our data that outliers are more likely to be found at the Advanced level than at other levels.

Table 31
P-values from Shaprio-Wilk test of normality

Level	Round 1	Round 2	Round 3	Round 4
Basic	0.239	0.230	0.054	0.299
Proficient	0.552	0.075	0.190	0.110
Advanced	0.000	0.029	0.025	0.028

Table 32 shows the average absolute difference (AAD) of cut scores from the median by round. Figure 31 is a plot of the AADs by round and level. The AAD was largest for the Advanced level due to the outliers seen at Round 1. In the Pilot Study, the AAD for the Advanced level was lower than the AAD for Basic and Proficient. One cannot generally conclude, therefore, that the variability of cut scores is typically higher for Advanced than for other levels.

The major result shown by the AADs is that differences among panelists' cut score recommendations get smaller over rounds. The most convergence occurs between Rounds 1 and 2. This finding is consistent with results ACT has obtained in previous standard setting work for NAGB.

There is a slight increase in the AAD of the Proficient cut scores from Round 3 to Round 4 and a slight increase in the AAD of the Basic cut scores from Round 2 to Round 3. Similar slight increases were observed in the Pilot Study, and are due to differences among panelists in their reaction to new information such as the student performance data and to discussion. The increases are not large and may occur at any level after Round 2. The lack of large increases in the AAD from Round 3 to Round 4 indicates that there are no extreme reactions among panelists to the student performance data.

Table 32
Average Absolute Difference of Cut Scores from the Median by Round

Level	Round 1	Round 2	Round 3	Round 4
Basic	13.4	6.9	7.3	6.9
Proficient	11.6	6.0	5.6	5.7
Advanced	15.1	7.9	6.8	6.3

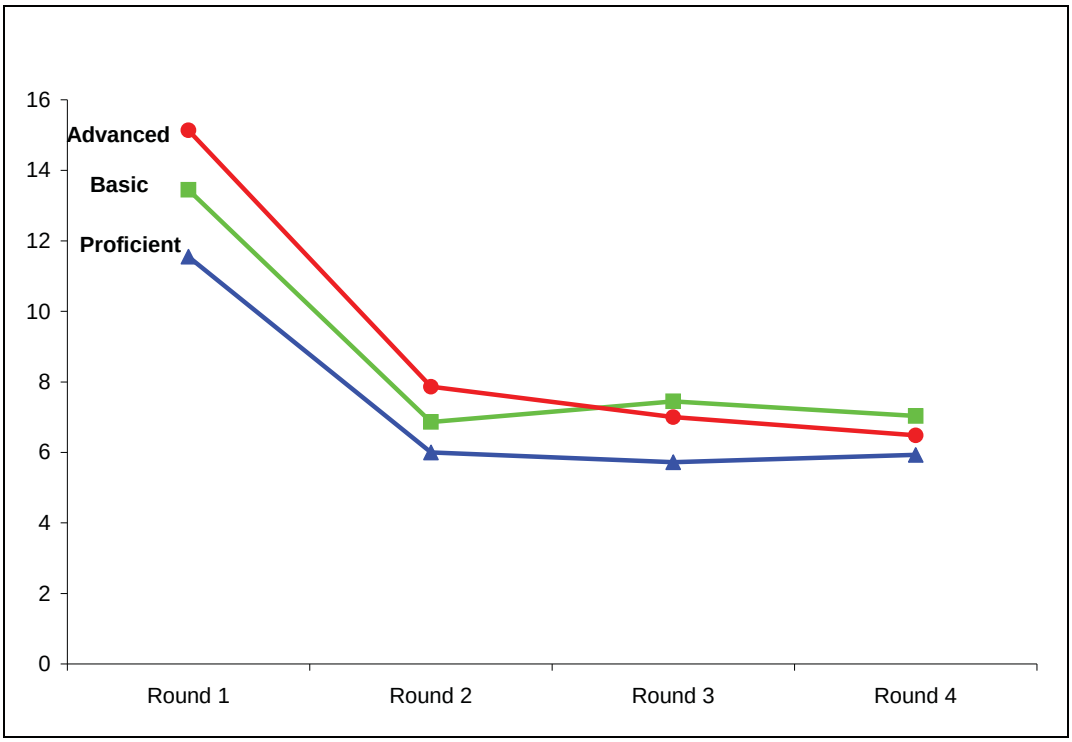


Figure 31. Average Absolute Difference (AAD) of cut scores from median by round.

Table 33 shows by round and level the number of panelists who increased, decreased, or did not change their cut score recommendations. The patterns seen in this table are similar to the patterns seen in previous standard settings for NAGB. The largest frequency of change is from Round 1 to Round 2. At each level, the majority of panelists did not change their cut scores in response to the student performance data. The cut scores tended to move in the direction of the most frequent change—higher from Round 1 to Round 2 for Basic and Proficient because more panelists increased than decreased their cut scores for these two levels. Small differences between the frequencies for higher and lower did not always affect the median cut score (e.g., for Advanced from Round 2 to Round 3 to Round 4).

Table 33
Number of Panelists who Increased, Decreased, or Did Not Change Their Cut Scores from Round to Round

Rounds of Change*	Basic			Proficient			Advanced		
	Increase	No Change	Decrease	Increase	No Change	Decrease	Increase	No Change	Decrease
R1 to R2	15	5	10	15	3	12	12	6	12
R2 to R3	11	9	10	8	13	8	6	13	11
R3 to R4	7	18	5	6	16	7	5	19	5

*R1 means "Round 1," etc.

As shown in Table 34, differences between the mean and median cut scores were generally small. The largest difference was three points. The largest difference occurred at the Advanced level, where the data was least normally distributed. In Round 1, where the Advanced cut scores were least normally distributed, however, the difference between the mean and median Advanced cut scores was only 1 point. This result shows that the mean and median can be the same or very similar even when the data are not normally distributed. This might be a reasonable conclusion to draw from all of the data in Table 34.

Another observation in Table 34 is that the median tends to be higher than the mean cut score. Eight of the ten signed differences in Table 34 are negative. Similar results were found in the Pilot Study and the Grade 8 Study: Nine of ten and six of eight signed differences in the Pilot and Grade 8 studies, respectively, were negative. In the field trials, however, the numbers were equal: 6 positive and 6 negative. It is possible that a fully operational, 4-round Mapmark process differs from the developmental versions in this respect. A predominance of negative values means that panelists cut scores are negatively skewed—the lowest cut scores recommended by panelists tend to be lower than one would expect in a normal distribution.

Table 34
Mean and Median Cut Scores and Difference by Round

Achievement Level	Mean				Median				Mean – Median			
	R1	R2	R3	R4	R1	R2	R3	R4	R1	R2	R3	R4
Basic	240	238	240	239	239	240	241	241	1	-2	-1	-2
Proficient	274	275	275	274	274	276	275	275	0	-1	0	-1
Advanced	315	313	311	311	314	314	314	314	1	-1	-3	-3

Reliability of Cut Scores

The reliability of a standard setting process is typically thought of in terms of how close the final cut scores from the process would be if the process were performed on two occasions with only certain differences allowed between occasions such as a different, randomly equivalent group of panelists and a different, but randomly equivalent set of items. If one were to repeat the process many times, the distribution of "final" cut scores would have a standard deviation—which is called the "standard error" of the cut score. The standard error concept applies equally well to median or mean cut scores.

Unfortunately, ACT can recommend no single, sure method for estimating the standard error of a cut score in the typical standard setting process in which panelists recommend cut scores over rounds, based in part on feedback they receive about the mean or median cut score from the previous round. Standard formulas for computing the standard error of a mean are too simple. After Round 1, the dispersion of cut scores over panelists is too small to be used to estimate the standard error of the mean cut score using the dispersion of cut scores from any subsequent round. Panelists' subsequent cut score recommendations are influenced by the fact that the mean or median cut score is normative. Panelists are generally more comfortable being close to the norm so there is a regression to the Round 1 mean or median cut score. Estimates of the standard error of the *final* mean or median cut score must somehow account for a fundamental regression to the mean (or median) of previous rounds, motivated by panelists' desire for conformity, as well as for the effects of criterion-referenced feedback.

Table 35
Estimates of Standard Error of Cut Scores for Mapmark Studies.

Achievement Level	SE Based on Single Study						SE Based on Replication	
	G8 Study		Pilot (Mapmark)		ALS		FT1--G8	Pilot--ALS
	SE_1	SE_2	SE_1	SE_2	SE_1	SE_2	SE_3	SE_3
Basic	3.46	5.25	3.35	3.00	2.91	5.50	0.50	1.00
Proficient	4.38	4.75	2.72	0.75	2.61	1.00	1.00	2.50
Advanced	1.95	0.75	2.02	0.00	4.32	0.25	0.50	1.00
Average	3.26	3.58	2.70	1.25	3.28	2.25	0.67	1.50

Note: G8 Study is the "Grade 8 Study." FT1 is "Field Trial 1."

Table 35 shows various estimates of standard error that were made over the course of the project. The cells highlighted in yellow show estimates pertaining to the final cut scores from the ALS meeting. The standard error estimates are explained as follows:

SE_1 The first type of standard error (SE_1) estimate is based on the standard deviation of cut scores at Round 1 for the whole group:

$$SE_1 \quad \frac{SD}{\sqrt{n_p - 1}}, \tag{1}$$

where SD is the standard deviation of Round 1 cut scores and n_p is the number of panelists. By using the Round 1 SD, SE_1 avoids the shrinkage in the standard deviation over rounds caused in part by panelists' desire to conform to whatever mean is presented to them in feedback.

SE_2 The second type of standard error estimate is:

$$SE_2 \quad \frac{|Cut_A - Cut_B|}{2}, \tag{2}$$

where Cut_A and Cut_B are final cut scores from Group A and B, respectively, within the same meeting or study. All of the meetings (or studies) from which estimates of SE_2 were obtained in Table 35 were based on the same, two-group (A and B), split-pool design as the ALS meeting. SE_2 represents effects of differences between item pools and groups of panelists. These effects tend to be underestimated by SE_2, however, because the two groups are not independent. They receive feedback about a common, whole-group cut score. Also, the item pools in ACT's designs overlap by approximately 30%.

$\underline{SE_3}$ The third type of standard error estimate is:

$$SE_3 \quad \frac{|Cut_1 - Cut_2|}{2}, \tag{3}$$

where Cut_1 and Cut_2 are final cut scores from two different standard setting meetings that apply the same method. SE_3 represents the effects of all differences between meetings. The Pilot Study and ALS meeting used different groups of panelists but the same item pool. FT1 and the Grade 8 Study used different groups of panelists and overlapping item pools.

SE_3 is probably the best estimate in terms of estimating the quantity of interest in questions about cut score reliability, but it may be unreliable because it is based on only two observations. Likewise, SE_2 is based on just two observations and may be unreliable.

Since there is no reason to suppose that the standard error of the cut score should be larger for one level than for another, and there is no consistent evidence in Table 35 for this, the standard errors within each column of Table 35 have been averaged. These averages range from 0.67 to 3.58, and themselves average 2.3 points. The average of estimates pertaining exclusively to the ALS meeting is also 2.3.

Overall, we conclude that the cut scores from the ALS meeting are reliable. With an average estimated standard error of 2.3 points, one would expect the average difference between cut scores from two independent meetings using the same or randomly equivalent item pools to be about 3.3 points, and for the difference to be less than 6.6 points 95% of the time.

This overall level of reliability is illustrated in Figure 32. It can be seen that by Round 4, cut scores from the ALS meeting ended up very close to the Round 4 cut scores from the Pilot Study Mapmark procedure. The differences between the Pilot Mapmark and ALS cut scores (Figure 32) were 2 points at Basic (239 vs. 241), five points at Proficient (270 vs. 275), and two points at Advanced (314 vs. 316). These differences are well within a 95% confidence interval if the standard error of a Round 4 cut score is assumed to be about 2.3 points.

ACT is studying other methods of evaluating the reliability of *median* cut scores across occasions where panelists may be considered to be randomly sampled for purposes of statistical hypothesis testing. No completely satisfactory method has yet been found for

estimating the reliability of the mean or median. However, ACT is confident in concluding that the ALS process produces reliable cut scores.

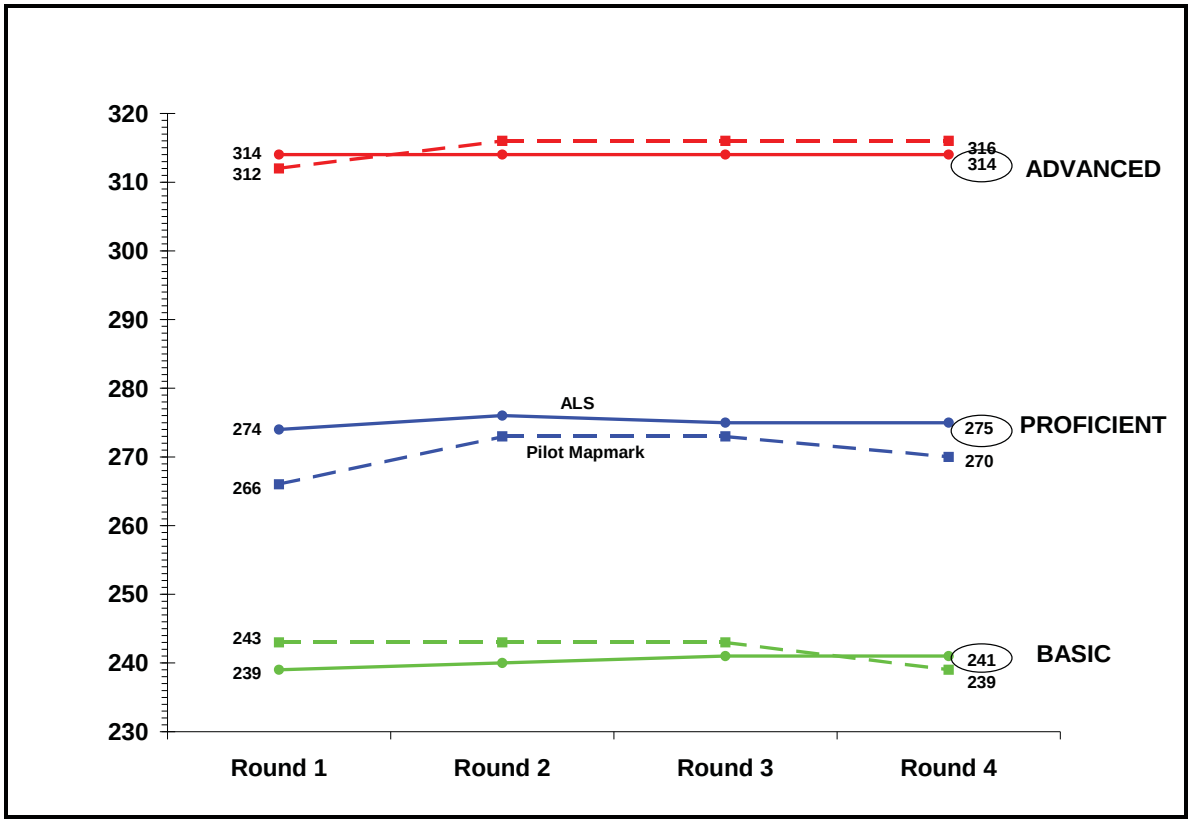


Figure 32. Pilot Study Mapmark and ALS meeting cut scores by round.

Design and Panelist Type Effects on Cut Scores

After a thorough review of the effects of design factors (tables and groups) and panelist characteristics on cut scores, ACT's Technical Advisory Committee on Standard Setting (TACSS) did not identify any effects that called the results of the ALS meeting into question or raised serious questions about the process.

As noted above with regard to estimating the reliability of median cut scores, there is no satisfactory method of estimating effects on median cut scores. As a substitute, ACT performed analyses of the effects on means. Very few statistically significant effects emerged from these analyses, but those that did will be mentioned along with a brief description of differences in medians.

Figure 33 shows the group median cut score by round and achievement level. Group differences at the Basic level were large at Rounds 1 and 3 in relation to corresponding AADs. As shown more precisely in Table 36, the group difference in the Basic median cut score (Group A minus Group B) was 19.0, 3.0, 10.5, and 11.0 points at Rounds 1, 2, 3, and 4, respectively. The group difference largely disappeared at Round 2, then reappeared at

Round 3 and remained at Round 4. This may not be unusual. Round 2 was based on domain scores. In Round 3 panelists could return to their Ordered Item Book which was used in Round 1. It seems natural for Round 3 to represent a compromise between two types of information. One would expect the groups to regain about half of the amount of difference they had given up in Round 2.

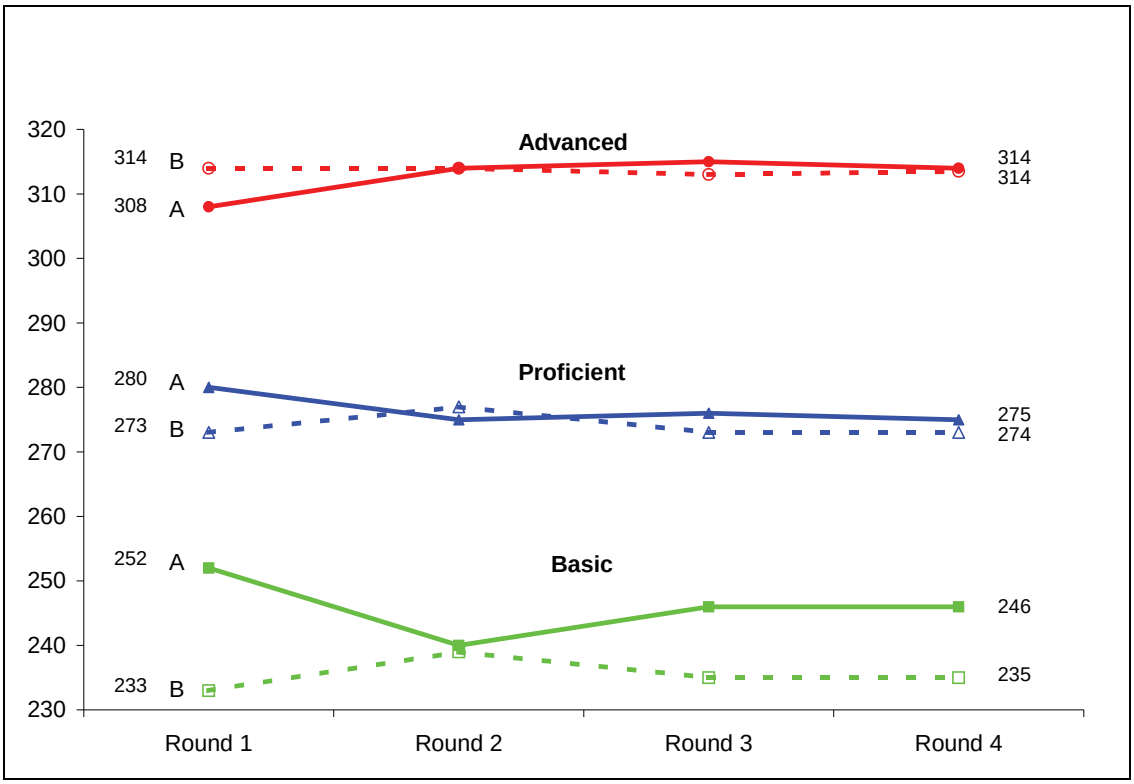


Figure 33. Median cut scores by group.

Figure 34 shows table medians by round and achievement level. Table effects tended to be large in Round 1, especially at the Advanced level, but decreased in later rounds. Table 36 shows that the largest within-group difference between table median Advanced cut scores was 29 points at Round 1 and only 8 points at Round 2.

The effects of panelist type and gender after Round 1 were small in relation to AADs for all achievement levels. Table 36 shows that the male/female difference in median Basic cut scores was 15.5 at Round 1, but only 3.0 points at Round 2. The largest difference between Basic median cut scores by panelist type was 19 points at Round 1, but only 6 points at Round 2.

Table 36
Medians and Average Absolute Difference (AAD) of Cut Scores by Factor Level

		Round 1						Round 2					
		Basic		Proficient		Advanced		Basic		Proficient		Advanced	
Factor	n	Md	AAD	Md	AAD	Md	AAD	Md	AAD	Md	AAD	Md	AAD
Type													
Teacher	17	233.0	11.5	270.0	11.2	314.0	14.1	240.0	6.2	275.0	4.4	316.0	6.1
Non-Tea	5	239.0	17.2	272.0	16.8	308.0	29.0	240.0	8.6	277.0	11.4	317.0	16.4
GP	9	252.0	10.0	282.0	7.1	314.0	7.7	246.0	5.0	280.0	4.8	310.0	4.8
Max. Diff		19.0		12.0		6.0		6.0		5.0		7.0	
Ethnicity													
White	22	235.0	12.3	271.0	12.7	314.0	17.3	242.5	7.3	278.0	7.0	316.5	9.0
Black		254.0	4.0	284.5	2.8	312.0	9.5	240.0	3.0	276.0	0.5	314.0	3.0
Asian		234.5	19.5	276.0	4.0	311.0	3.0	236.0	6.5	276.0	4.0	312.0	2.0
Hispanic		227.0	14.0	265.0	8.7	300.0	6.7	237.0	2.0	272.0	1.3	309.0	4.7
Max. Diff		27.0		19.5		14.0		6.5		6.0		7.5	
Gender													
Male	18	248.5	14.8	277.0	12.8	311.0	19.8	240.0	6.5	275.5	5.7	315.0	9.8
Female	13	233.0	9.3	272.0	10.5	314.0	7.9	243.0	6.4	278.0	5.8	314.0	4.7
Max. Diff		15.5		5.0		3.0		3.0		2.5		1.0	
Region													
Midwest	8	244.5	10.8	271.5	11.8	308.0	10.4	244.5	5.8	274.5	8.0	309.5	7.1
Northeast	10	236.0	12.7	273.0	10.6	318.5	20.4	240.0	4.7	278.5	3.1	316.5	6.7
South	6	248.5	12.7	282.0	9.7	315.5	10.8	242.5	7.3	277.0	3.7	314.5	6.0
West	7	233.0	16.3	265.0	12.9	310.0	10.7	240.0	9.4	273.0	7.1	314.0	7.6
Max. Diff		15.5		17.0		10.5		4.5		5.5		7.0	
Group													
A	15	252.0	11.8	280.0	12.5	308.0	11.3	242.0	4.7	275.0	4.3	316.0	5.9
B	16	233.0	12.2	273.0	11.0	314.0	17.0	239.0	8.3	277.0	7.1	314.0	9.2
Max. Diff		19.0		7.0		6.0		3.0		2.0		2.0	
Table													
A:1	5	237.0	13.6	280.0	8.4	308.0	6.2	240.0	1.6	275.0	2.2	317.0	4.4
A:2	5	256.0	1.6	284.0	5.8	323.0	5.2	248.0	4.4	279.0	1.8	316.0	5.0
A:3	5	232.0	11.2	257.0	9.4	294.0	7.8	242.0	4.2	271.0	5.4	309.0	6.6
Max. Diff		24.0		27.0		29.0		8.0		8.0		8.0	
B:4	5	217.0	13.0	272.0	8.2	302.0	8.4	225.0	11.0	272.0	8.0	309.0	7.6
B:5	5	233.0	13.2	274.0	15.8	321.0	8.6	245.0	3.4	279.0	6.6	317.0	4.8
B:6	6	239.0	6.0	276.0	9.0	319.0	27.0	239.0	4.0	277.0	4.3	311.0	11.5
Max. Diff		22.0		4.0		19.0		20.0		7.0		8.0	
Overall	31	239.0	13.4	274.0	11.6	314.0	15.1	240.0	6.9	276.0	6.0	314.0	7.9

Table 36 (continued)
Medians and Average Absolute Difference(AAD) of Cut Scores by Factor Level

		Round 3						Round 4					
		Basic		Proficient		Advanced		Basic		Proficient		Advanced	
Factor	n	Md	AAD	Md	AAD	Md	AAD	Md	AAD	Md	AAD	Md	AAD
Type													
Teacher	17	240.0	7.4	275.0	5.0	314.0	6.7	240.0	6.8	275.0	4.9	314.0	6.4
Non-Tea	5	240.0	8.6	277.0	9.8	317.0	8.2	246.0	9.6	277.0	11.2	316.0	8.0
GP	9	246.0	4.4	274.0	4.0	312.0	5.6	243.0	4.6	274.0	3.9	312.0	4.6
Max. Diff		6.0		3.0		5.0		6.0		3.0		4.0	
Ethnicity													
White	22	239.0	8.2	276.0	6.4	314.0	7.7	238.5	7.9	275.5	6.6	314.0	6.9
Black		243.0	2.8	277.0	1.8	315.0	1.0	242.0	3.5	275.0	3.0	316.0	0.5
Asian		236.0	10.0	276.0	4.0	313.0	1.0	240.0	7.0	276.0	3.0	312.0	2.0
Hispanic		246.0	2.3	270.0	3.0	300.0	6.7	243.0	2.3	273.0	3.0	300.0	6.7
Max. Diff		10.0		7.0		15.0		4.5		3.0		16.0	
Gender													
Male	18	242.0	7.3	275.5	5.4	314.0	8.3	242.5	7.5	274.5	5.9	314.5	7.9
Female	13	238.0	7.0	275.0	5.8	314.0	4.6	237.0	5.4	275.0	5.5	313.0	3.9
Max. Diff		4.0		0.5		0.0		5.5		0.5		1.5	
Region													
Midwest	8	241.5	7.3	274.5	7.1	309.0	8.0	241.5	8.4	274.5	8.9	307.5	8.0
Northeast	10	240.0	6.6	276.0	4.7	315.0	7.2	241.0	6.1	276.0	4.8	314.5	5.6
South	6	243.0	6.7	277.0	4.8	315.0	4.0	242.0	6.0	274.0	4.8	315.0	3.8
West	7	246.0	7.9	274.0	5.3	314.0	6.6	239.0	6.7	274.0	4.1	312.0	6.0
Max. Diff		6.0		3.0		6.0		3.0		2.0		7.5	
Group													
A	15	246.0	3.4	276.0	5.0	315.0	5.9	246.0	4.6	275.0	5.7	314.0	5.7
B	16	235.5	8.1	273.5	5.9	313.0	7.6	235.0	6.8	273.5	5.7	313.5	6.9
Max. Diff		10.5		2.5		2.0		11.0		1.5		0.5	
Table													
A:1	5	245.0	4.6	275.0	3.8	316.0	4.2	243.0	3.4	274.0	4.0	315.0	4.0
A:2	5	247.0	2.8	279.0	1.2	315.0	0.8	247.0	4.2	279.0	2.2	316.0	0.8
A:3	5	246.0	2.4	274.0	7.6	302.0	9.8	246.0	3.0	274.0	8.0	302.0	8.4
Max. Diff		2.0		5.0		14.0		4.0		5.0		14.0	
B:4	5	226.0	7.0	272.0	5.2	301.0	7.4	226.0	4.4	272.0	4.8	305.0	7.4
B:5	5	239.0	4.0	272.0	7.4	314.0	5.2	240.0	3.2	273.0	5.6	314.0	2.8
B:6	6	236.5	6.3	277.5	3.7	313.0	7.5	237.0	7.5	276.0	5.7	312.0	8.3
Max. Diff		13.0		5.5		13.0		14.0		4.0		9.0	
Overall	31	241.0	7.3	275.0	5.6	314.0	6.8	241.0	6.9	275.0	5.7	314.0	6.3

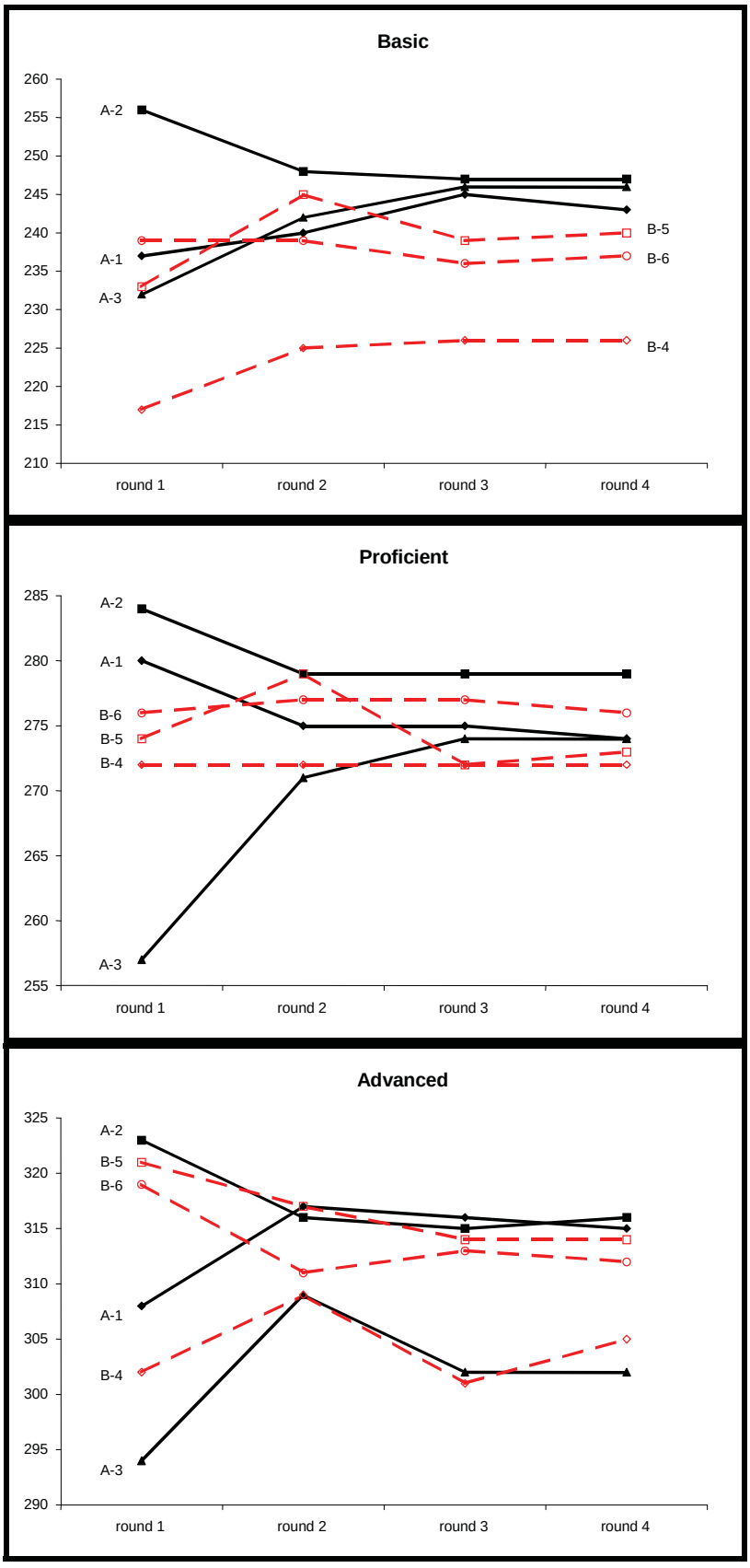


Figure 34. Median cut scores by level and table.

Exemplar Item Ratings

Exemplar item ratings were gathered in the ALS meeting to provide NAGB with information concerning the suitability of assessment items for illustrating what students know and can do at each level of achievement.

Potential exemplar items were drawn from three blocks of the assessment selected for eventual release to the public. There were a total of 53 potential exemplar items. There were 38 multiple choice items, 4 dichotomously-scored constructed response items, and 11 polytomously-scored constructed response items. These represented a total of 68 steps, or score points. Items/score points were mapped to the first, or easiest, achievement level at which the probability was 0.67 or higher that a student at the midpoint or median of the level could correctly answer the item or attain the score point. Recall that each score point of a polytomously-scored item was mapped independently of other score points by the probability of scoring at-or-above the score point.

Using the 0.67 probability rule, 65 of the 68 score points were mapped to the achievement levels. Three score points were too difficult for any achievement level. That is, the probability of earning those score points was less than 0.67 for a median Advanced student. Of the three "too difficult" score points, one was a multiple choice item, one was a dichotomously-scored item, and one was the highest score point on a polytomously-scored item.

The number of score points per level is shown in the "Overall" column of Table 37. There were 17 score points mapped to Basic, 31 mapped to Proficient, and 17 mapped to Advanced. The number of score points per level by item type is shown by the right-most number in the remaining columns.

Table 37
Number of Score Points Meeting Rating Criteria Out of Total Mapped to Level Overall and by Item Type

Level	Overall	By Item Type		
		Multiple Choice	Dichotomously Scored	Polytomously Scored
Basic	12 of 17	9 of 10	0 of 0	3 of 7
Proficient	20 of 31	13 of 16	2 of 3	5 of 12
Advanced	10 of 17	6 of 11	0 of 0	3 of 6
Total	42 of 65	28 of 37	2 of 3	11 of 25

Detailed results of the exemplar item ratings are shown in Appendix I. The percent of responses falling into each of the rating categories is shown separately for Pilot Study and ALS panelists. In the Pilot Study, only Blocks 3 and 4 were rated as potential exemplar items, so only ALS results are shown for items in Block 12.

One set of ratings criteria ACT suggests be considered for dividing the potential exemplars mapped to an achievement level into "use" and "do not use" categories is that: 1) at least 33% of panelists checked the "very good" category, and 2) fewer than 33% of panelists checked the "do not use" category.

Results of applying this criterion using only the ALS ratings data (not the Pilot Study ratings) are summarized in Table 37. Overall, 42 of the 65 score points mapped to achievement levels (65%) met the criterion. These included 28 of 37 multiple choice items (76%), 2 of 3 dichotomously-scored items (67%), and 11 of 25 points on polytomously-scored items (44%).

It might be noteworthy that the percentage of score points meeting ACT's suggested ratings criteria was substantially higher for multiple choice items than for polytomously-scored items. This is likely due to the nature of the items themselves rather than to the specific ratings criterion. Score points associated with polytomously-scored items simply receive fewer "very good" ratings and more "do not use" ratings.

Since there were so few dichotomously-scored items, and the percentage meeting the criterion was intermediate with multiple choice and polytomously-scored items (66%), it is unclear whether the lower ratings for polytomously-scored items are due to their criterion-referenced nature (which they share with dichtomously-scored items) or to the partial-credit scoring.

ALS panelists' responses to process evaluation questions concerning the exemplar items are shown in Table 38. These were questions 20 and 21 on the last process evaluation questionnaire (#6). The average rating to the question, "The exemplar items I reviewed seemed appropriately matched to their achievement level," 3.55, was acceptable, but lower than one would expect from Pilot Study panelists' responses to similar questions, which are shown in Table 39.

Table 38
Responses of ALS Panelists to Questions about Exemplar Items

Question	5	4	3	2	1	Mean Score
6-20. I believe the exemplar items will be useful for describing the achievement levels.	Totally Agree 11	10	Somewhat Agree 7	1	Totally Disagree 0	4.07
6-21. The exemplar items I reviewed seemed appropriately matched to their achievement level.	Totally Agree 4	12	Somewhat Agree 9	4	Totally Disagree 0	3.55

Table 39
Responses of Pilot Study Panelists to Questions about Exemplar Items

28. Overall, the Exemplar Items I rated to illustrate student performance at the Basic Achievement level are appropriate. (IR7-16)	Totally Agree		Somewhat Agree		Totally Disagree	
	10	11	0	0	0	4.48
29. Overall, the Exemplar Items I rated to illustrate student performance at the Proficient Achievement level are appropriate. (IR7-17)	Totally Agree		Somewhat Agree		Totally Disagree	
	10	11	0	0	0	4.48
30. Overall, the Exemplar Items I rated to illustrate student performance at the Advanced Achievement level are appropriate. (IR7-18)	Totally Agree		Somewhat Agree		Totally Disagree	
	8	10	3	0	0	4.24

The difference between Pilot Study and ALS ratings with regard to appropriateness of exemplar items for the achievement levels may be related to a point made earlier, that ALS panelists were generally more critical than Pilot Study panelists in their content-based judgments. Pilot Study panelists rated two of the three blocks of items rated by ALS panelists. Except for one score point, the match between items/score points and achievement levels was the same for items in those blocks. This is shown by the overlap of Pilot Study and ALS data for each item in Appendix I (except for items in Block 12). Table 40 shows that for the common items in the two studies, the rate of unacceptable percentages flagged in yellow in Appendix I (33% or higher "do not use" or less than 33% "very good") is twice as high in the ALS meeting as in the Pilot Study.

Table 40
Rate of "Unacceptable" Percentages* of Exemplar Item Ratings in the Pilot Study and ALS Meeting

Achievement Level	Blocks 3, 4, and 12	Blocks 3 and 4	
	ALS	ALS	Pilot
Basic	**8/34 (24%)**	**5/24 (21%)**	**4/24 (17%)**
Proficient	**16/62 (26%)**	**16/42 (38%)**	**4/42 (10%)**
Advanced	**13/34 (38%)**	**11/24 (45%)**	**7/24 (29%)**
Total:	**37/130 (28%)**	**32/90 (36%)**	**15/90 (17%)**

*Less than 33% "Very Good" or 33% or higher "Do Not Use."

ALS panelists' average response to the statement shown in Table 38, "I believe the exemplar items will be useful for describing the achievement levels," is higher (4.07) and may reflect their expectation that their ratings will be used to select exemplar items from the pool of potential exemplar items.

ACT's suggested ratings criteria (at least 33% "very good" *and* fewer than 33% "do not use") leaves a sufficient number of potential exemplar items of both multiple choice and constructed response type for NAGB to choose exemplar items from. Even with the lower ratings of polytomously-scored items, each achievement level was associated with at least three score points on partial credit (polytomously-scored) items that met ACT's suggested ratings criteria. Moreover, there was one polytomously-scored item that met the criteria with one or more of its score points at each level.

The ratings information in Appendix I actually allows NAGB to apply any ratings criteria to select exemplar items from the pool mapped to an achievement level. For example, NAGB could apply different ratings criteria to different types of items, depending on the need to include certain types of items, items of a particular content area or subscale, or items of certain difficulty, among the exemplar items.

ACT recommends that NAGB use the lists of items mapped to the achievement levels in the ALS meeting, and the exemplar item ratings data, along with other criteria of its choosing, to select exemplar items for the achievement levels. This recommendation is based on the following:

- The statistical criteria used to map items to achievement levels results in a sufficient number of items being associated with each level (shown in Table 37).
- Reasonable criteria applied to the exemplar item ratings leaves a sufficient number of items from which NAGB can choose from and/or apply additional criteria to in order to select exemplar items (shown in Table 37).
- Panelists understood their exemplar item rating task.
- Panelists felt that the exemplar items they reviewed seemed appropriately matched to their achievement level.
- Panelists' felt that the exemplar items will be useful for illustrating the achievement levels.

The specific items or score points meeting ACT's suggested criteria are identified by color code and item handle in Appendix I. The color codes are explained in footnotes to the tables in the appendix. Item handles are also associated with other item identification data in the Technical Report.

CONCLUSIONS DRAWN FROM THE ALS PROCESS

Many issues concerning the Mapmark method were investigated over the course of the project. This section presents issues and conclusions that are most essential to the Mapmark method, as performed in the ALS meeting, and to the recommendations ACT offers for future standard setting efforts.

The Response Probability Criterion

The response probability (RP) criterion is a conditional probability by which a scale value, and consequently a cut score, is associated with an item in the bookmark and Mapmark methods. A panelist's recommended cut score in Round 1 of the Mapmark method is the scale value of the item that receives the bookmark. Items are located on item maps and ordered in the Ordered Item Book by their RP-criterion-based scale value.

The RP criterion is recognized as a critical issue because different choices of an RP criterion can lead to different cut scores. In order to obtain the same cut score, the higher the RP criterion, the more difficult the item that receives the bookmark should be. But panelists typically under-adjust for differences in the RP criterion. Higher RP criteria typically produce higher cut scores.

In recognition of the RP criterion issue, ACT and NAGB decided to treat the choice of the RP criterion as a policy decision. ACT investigated and chose RP criteria for its achievement level studies, but the choice of the RP criterion for the operational ALS meeting was ultimately a NAGB decision. NAGB chose a 0.67 RP based on ACT's recommendation. At the outset of the project, ACT provided the following reasons for using a 0.67 RP:

- A 0.67 RP is traditionally used in the bookmark procedure. Use of the same RP facilitates comparisons of standards across different assessment programs.

- A 0.67 RP corresponds to a simple fraction, 2/3. Use of a simple fraction facilitates understanding of the probability by more panelists and allows panelists to apply this probability to their task in ways that are not possible with a more complex fraction. [A 0.5 RP criterion also corresponds to a simple fraction, so this consideration does not point exclusively to a 0.67 RP.]

- ACT's previous research under NAGB contract (Webb & Loomis, 1993; American College Testing, 1995; Loomis, 1999; Yang, 1999) indicated that an RP of 0.67 would produce cut scores close to those obtained with ACT's Item Rating method.

- A 0.67 RP is close enough to optimal in terms of the theoretical psychometric relationship between RP criterion and cut score reliability.

- A 0.67 RP criterion tends to produce reasonable standards in other assessment programs.

In addition, ACT decided to make the RP criterion more visible to panelists than it is typically made in the traditional bookmark method. The RP criterion was used to explain to panelists the location of items on their item maps and the order of items in their ordered item booklet. Panelists were repeatedly reminded of the RP in their instructions for placing their bookmarks.

ACT's research findings on the RP issue early in the project may be related to the emphasis that was placed on making panelists aware of the RP criterion:

- Panelists using a 0.5 RP expressed difficulty and confusion in placing their bookmark relative to items whose content students "should know and be able to do" according to the achievement level description. An "even chance" of getting an item correct did not seem high enough to signify satisfactory performance on content specified in the achievement level description.

- Panelists using a 0.67 RP reported no difficulty placing their bookmarks relative to content that students should know and be able to do according to the achievement level descriptions.

ACT's subsequent use of a 0.67 RP for the remainder of the project produced the following additional findings:

- Panelists were comfortable using a 0.67 RP to define mastery in placing their bookmarks.

- The cut scores obtained with a 0.67 RP were reasonably close to those obtained using the Item Rating method.

- The cut scores obtained with a 0.67 RP are likely to be considered reasonable in terms of achievement level percentages.

The reasonableness of results obtained in this project supports the choice of a 0.67 RP for the Grade 12 mathematics assessment and suggests that this RP may be a reasonable choice or starting point for investigating the issue in future standard setting work. Whatever the RP, ACT recommends that it be made highly visible to panelists in their training.

The Use of Domains in Mapmark

The use of domains and domain score feedback was considered an optional feature of the Mapmark method. ACT was prepared to develop the Mapmark method solely around item maps and the bookmark method if the use of domains proved to be too complicated or burdensome to the process.

Based on results of field trials conducted early in the project, ACT decided to fully incorporate domains into the Mapmark procedure. Data collected on the use of domains in fully-developed, five-day and four-day Mapmark procedures involving four rounds of ratings indicated that panelists are capable of understanding domain scores and using them in a logical and defensible fashion to recommend cut scores. Percent correct score feedback on domains provides panelists with a useful, additional perspective on their cut scores, thereby increasing the overall validity of the process. These claims are supported by the following process evaluation results obtained at the conclusion of Round 2 of the ALS meeting, where panelists used domain score feedback exclusively to recommend cut scores. Panelist's responses indicated that:

- They understood the domain score feedback.

- As a result of the domain score tasks, they understood the concept of domain scores.

- They were comfortable thinking about whether an item was like other items in its domain.

- They were confident deciding whether domain scores should be higher/OK/lower relative to the ALDs.

- They understood how to use the Domain Item Maps, the Domain Ordered Item Booklet, and the Domain Score Chart.

The tasks necessary to help panelists understand the domains and domain scores also help panelists understand the limitations of inferences based on single test items. This understanding helps panelists avoid placing too much importance on where their cut scores lie with respect to individual test items, which increases the overall validity of the process.

ACT's TACSS favored the Mapmark procedure in part because domains have the potential for improving the interpretation of scores and understanding of achievement levels in NAEP reports. Based on this potential, ACT recommends that NAGB explore the use of domains in describing the achievement levels that NAGB will set.

In supporting this recommendation before the COSDAM, ACT noted that:

- the standard setting panel is representative of the audience for NAEP—teachers, nonteacher-educators, and the general public. If domains can be understood and used by the panelists to understand student achievement, it is likely they can also be useful to the audience targeted by NAEP reports; and

- the use of domains and percent correct scores can help NAGB avoid the misconceptions that might be associated with the use of individual items to illustrate what students in achievement levels can or cannot do.

ACT's TACSS was unsure how well domains would work for setting standards or describing achievement levels in other subject areas, such as Reading, where skills may be less sequential or hierarchical in difficulty.

Identification of Item Knowledge, Skills, and Abilities (KSAs)

Besides the actual placing of the bookmark, the other key component of bookmark incorporated into Mapmark is the identification of the knowledge, skills, and abilities (KSAs) required by test items. This is done by panelists in a number of distinct stages and activities that are performed over a total of approximately ten hours time-on-task. First, panelists study the progression of knowledge, skills, and abilities needed to reach

successively higher score levels on each of the polytomously-scored items in their pool. Then, they identify the KSAs in all of the items in their pool in the context of the Ordered Item Booklet. Polytomously-scored items are included in the OIB, but are seen for the first time in the context of the overall progression of item difficulty in the assessment. Panelists begin with the easiest, or first item in the book, and proceed sequentially through the book, identifying the KSAs required by each item and what additional KSAs an item appears to require that were not required by easier items representing similar content. This review is done independently first, and then in a table-group discussion process in which panelists share their thoughts on the KSAs.

ACT, its TACSS, and the COSDAM, felt that the KSA review was the most attractive feature of the bookmark method, and cited it as an important factor in preferring the Mapmark method over Item Rating. It was generally felt that the amount of time panelists spend on the KSA review (approximately 10 hours total) helps panelists acquire the familiarity and knowledge they need with the assessment in order to recommend cut scores. It was noted that the KSA review helps panelists see the "big picture" of progression in knowledge, skills, and abilities that is required for students to obtain higher scores on the assessment. Also, since the achievement level descriptions themselves collectively represent a progression of KSAs in student achievement (Basic, Proficient, and Advanced), the KSA review helps panelists match the ALDs to the progression of student achievement on the test. Finally, it was noted that the component of the KSA review concerning polytomously-scored items helps panelists understand and use these types of items in their subsequent tasks. These perceptions were supported by the following findings in this project:

- Panelists gave the Ordered Item Book a high average rating for being helpful in the standard setting process.

- Panelists gave the table discussion of the OIB in which they shared ideas about KSAs required by test items a high average rating for usefulness.

- Following the KSA review, panelists reported that they understood the score levels of the polytomous items.

- Panelists reported that the KSA review helped them understand what makes one item harder than another.

The KSA review pays dividends later in the Mapmark method when panelists review the items in tasks designed to help them understand the domains and domain scores. This is one reason why ACT selected the bookmark kernel over an Item Rating kernel in its proposal to incorporate domains into a standard setting process for NAEP assessments.

The Use of Item Maps

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, but it is not spatially representative. Item maps were added in the

Mapmark method to provide a visual representation of differences in item difficulty (on the vertical scale) and of similarity in content (by organizing items into columns). These representations were expected to be useful to panelists in their KSA review and in placing their bookmarks. The organization of items into content-columns helps panelists compare one item to others of similar content in the KSA review. The spatial representation of differences in item difficulty on the same scale as student achievement helps panelists evaluate the difference between items and their associated KSAs when placing their bookmarks.

Results from the ALS meeting and developmental Mapmark studies showed that panelists understood their item maps, found them useful, and were able to integrate them with the other materials they were given. Specifically,

- panelists gave high ratings of helpfulness to their item map, and

- panelists reported that they understood how to use their item map and Ordered Item Booklet.

The Concept of Borderline Performance

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 perform 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.

Results from the Pilot Study and the ALS meeting show that Mapmark panelists were able to successfully develop and apply their concept of borderline performance. Specifically,

- at the conclusion of Round 1, Mapmark panelists' ratings of their comfort level in using the concept of borderline performance to place their bookmarks is comparable to Item Rating panelists ratings of how well formed their concept of borderline performance was at the time they provided their Round 1 ratings;

- in subsequent rounds, Mapmark panelists reported that their concept of borderline performance was well formed. Average ratings were 4.34, 4.41, and 4.54 for Rounds 2, 3, and 4, respectively. These averages compared favorably to Item Rating results from the Pilot Study and previous ALS meetings; and

- Mapmark panelists understood the distinction between borderline performance and typical performance in the achievement level. (Average ratings of understanding

this distinction were 4.52, 4.58, and 4.47, respectively, at the conclusion of Rounds 1 to 3.)

Independence Among Panelists

For NAGB, the Mapmark process was implemented in a way that encourages panelists to learn from the perspective and experience of other panelists, but to basically maintain their own perspective and independent judgment. It is important to give panelists reasons for the importance of independent judgement and to provide specific instructions on this point because panelists can easily see the bookmark placements and cut score recommendations of other panelists at their table. The following reasons for valuing independent cut score recommendations were given to Mapmark panelists:

- 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.

In summary, the group is wiser than any individual within the group. 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.

ACT used questions from the process evaluation questionnaires to reinforce this perspective as well as to evaluate whether it was accepted by panelists. Indications that the panelists did in fact accept this perspective included the following:

- At the conclusion of each round, panelists indicated near total *disagreement* with the statement, “I felt pressure to recommend bookmarks/cut scores that were close to those recommended by other panelists.”

- At the end of the process, panelists indicated that their input was valued and considered by other members of the group.

- The variability of cut scores within table groups was smaller than the overall variability, but still substantial.

- There was no instance in any round where all of the panelists at a table, or even more than three panelists, recommended the same cut score.

Identifying a Range of Uncertainty for Bookmark Placements

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.

Unlike in the typical Bookmark procedure, 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 a little further, beyond the first item they judge to be too difficult, to see if there are any 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 20 shows a slide that was used to illustrate this concept for panelists.

In achievement level setting studies leading up to the operational ALS meeting, ACT found that the supplemental instructions were very important and needed to be emphasized through the slide and other means. This is because some panelists, through observation and self-report, were known to have felt, in retrospect, that they placed their bookmark too early in the OIB because they stopped at the very first item that seemed "too difficult."

SUMMARY AND CONCLUSION

For the purposes of helping NAGB set achievement levels for the 2005 NAEP in Grade 12 Mathematics, ACT:

- developed a new standard setting method, Mapmark, based on the bookmark method;

- conducted a pilot test in which cut scores for the 2005 NAEP in Grade 12 mathematics were set using Mapmark and an Item Rating procedure fundamentally

similar to the procedure ACT used to set achievement levels for the 1998 NAEP in Civics; and

- implemented Mapmark for the operational ALS meeting.

Based on these activities, ACT provided NAGB with the following input regarding the three recognized outcomes of the achievement level setting process: 1) achievement level descriptions, 2) cut scores, and 3) exemplar items.

- ACT endorses the achievement level descriptions that were used in the operational ALS meeting.

- ACT recommends the cut scores from Round 4 of the operational ALS meeting. These cut scores are currently not on the scale that will be used to report the 2005 assessment results.

- ACT recommends that NAGB use the lists of potential exemplar items and panelists' ratings from the ALS meeting in the process of selecting exemplar items.

A recommendation concerning the use of domains was also offered:

- ACT recommends that NAGB explore the use of domains for describing the achievement levels.

These recommendations and endorsements are based on positive evaluations and conclusions concerning relevant elements of the process by panelists, ACT's Technical Advisory Committee on Standard Setting, and by members of the NAGB Committee on Standards, Design, and Methodology.

REFERENCES

American College Testing (1995, February). *NAEP reading revisiedt: An evaluation of the 1992 achievement levels descriptions.* Iowa City, IA: Author.

American College Testing (1993, September). *Setting achievement levels on the 1992 National Assessment of Educational Progress in Mathematics, Reading, and Writing.* A report presented to the National Assessment Governing Board. Iowa City, IA: Author.

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

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.

Loomis, S. C. (1999, April). *Civics item classification study: 1998 Civics NAEP*. Paper presented at the meeting of the Technical Advisory Committee for Standard Setting, Saint Paul, MN.

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.

National Assessment Governing Board (August, 2003). *NAGB policy*. (Appendix A in Attachment A (*Statement of Work*) to Solicitation No. ED-03-R-0021. *Twelfth Grade Mathematics Achievement Levels*). Washington, D.C.: Author.

Reckase, M. (2000, June). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT.* Iowa City: ACT, Inc.

Schulz, E. M., Lee, W., & Mullen, K. (2005). A domain-level approach to describing growth in achievement. *Journal of Educational Measurement, 42,* 1-26.

Schulz, E. M. (2003). *Describing Achievement Levels on the Grade 8 NAEP Mathematics Assessment Using Multiple Domain Scores*. Unpublished report to the National Center for Education Statistics.

Schulz, E. M., Kolen, M. & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement, 23,* 347-362.

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality. *Biometrika, 52,* 591-611.

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

Webb, M. W. II & Loomis, S. C. (1993, April). *Content validity studies related to setting achievement levels on the 1992 National Assessment of Educational Progress in mathematics, reading, and writing.* Paper presented at the annual meeting of the American Educational Research Association, Atlanta.

Yang, W. (1999, June). Do teachers agree with one another for their classification for the 1998 NAEP Civics items? Paper presented at the meeting of the Technical Advisory Committee on Standard Setting, Atlanta.

Appendix A

Achievement Level Descriptions

Mathematics Achievement Level Descriptions

The NAEP framework was developed under the assumption that most students in the twelfth grade will have taken courses that contain the mathematics of algebra I, algebra II, and geometry; they would also have received instruction in specified topics in data analysis and probability. The framework reaffirms the importance of understanding numbers and operations as computation is used in all of framework content areas. It is also important to note that achievement levels are not synonymous with levels of complexity. For example, there may be a high complexity item that falls within the basic range, and conversely, a low complexity item that maps at the advanced level.

BASIC

Twelfth-grade students performing at the basic level should be able to solve mathematical problems that require the direct application of concepts and procedures in familiar situations. For example, they should be able to perform computations with real numbers and estimate the results of numerical calculations. These students should also be able to estimate, calculate, and compare measures and identify and compare properties of two- and three-dimensional figures, and solve simple problems using two-dimensional coordinate geometry. At this level, students should be able to identify the source of bias in a sample and make inferences from sample results, calculate, interpret, and use measures of central tendency and compute simple probabilities. They should understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities. They should be able to solve problems involving linear relations using tables, graphs, or symbols; and solve linear equations involving one variable.

Number Properties and Operations

- perform computations with real numbers including common irrational numbers or the absolute value of numbers
- solve problems involving factorization and divisibility
- estimate the results of numerical calculations including square and cube roots of numbers, or very small and very large numbers

Measurement and Geometry

- recognize, define, and describe properties of two and three dimensional figures
- estimate, calculate, and compare measures of two and three dimensional figures
- draw or sketch a geometric figure from a description
- use the Pythagorean Theorem to solve problems in two dimensions
- solve problems in coordinate geometry (two dimensions)

Data Analysis and Probability

- evaluate a sample for bias and make inferences from sample results
- describe the impact of outliers on measures of central tendency and variability
- calculate, interpret, and use measures of central tendency and variability
- understand the use of correlation coefficients to describe the relation between two data sets
- compute simple probabilities
- distinguish between experimental and theoretical probability

Algebra

- understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities
- solve problems involving linear relations expressed in algebraic, verbal, tabular, or graphical forms
- solve linear equations in one variable
- perform basic operations on algebraic expressions
- recognize, describe, and extend arithmetic or geometric progressions

PROFICIENT

Students in the twelfth grade performing at the proficient level should be able to select strategies to solve problems and integrate concepts and procedures. These students should be able to interpret an argument, justify a mathematical process, and make comparisons dealing with a wide variety of mathematical tasks. They should also be able to perform calculations involving similar figures including right triangle trigonometry. They should understand and apply properties of geometric figures and relationships between figures in two and three dimensions. Students at this level should select and use appropriate units of measure as they apply formulas to solve problems. Students performing at this level should be able to use measures of central tendency and variability of distributions to make decisions and predictions; calculate combinations and permutations to solve problems, and understand the use of the normal distribution to describe real-world situations. Students performing at the proficient level should be able to identify, manipulate, graph, and apply linear, quadratic, exponential, and inverse functions ($y = k/x$); solve routine and non-routine problems involving functions expressed in algebraic, verbal, tabular, and graphical forms; and solve quadratic and rational equations in one variable and solve systems of linear equations.

Number Properties and Operations

write, rename, represent, or compare real numbers

- model and solve problems using proportions, percents, or absolute values
- solve problems involving factors, multiples, or prime factorization

Measurement and Geometry

- perform calculations involving similar figures including right triangle trigonometry
- understand and apply properties of geometric figures and relationships between figures in two and three dimensions, especially under transformations of the plane
- solve problems involving indirect measurement
- select and use appropriate units of measurement to solve problems and convert between different measurement systems
- represent problem situations with geometric models to solve mathematical and real world problems

Data Analysis and Probability

- analyze and interpret data presented in multiple formats
- describe, select, and use measures of central tendency and variability of distributions to make decisions and predictions
- calculate combinations and permutations to solve problems
- determine probabilities of independent and dependent events, and interpret them within a given context
- compare two or more data sets using measures of central tendency and variability
- make judgements about the appropriateness of different representations of data
- understand the use of the normal distribution to describe real-world situations

Algebra

- use algebraic functions and function notation to represent relationships or solve problems
- identify, manipulate, graph, and apply linear, quadratic, exponential, and inverse functions ($y = k/x$)
- write algebraic expressions, equations, or inequalities to model a given situation
- analyze, interpret, and translate among various relations represented in verbal, tabular, graphical, or algebraic form
- solve routine and non-routine problems involving functions expressed in algebraic, verbal, tabular, and graphical forms
- solve quadratic and rational equations in one variable and solve a system of linear equations

ADVANCED

Twelfth-grade students performing at the advanced level should demonstrate in-depth knowledge of the mathematical concepts and procedures represented in the framework. They can integrate knowledge to solve complex problems and justify and explain their thinking. These students should be able to analyze, make and justify mathematical arguments, and communicate their ideas clearly. Advanced level students should be able to describe the intersections of geometric figures in two and three dimensions, and use vectors to represent velocity and direction. They should also be able to describe the impact of linear transformations and outliers on measures of central tendency and variability; analyze predictions based on multiple data sets; and apply probability and statistical reasoning in more complex problems. Students performing at the advanced level should be able to solve or interpret systems of inequalities; and formulate a model for a complex situation (e.g., exponential growth and decay) and make inferences or predictions using the mathematical model.

Number Properties and Operations

- provide a mathematical justification involving numerical properties or complex relationships
- analyze the effect of an estimation method on the accuracy of results for very small and very large numbers in a given context
- describe, generalize, analyze, and solve problems requiring the integration of numerical properties and operations across content areas

Measurement and Geometry

- make, test, and validate conjectures about two and three dimensional figures
- determine appropriate accuracy of measurement in problem situations
- describe the intersections of geometric figures in two and three dimensions (e.g., conic sections as intersection of plane and cone)
- use vectors to represent velocity and direction

Data Analysis and Probability

- use or interpret a normal distribution for summarizing sets of data
- evaluate the characteristics of a good survey or of a well-designed experiment
- describe the impact of linear transformation and outliers on measures of central tendency and variability
- analyze predictions on multiple data sets
- apply probability and statistical reasoning in more complex problems

Algebra

- solve or interpret systems of inequalities
- formulate a model for a complex situation (e.g., exponential growth and decay, inverse relations) and make inferences or predictions using the model
- analyze or produce a deductive argument or mathematical justification in various content areas

Appendix B

Technical Advisors

List of Technical Advisors

Technical Advisory Committee on Standard Setting (TACSS)

Gregory Cizek
University of North Carolina-Chapel Hill

Barbara Dodd
University of Texas-Austin

John Dossey
Illinois State University (retired)

Robert Forsyth
University of Iowa (retired)

Mary Pitoniak
Educational Testing Service

Technical Advisory Team (TAT)

Robert Forsyth, University of Iowa (retired)
Nancy Petersen, ACT
James Sconing, ACT

Appendix C

Panelist Information

2005 NAEP Mathematics Achievement Level Setting
List of Panelists

Table C-1
Panelist Demographic Information, by Region

Nominee Type	Northeast	South	Midwest	West	Total
Teacher	7	4	4	2	17 (55%)
Nonteacher Educator	2	0	2	1	5 (16%)
General Public	1	2	2	4	9 (29%)
Total	10 (32%)	6 (19%)	8 (26%)	7 (23%)	31

Table C-2
Panelist Demographic Information, by Gender

Nominee Type	Male	Female	Total
Teacher	9	8	17
Nonteacher Educator	5	0	5
General Public	4	5	9
Total	18 (58%)	13 (42%)	31

Table C-3
Panelist Demographic Information, by Race/Ethnicity

Nominee Type	White	Black	Asian	Native	Hispanic	Total
Teacher						17
Nonteacher Educator						5
General Public						9
Total	22 (71%)	4 (13%)	2 (6%)	0 (0%)	3 (10%)	31

Appendix D

Agenda

AGENDA FOR THE 2005 GRADE 12 NAEP MATHEMATICS
ACHIEVEMENT LEVEL SETTING MEETING

November 12-15, 2004
St. Louis, Missouri

DAY 1 **Friday, November 12**	
Cupples Foyer 8:00 AM	Registration and Continental Breakfast
Salon B/C 8:30 AM	Welcome and Introductions, *Matt Schulz*, ACT
8:40 AM	General Orientation to the NAEP *Sharif Shakrani*, NAGB
9:30 AM	General Introduction to NAEP Achievement Level Setting Process, *Matt Schulz*
10:00 AM	Break
10:15 AM	NAEP Exam and Panelists' Introductions, *Matt Schulz*
Salon A 11:45 AM	LUNCH
Salon B/C 12:45 PM	Orientation to the Mapmark Method, *Matt Schulz*
1:45 PM	The NAEP Mathematics Framework, *Mary Jo Messenger*
2:30 PM	Break
2:45 PM	Whole Group KSA Review of Common Extended Constructed Response Items, *Jason Schwartz*, Pacific Metrics
4:15 PM	Table KSA Review of Remaining Extended Constructed Response Items, *Jason Schwartz*
5:15 PM	Evaluation #1
Concourse A/B 6:30 PM	Get-Acquainted Social Time
7:00 PM	DINNER
8:30 PM	Adjourn

DAY 2 **Saturday, November 13**	
Cupples Foyer 7:30 AM	Continental Breakfast
Salon B/C 8:00 AM	Continue Table KSA Review of Remaining Extended Constructed Response Items, *Jason Schwartz*
9:15 AM	Independent Review of Ordered Item Booklet, *Howard Mitzel*, Pacific Metrics
	Break as needed
Salon A 11:45 AM	LUNCH
Salon B/C 12:45 PM	Table Discussion of Ordered Item Booklet, *Howard Mitzel*
2:45 PM	Break
3:00 PM	ALD Presentation, *Mary Jo Messenger*
4:15 PM	Round 1 Bookmark Training, *Howard Mitzel*
4:45 PM	Round 1 Bookmarks
5:45 PM	Evaluation #2
6:00 PM	Adjourn

DAY 3
Sunday, November 14

Cupples Foyer
7:30 AM — Continental Breakfast

Salon B/C
8:00 AM — Feedback from Round 1, *Matt Schulz*
- Cut scores
- Domain scores
- Rater locations

Domain Task 1, *Matt Schulz*

9:30 AM — Break

9:45 AM — Domain Task 2, *Matt Schulz*

11:00 AM — Break

11:15 AM — Instructions for Round 2 Cut Score Recommendations, *Matt Schulz*

11:30 AM — Round 2 Cut Score Recommendations

Evaluation #3

Salon A
12:45 PM — LUNCH

Salon B/C
2:00 PM — Feedback from Round 2, *Matt Schulz*
- Cut scores
- Domain scores
- Rater locations

Whole Group Discussion

Rater Group Discussion

4:00 PM — Break

4:15 PM — Instructions for Round 3 Cut Score Recommendations, *Matt Schulz*

Round 3 Cut Score Recommendations

4:45 PM — Evaluation #4

5:00 PM — Adjourn

DAY 4
Monday, November 15

Cupples Foyer
8:00 AM — Continental Breakfast

Salon B/C
8:30 AM — Feedback from Round 3, *Matt Schulz*
- Cut scores
- Domain scores
- Rater locations
- Consequences data

Whole Group Discussion

9:00 AM — Instructions for Round 4 Cut Score Recommendations, *Matt Schulz*

Round 4 Cut Score Recommendations

Evaluation #5

Salon A
9:30 AM — BRUNCH

Salon B/C
10:30 AM — Feedback from Round 4, *Matt Schulz*
- Cut scores
- Domain scores
- Consequences data

Consequences Questionnaire

11:00 AM — Exemplar Item Ratings, *Matt Schulz*

12:15 PM — Evaluation #6

12:30 PM — Adjourn

Appendix E

Item Pool Information

Table E-1
Comparability of Group A and B Item Pools With Regard to Subscale Representation

	Full Item Pool for 2005 Assessment					Group A Pool					Group B Pool				
Block	1	2	3	4	Total	1	2	3	4	Total	1	2	3	4	Total
	3	6	3	5	17	3	6	3	5	17	3	6	3	5	17
	4	5	4	4	17	4	5	4	4	17	4	5	4	4	17
	4	8	3	3	18	4	8	3	3	18	0	0	0	0	0
	2	6	4	5	17	1	0	0	0	1	1	6	4	5	16
	2	6	3	7	18	1	0	0	0	1	1	6	3	7	17
	7	7	4	3	21	0	0	0	0	0	7	7	4	3	21
	4	6	3	5	18	4	6	3	5	18	0	0	0	0	0
	1	4	3	10	18	0	0	2	0	2	1	4	3	10	18
	2	6	4	6	18	0	6	4	5	15	2	0	0	1	3
	2	6	4	6	18	2	6	4	6	18	0	0	0	0	0
Sum	31	60	35	54	180	19	37	23	28	107	19	34	21	35	109
%	17%	33%	19%	30%	100%	18%	35%	21%	26%	100%	17%	31%	19%	32%	100%

Note. Subscales 1 to 4 stand for Number Properties and Operations, Measurement and Geometry, Data Analysis and Probability, and Algebra, respectively.

Table E-2
Comparability of Group A and B Item Pools With Regard to Item Types

	Full Item Pool for 2005 Assessment						Group A Pool						Group B Pool					
	Item				Step		Item				Step		Item				Step	
Block	MC	DI	Poly	Total[a]	Poly	Total[b]	MC	DI	Poly	Total	Poly	Total	MC	DI	Poly	Total	Poly	Total
	12	1	4	17	8	21	12	1	4	17	8	21	12	1	4	17	8	21
	11	2	4	17	12	25	11	2	4	17	12	25	11	2	4	17	12	25
	11	3	4	18	12	26	11	3	4	18	12	26	0	0	0	0	0	0
	14	1	2	17	4	19	1	0	0	1	0	1	13	1	2	16	4	18
	13	0	5	18	11	24	0	0	1	1	2	2	13	0	4	17	9	22
	10	4	7	21	18	32	0	0	0	0	0	0	10	4	7	21	18	32
	7	5	6	18	13	25	7	5	6	18	13	25	0	0	0	0	0	0
	11	5	2	18	6	22	1	1	0	2	0	2	11	5	2	18	6	22
	15	2	1	18	4	21	12	2	1	15	4	18	3	0	0	3	0	3
	15	1	2	18	6	22	15	1	2	18	6	22	0	0	0	0	0	0
Sum	119	24	37	180	94	237	70	15	22	107	57	142	73	13	23	109	57	143
%	66%	13%	21%	100%	40%	100%	65%	14%	21%	100%	40%	100%	67%	12%	21%	100%	40%	100%

Note. “MC”, “DI”, and “Poly” in table stand for “multiple-choice”, “dichotomously-scored constructed response”, and “polytomously-scored constructed response”. [a] is the sum of items for MC, DI and Poly. [b] is the sum of steps for MC, DI, and Poly.

Table E-3
Comparability of Group A and B Item Pools With Regard to Item/Step Scale Values Using 0.67 Mapping Criterion

Block	Full Item Pool for 2005 Assessment					Group A Pool					Group B Pool				
	N*	Mean	SD	Min	Max	N*	Mean	SD	Min	Max	N*	Mean	SD	Min	Max
	21	269	35	159	305	21	269	35	159	305	21	269	35	159	305
	25	279	37	215	401	25	279	37	215	401	25	279	37	215	401
	26	282	39	223	357	26	282	39	223	357	0	---	---	---	---
	19	285	32	212	330	1	212	NA	212	212	18	289	28	215	330
	24	276	39	141	337	2	193	74	141	245	22	283	26	230	337
	32	274	54	157	375	0	---	---	---	---	32	274	54	157	375
	25	281	31	214	328	25	281	31	214	328	0	---	---	---	---
	22	285	45	201	401	2	330	8	324	336	22	285	45	201	401
	21	278	28	236	339	18	279	30	236	339	3	276	21	252	291
	22	281	37	216	380	22	281	37	216	380	0	---	---	---	---
Overall:	237	279	39	141	401	142	278	37	141	401	143	279	40	157	401

*N = Number of steps or score points greater than zero.

Appendix F

Consequences Data Questionnaire

Table F-1
Responses to Consequences Questionnaire

1. Given your understanding of borderline student performance at each of the three achievement levels, do these percentages reflect your expectations about the proportions of students whose NAEP score would be at or above the cut score your group set for each of these achievement levels?

	ALS
❑ Yes (Please skip to Number 4.)	13
❑ No (Please continue to Number 2.)	18

2. Having seen the data on the percentages of students whose score on the NAEP was at or above the cut score your group set for each achievement level, would you change one or more of the achievement levels you have set if you could?

	ALS
❑ Yes (Please continue to Number 3.)	12
❑ No (Please skip to Number 4.)	6

3. Please mark the box corresponding to the response that indicates *how* you would *change the final cut scores* for each level. Changing the final cut scores would make these percentages more in line with your expectations about the proportions of students taking the Mathematics NAEP who would score at or above the cut score of each of the achievement levels. *You must give a cut score if you recommend a change.*

Basic	ALS
❑ Make no change. I am satisfied with the Basic cutscore.	3
❑ Raise the cut score for the Basic level so that a *smaller* percentage of students score at or above the Basic level. I want to raise the Basic cut score to _____.	4
❑ Lower the cut score for the Basic level so that a *larger* percentage of students would score at or above the Basic level. I want to lower the Basic cut score to _____.	6

Proficient	ALS
❑ Make no change. I am satisfied with the Proficient cutscore.	5
❑ Raise the cut score for the Proficient level so that a *smaller* percentage of students score at or above the Proficient level. I want to raise the Proficient cut score to _____.	3
❑ Lower the cut score for the Proficient level so that a *larger* percentage of students would score at or above the Proficient level. I want to lower the Proficient cut score to _____.	5

Advanced	ALS
❑ Make no change. I am satisfied with the Advanced cutscore.	4
❑ Raise the cut score for the Advanced level so that a *smaller* percentage of students score at or above the Advanced level. I want to raise the Advanced cut score to _____.	2
❑ Lower the cut score for the Advanced level so that a *larger* percentage of students would score at or above the Advanced level. I want to lower the Advanced cut score to _____.	7

4. What recommendations do you wish to make to the National Assessment Governing Board regarding the cut scores set for achievement levels?

	ALS
❑ I recommend that the achievement levels be reported as set.	19
❑ I recommend changes consistent with my answers above. If you wish, comment on the magnitude of change you would recommend.	11
No Response	1

Appendix G

Process Evaluation Questionnaires

Table G-1
Results from Process Evaluation Questionnaire No. 1

Question	5	4	3	2	1	Mean Score
1. The advance materials I received were adequate to prepare me to fulfill my role in this meeting:	Totally Agree		Somewhat Agree		Totally Disagree	
	13	8	8	2	0	4.03
2. The organization of the advance materials I received for this meeting was:	Very Good		Acceptable		Very Poor	
	20	5	4	2	0	4.39
3. The amount of time allocated for the General Orientation to the NAEP Program was:	Far Too Long		About Right		Far Too Short	
	3	6	21	1	0	3.35
4. The explanation of the development of the NAEP in general was:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	11	17	3	0	0	4.26
5. The explanation of the development of the Mathematics NAEP was:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	14	14	1	1	0	4.37
6. The major organizations involved in NAEP and the roles of each was:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	14	10	7	0	0	4.23
7. I understand the purpose of the NAEP achievement level setting meeting.	Totally Agree		Somewhat Agree		Totally Disagree	
	15	12	4	0	0	4.35
8. The amount of time allocated for the General Introduction to the NAEP achievement level setting process was:	Far Too Long		About Right		Far Too Short	
	4	7	18	2	0	3.42
9. I believe my perspectives and experiences will be important in the NAEP standard setting process.	Totally Agree		Somewhat Agree		Totally Disagree	
	14	11	5	1	0	4.23
10. I understand the difference between criterion-referenced and norm-referenced standards.	Totally Agree		Somewhat Agree		Totally Disagree	
	23	3	4	0	0	4.63
11. I will not allow my judgments in this meeting to be influenced by my personal feelings about the No Child Left Behind (NCLB) law.	Totally Agree		Somewhat Agree		Totally Disagree	
	25	2	3	0	0	4.73
12. Taking the Mathematics NAEP was an informative experience.	Totally Agree		Somewhat Agree		Totally Disagree	
	24	5	1	0	0	4.77
13. Taking the Mathematics NAEP gave me a good idea of what is expected of students.	Totally Agree		Somewhat Agree		Totally Disagree	
	22	6	2	0	0	4.67

14. The amount of time allocated for the Mapmark method orientation was:	Far Too Long 2	3	About Right 19	6	Far Too Short 0	3.03
15. The overview of the method to be followed in this meeting was:	Absolutely Clear 5	16	Somewhat Clear 5	3	Not at All Clear 0	3.79
16. The explanation of how an item map is constructed was:	Absolutely Clear 8	10	Somewhat Clear 9	3	Not at All Clear 0	3.77
17. I think I will be comfortable using a 2/3 or 0.67 probability to interpret the location of an item on my map.	Totally Agree 12	13	Somewhat Agree 5	0	Totally Disagree 0	4.23
18. The explanation of the information in my Ordered Item Booklet (OIB) was:	Absolutely Clear 10	13	Somewhat Clear 6	1	Not at All Clear 0	4.07
19. The amount of time allocated for the Framework presentation was:	Far Too Long 3	3	About Right 21	3	Far Too Short 0	3.20
20. The presentation of the Mathematics Framework was:	Absolutely Clear 8	15	Somewhat Clear 6	1	Not at All Clear 0	4.00
21. The presentation of the Mathematics Framework had about the right level of detail.	Totally Agree 7	14	Somewhat Agree 6	3	Totally Disagree 0	3.83
22. The amount of time allocated for the whole group KSA review was:	Far Too Long 1	0	About Right 18	10	Far Too Short 2	2.61
23. The instructions on what I was to do in the KSA review were:	Absolutely Clear 9	14	Somewhat Clear 3	4	Not at All Clear 1	3.84
24. My understanding of our tasks in the KSA review was:	Totally Adequate 13	10	Somewhat Adequate 5	2	Totally Inadequate 1	4.03
25. The whole group work on the common constructed response items was:	Very Useful 17	7	Somewhat Useful 5	1	Not at All Useful 1	4.23

Table G-2
Results from Process Evaluation Questionnaire No. 2

Question	5	4	3	2	1	Mean Score
1. The amount of time allocated for the table group KSA review was:	Far Too Long		About Right		Far Too Short	
	2	4	17	5	2	2.97
2. The table group review of the remaining constructed response items was:	Very Useful		Somewhat Useful		Not at All Useful	
	19	4	6	1	0	4.37
3. I understand the score levels of polytomous items.	Totally Agree		Somewhat Agree		Totally Disagree	
	13	11	5	1	1	4.10
4. The amount of time allocated for the independent OIB review was:	Far Too Long		About Right		Far Too Short	
	0	3	22	5	1	2.87
5. The instructions on what I was to do for the independent OIB review were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	16	11	3	1	0	4.35
6. I understood how to use my item map and ordered item booklet.	Totally Agree		Somewhat Agree		Totally Disagree	
	20	8	1	0	2	4.42
7. I was comfortable working through the ordered item booklet on my own.	Totally Agree		Somewhat Agree		Totally Disagree	
	19	7	3	2	0	4.39
8. The ordering of the items in the ordered item booklet agreed with my perceptions of the relative difficulty of the items.	Totally Agree		Somewhat Agree		Totally Disagree	
	4	11	14	2	0	3.55
9. The KSA work with the OIB helped me understand what can make one item harder than others.	Totally Agree		Somewhat Agree		Totally Disagree	
	10	10	10	1	0	3.94
10. The amount of time allocated for the table discussion of the OIB was:	Far Too Long		About Right		Far Too Short	
	1	0	24	5	0	2.90
11. The instructions on we were to do in the table discussion of the OIB were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	12	11	7	0	0	4.17
12. The table discussion of the ordered item booklet was:	Very Useful		Somewhat Useful		Not at All Useful	
	16	9	5	0	0	4.37

13. I feel I made a valuable contribution to my table group's discussion.	Totally Agree		Somewhat Agree		Totally Disagree	
	13	11	4	1	1	4.13
14. I feel my perspective is being heard by others in my table group.	Totally Agree		Somewhat Agree		Totally Disagree	
	20	6	3	1	0	4.50
15. I feel that I was being pressured to agree with others in my table group.	Totally Agree		Somewhat Agree		Totally Disagree	
	0	1	0	5	24	1.27
16. The amount of time allocated for the ALD presentation was:	Far Too Long		About Right		Far Too Short	
	4	4	22	1	0	3.35
17. The ALDs appear to be reasonably complete and comprehensive statements of what students should know and be able to do at each level of achievement.	Totally Agree		Somewhat Agree		Totally Disagree	
	9	11	6	3	2	3.71
18. My own level of satisfaction with the Basic achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	10	10	9	2	0	3.90
19. My own level of satisfaction with the Proficient achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	10	10	7	3	1	3.81
20. My own level of satisfaction with the Advanced achievement level description is:	Very Satisfied		Somewhat Satisfied		Not at All Satisfied	
	9	11	8	1	2	3.77
21. At the time I provided the Round 1 bookmark placements, my understanding of the Basic achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	8	13	9	1	0	3.90
22. At the time I provided the Round 1 bookmark placements, my understanding of the Proficient achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	7	11	10	2	0	3.77
23. At the time I provided the Round 1 bookmark placements, my understanding of the Advanced achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	7	12	8	2	1	3.73
24. I was comfortable using the concept of performance at the lower borderline of Basic.	Totally Agree		Somewhat Agree		Totally Disagree	
	10	10	8	3	0	3.87
25. I was comfortable using the concept of performance at the lower borderline of Proficient.	Totally Agree		Somewhat Agree		Totally Disagree	
	10	9	8	4	0	3.81

26. I was comfortable using the concept of performance at the lower borderline of Advanced.	Totally Agree	10	9	Somewhat Agree	9	3	Totally Disagree	0	3.84
27. I believe my Round 1 bookmark placements are consistent with the achievement level descriptions.	Totally Agree	10	10	Somewhat Agree	10	1	Totally Disagree	0	3.94
28. The amount of time allocated for placing the bookmarks was:	Far Too Long	1	2	About Right	23	3	Far Too Short	1	2.97
29. The instructions on how I was to place my bookmarks were:	Absolutely Clear	15	12	Somewhat Clear	2	0	Not at All Clear	0	4.45
30. My understanding of how to use the ALDs to choose my bookmarks was:	Totally Adequate	11	13	Somewhat Adequate	6	0	Totally Inadequate	0	4.17
31. The most accurate description of my level of confidence in my Round 1 bookmark placements is:	Totally Confident	4	8	Somewhat Confident	10	6	Not at All Confident	1	3.28
32. I felt pressure to recommend bookmarks that were close to those recommended by other panelists.	Totally Agree	0	2	Somewhat Agree	1	3	Totally Disagree	24	1.37
33. I was comfortable using a 0.67 probability to define "mastery" in placing my bookmarks.	Totally Agree	12	10	Somewhat Agree	6	3	Totally Disagree	0	4.00
34. The KSAs required by the items around my bookmarks seemed to be appropriate for the borderline of the corresponding achievement level description.	Totally Agree	7	13	Somewhat Agree	10	0	Totally Disagree	0	3.90

Table G-3
Results from Process Evaluation Questionnaire No. 3

Question	5	4	3	2	1	Mean Score
1. I understood the Round 1 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	21	7	3	0	0	4.58
2. I understood what students at the Round 1 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree	
	18	9	4	0	0	4.45
3. I understood the Rater Location Feedback (where my Round 1 cut scores were in comparison to the Round 1 median cut scores).	Totally Agree		Somewhat Agree		Totally Disagree	
	22	8	1	0	0	4.68
4. I understood the domain score feedback (in the Percent Correct Table).	Totally Agree		Somewhat Agree		Totally Disagree	
	19	10	2	0	0	4.55
5. The amount of time allocated for our work with domains was:	Far Too Long		About Right		Far Too Short	
	1	3	23	4	0	3.03
6. The instructions I received for our domain score tasks were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	11	10	9	1	0	4.00
7. As a result of performing the domain score tasks, I now understand the concept of domain scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	15	10	4	1	0	4.30
8. I was comfortable thinking about whether an item was like other items in its domain (Domain Task 1).	Totally Agree		Somewhat Agree		Totally Disagree	
	19	7	4	0	1	4.39
9. The most accurate description of my level of confidence in deciding whether domain scores should be higher/OK/lower (Domain Task 2), relative to the ALDs.	Totally Confident		Somewhat Confident		Not at All Confident	
	5	18	6	2	0	3.84
10. I understood how to use the Domain Item Maps.	Totally Agree		Somewhat Agree		Totally Disagree	
	15	11	2	2	1	4.19
11. I understood how to use the Domain Ordered Item Booklet.	Totally Agree		Somewhat Agree		Totally Disagree	
	19	10	1	1	0	4.52

12. I understood how to use the Domain Score Chart.	Totally Agree		Somewhat Agree		Totally Disagree	
	16	11	4	0	0	4.39
13. At the time I provided the Round 2 cut score recommendations, my understanding of the Basic achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	17	8	6	0	0	4.35
14. At the time I provided the Round 2 cut score recommendations, my understanding of the Proficient achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	18	7	6	0	0	4.39
15. At the time I provided the Round 2 cut score recommendations, my understanding of the Advanced achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	17	8	5	0	1	4.29
16. At the time I provided the Round 2 cut score recommendations, my concept of the lower borderline performance at the Basic level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	14	11	6	0	0	4.26
17. At the time I provided the Round 2 cut score recommendations, my concept of the lower borderline performance at the Proficient level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	14	11	6	0	0	4.26
18. At the time I provided the Round 2 cut score recommendations, my concept of the lower borderline performance at the Advanced level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	13	13	5	0	0	4.26
19. I understand the difference between borderline performance and typical performance within an achievement level.	Totally Agree		Somewhat Agree		Totally Disagree	
	19	9	3	0	0	4.52
20. I believe my Round 2 cut score recommendations are consistent with the achievement level descriptions.	Totally Agree		Somewhat Agree		Totally Disagree	
	10	15	4	1	0	4.13
21. The amount of time allocated for my Round 2 cut score recommendations was:	Far Too Long		About Right		Far Too Short	
	5	7	15	3	1	3.39
22. The instructions I received for recommending Round 2 cut scores were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	14	9	7	1	0	4.16
23. My level of understanding of how I was to choose cut scores for Round 2 was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	15	6	9	1	0	4.13
24. The most accurate description of my level of confidence in my Round 2 cut score recommendations is:	Totally Confident		Somewhat Confident		Not at All Confident	
	7	13	8	3	0	3.77

25. I felt pressure to recommend cut scores that were close to those recommended by other panelists.	Totally Agree		Somewhat Agree		Totally Disagree	
	2	2	3	2	22	1.71
26. I was comfortable choosing scale values instead of placing bookmarks to recommend cut scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	11	10	6	4	0	3.90

Table G-4
Results from Process Evaluation Questionnaire No. 4

Question	5	4	3	2	1	Mean Score
1. I understood the Round 2 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	22	8	1	0	0	4.68
2. I understood what students at the Round 2 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree	
	15	15	1	0	0	4.45
3. I understood the Rater Location Feedback (where my Round 2 cut scores were in comparison to the Round 2 median cut scores).	Totally Agree		Somewhat Agree		Totally Disagree	
	21	10	0	0	0	4.68
4. I understood the domain score feedback (in the Percent Correct Table).	Totally Agree		Somewhat Agree		Totally Disagree	
	16	15	0	0	0	4.52
5. The amount of time allocated for the rater group discussion was:	Far Too Long		About Right		Far Too Short	
	2	3	23	1	1	3.13
6. The instructions on what I was to do in the rater group discussion were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	14	9	7	1	0	4.16
7. I would characterize the rater group discussions as extensive.	Totally Agree		Somewhat Agree		Totally Disagree	
	9	10	9	3	0	3.81
8. I would characterize the rater group discussions as balanced and reasoned.	Totally Agree		Somewhat Agree		Totally Disagree	
	12	6	10	1	1	3.90
9. I would characterize the rater group discussions as helpful and informative.	Totally Agree		Somewhat Agree		Totally Disagree	
	13	10	7	1	0	4.13
10. At the time I provided the Round 3 cut score recommendations, my understanding of the Basic achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	16	13	2	0	0	4.45
11. At the time I provided the Round 3 cut score recommendations, my understanding of the Proficient achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	15	14	2	0	0	4.42
12. At the time I provided the Round 3 cut score recommendations, my understanding of the Advanced achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	16	13	2	0	0	4.45

13. At the time I provided the Round 3 cut score recommendations, my concept of the lower borderline performance at the Basic level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	16	9	5	0	0	4.37
14. At the time I provided the Round 3 cut score recommendations, my concept of the lower borderline performance at the Proficient level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	16	11	4	0	0	4.39
15. At the time I provided the Round 3 cut score recommendations, my concept of the lower borderline performance at the Advanced level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	17	10	3	0	0	4.47
16. I understand the difference between borderline performance and typical performance within an achievement level.	Totally Agree		Somewhat Agree		Totally Disagree	
	21	7	3	0	0	4.58
17. My Round 3 cut score recommendations are consistent with the achievement level descriptions.	Totally Agree		Somewhat Agree		Totally Disagree	
	18	10	3	0	0	4.48
18. The instructions I received for recommending Round 3 cut scores were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	15	11	5	0	0	4.32
19. My level of understanding of how I was to choose cut scores for Round 3 was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	16	12	3	0	0	4.42
20. The most accurate description of my level of confidence in my Round 3 cut score recommendations is:	Totally Confident		Somewhat Confident		Not at All Confident	
	10	17	4	0	0	4.19
21. I felt pressure to recommend cut scores that were close to those recommended by other panelists.	Totally Agree		Somewhat Agree		Totally Disagree	
	1	1	2	2	24	1.43

Table G-5
Results from Process Evaluation Questionnaire No. 5

Question	5	4	3	2	1	Mean Score
1. I understood the Round 3 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	23	5	2	0	0	4.70
2. I understood what students at the Round 3 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree	
	19	9	2	0	0	4.57
3. I understood the Rater Location Feedback (where my Round 3 cut scores were in comparison to the Round 3 median cut scores).	Totally Agree		Somewhat Agree		Totally Disagree	
	22	6	1	0	0	4.72
4. I understood the domain score feedback (in the Percent Correct Table).	Totally Agree		Somewhat Agree		Totally Disagree	
	21	8	1	0	0	4.67
5. I understood the consequences data.	Totally Agree		Somewhat Agree		Totally Disagree	
	22	7	1	0	0	4.70
6. The instructions I received for using consequences data during Round 4 were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	19	6	5	0	0	4.47
7. The amount of time allocated for discussing the consequences data was:	Far Too Long		About Right		Far Too Short	
	3	2	23	2	0	3.20
8. The most accurate description of my level of confidence in using the consequences data to recommend cut scores in Round 4.	Totally Confident		Somewhat Confident		Not at All Confident	
	14	11	5	0	0	4.30
9. At the time I provided the Round 4 cut score recommendations, my understanding of the Basic achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	18	9	3	0	0	4.50
10. At the time I provided the Round 4 cut score recommendations, my understanding of the Proficient achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	17	10	3	0	0	4.47
11. At the time I provided the Round 4 cut score recommendations, my understanding of the Advanced achievement level description was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	17	10	3	0	0	4.47

12. At the time I provided the Round 4 cut score recommendations, my concept of the lower borderline performance at the Basic level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	19	8	2	0	0	4.59
13. At the time I provided the Round 4 cut score recommendations, my concept of the lower borderline performance at the Proficient level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	18	10	2	0	0	4.53
14. At the time I provided the Round 4 cut score recommendations, my concept of the lower borderline performance at the Advanced level was:	Very Well Formed		Moderately Formed		Not Well Formed	
	18	10	1	1	0	4.50
15. I understand the difference between borderline performance and typical performance within an achievement level.	Totally Agree		Somewhat Agree		Totally Disagree	
	18	8	4	0	0	4.47
16. I believe my Round 4 cut score recommendations are consistent with the achievement level descriptions.	Totally Agree		Somewhat Agree		Totally Disagree	
	21	7	2	0	0	4.63
17. The instructions I received for recommending Round 4 cut scores were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	18	10	2	0	0	4.53
18. My level of understanding of how I was to choose cut scores for Round 4 was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	18	10	2	0	0	4.53
19. The most accurate description of my level of confidence in my Round 4 cut score recommendations is:	Totally Confident		Somewhat Confident		Not at All Confident	
	17	9	4	0	0	4.43
20. I felt pressure to recommend cut scores that were close to those recommended by other panelists.	Totally Agree		Somewhat Agree		Totally Disagree	
	2	1	3	2	22	1.63

Table G-6
Results from Final Process Evaluation Questionnaire (No. 6)

Question	5	4	3	2	1	Mean Score
1. I understood the Round 4 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree	
	23	6	1	0	0	4.73
2. I understood what students at the Round 4 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree	
	21	8	1	0	0	4.67
3. I understood the domain score feedback (in the Percent Score Table).	Totally Agree		Somewhat Agree		Totally Disagree	
	22	7	1	0	0	4.70
4. The amount of time allocated for the Consequences Questionnaire was:	Far Too Long		About Right		Far Too Short	
	3	1	25	1	0	3.20
5. I understood the Round 4 consequences data.	Totally Agree		Somewhat Agree		Totally Disagree	
	19	5	3	1	0	4.50
6. The instructions I received for completing the Consequences Questionnaire were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	18	9	1	1	0	4.52
7. I understood how to complete the Consequences Questionnaire.	Totally Agree		Somewhat Agree		Totally Disagree	
	21	4	4	1	0	4.50
8. The instructions on what I was to do during each round were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	12	13	3	2	0	4.17
9. My understanding of the tasks I was to accomplish during each round was:	Totally Adequate		Somewhat Adequate		Totally Inadequate	
	14	12	2	2	0	4.27
10. The most accurate description of my level of confidence in the cut score recommendations I provided was:	Totally Confident		Somewhat Confident		Not at All Confident	
	14	14	1	1	0	4.37
11. The amount of time I had to complete the tasks I was to accomplish during each round was:	Far Too Long		About Right		Far Too Short	
	2	2	21	5	0	3.03
12. I would describe the effectiveness of this achievement level setting method as:	Highly Effective		Somewhat Effective		Not at All Effective	
	15	9	3	2	0	4.28
13. I was comfortable choosing scale values to represent my judgments instead of page numbers.13. I felt my input was valued and considered by others in my group.	To a Great Extent		Somewhat		Not at All	
	16	9	0	2	1	4.32

14. I felt pressured by others in my group to make my cut score recommendations agree with theirs.	To a Great Extent		Somewhat		Not at All	
	1	0	2	2	25	1.33
15. I felt pressured by staff to make cut score recommendations higher or lower.	To a Great Extent		Somewhat		Not at All	
	1	0	0	1	28	1.17
16. I felt pressured by staff to keep my cut score recommendations the same.	To a Great Extent		Somewhat		Not at All	
	1	0	0	1	28	1.17
17. The amount of time allocated for the Exemplar Item Rating Task was:	Far Too Long		About Right		Far Too Short	
	2	2	25	0	1	3.13
18. The instructions I received for the Exemplar Item Rating Task were:	Absolutely Clear		Somewhat Clear		Not at All Clear	
	13	6	4	4	0	4.04
19. My understanding of how I was to perform the Exemplar Item Rating Task was:	Totally Adequate		Marginally Adequate		Totally Inadequate	
	16	8	4	1	0	4.34
20. I believe the exemplar items will be useful for describing the achievement levels.	Totally Agree		Somewhat Agree		Totally Disagree	
	11	10	7	1	0	4.07
21. The exemplar items I reviewed seemed appropriately matched to their achievement level.	Totally Agree		Somewhat Agree		Totally Disagree	
	4	12	9	4	0	3.55
22. I understood the purpose of this meeting.	Totally Agree		Somewhat Agree		Totally Disagree	
	24	6	0	0	0	4.80
23. I feel that this ALS process provided me an opportunity to use my best judgment to recommend cut scores for the NAEP mathematics assessment.	To a Great Extent		Somewhat		Not at All	
	19	9	2	0	0	4.57
24. I feel that this ALS process has produced achievement levels that are defensible.	To a Great Extent		Somewhat		Not at All	
	14	14	2	0	0	4.40
25. I feel that this ALS process has produced achievement levels that will generally be considered reasonable.	To a Great Extent		Somewhat		Not at All	
	14	11	5	0	0	4.30
26. I believe that the achievement levels capture meaningful distinctions in mathematics performance as described in the ALDs.	Totally Agree		Somewhat Agree		Totally Disagree	
	9	15	4	1	1	4.00

27. I feel that the panel in this meeting is widely inclusive of groups that should have a say in setting NAEP achievement levels.	To a Great Extent		Somewhat		Not at All	
	19	5	3	1	0	4.50
28. I feel that the panelists in this meeting are appropriately qualified for setting NAEP achievement levels.	To a Great Extent		Somewhat		Not at All	
	22	5	1	1	0	4.66
29. I would be willing to sign a statement (after reading it of course) recommending the use of the cut scores resulting from this ALS process.		Yes, definitely	Yes, probably	No, probably	No, definitely	
		19	9	1	0	3.62
30. Having observers present influenced my judgments.	To a Great Extent		Somewhat		Not at All	
	0	0	0	2	27	1.07
31. During the ALS process, I found the Achievement Level Descriptions:	Very Helpful		Somewhat Helpful		Not at All Helpful	
	17	8	2	2	0	4.38
32. During the ALS process, I found the Ordered Item Booklet:	Very Helpful		Somewhat Helpful		Not at All Helpful	
	22	7	0	0	0	4.76
33. During the ALS process, I found the Primary Item Map:	Very Helpful		Somewhat Helpful		Not at All Helpful	
	16	7	3	3	0	4.24
34. During the ALS process, I found the Domain-Ordered Item Maps:	Very Helpful		Somewhat Helpful		Not at All Helpful	
	17	5	4	3	0	4.24
35. During the ALS process, I found the Rater Location Data (the location of my cut scores relative to the median cut scores):	Very Helpful		Somewhat Helpful		Not at All Helpful	
	16	9	3	0	0	4.46
36. During the ALS process, I found the Domain Score Feedback (in the Percent Correct Table):	Very Helpful		Somewhat Helpful		Not at All Helpful	
	14	10	3	1	1	4.21
37. During the ALS process, I found the Consequences Data:	Very Helpful		Somewhat Helpful		Not at All Helpful	
	13	9	3	1	2	4.07
38. How would you rate the amount of personal attention and assistance you received from the process facilitator?	Too Much		About Right		Too Little	
	0	3	26	0	0	3.10
39. How would you rate the amount of personal attention and assistance you received from the content staff?	Too Much		About Right		Too Little	
	0	3	25	1	0	3.07

Appendix H

ALD Evaluation Task

Table H-1
ALD Statement Ratings

Basic

Statement	*Did you see test items related to this statement?* Yes	No	Not Sure	If you would modify this statement, please indicate how.
Twelfth-grade students performing at the basic level should be able to solve mathematical problems that require the direct application of concepts and procedures in familiar situations.	21	0	0	
For example, they should be able to perform computations with real numbers and estimate the results of numerical calculations.	21	0	0	
These students should also be able to estimate, calculate, and compare measures and identify and compare properties of two- and three-dimensional figures, and solve simple problems using two-dimensional coordinate geometry.	21	0	0	
At this level, students should be able to identify the source of bias in a sample and make	19	1	0	
They should understand the use of variables, expressions, and equations to represent unknown quantities and relationships among unknown quantities.	21	0	0	
They should be able to solve problems involving linear relations using tables, graphs, or symbols; and solve linear equations involving one variable.	20	0	0	
Totals	123	1	0	

Table H-2
ALD Statement Ratings

Proficient

Statement	*Did you see test items related to this statement?* Yes	No	Not Sure	If you would modify this statement, please indicate how.
Students in the twelfth grade performing at the proficient level should be able to select strategies to solve problems and integrate concepts and procedures.	21	0	0	
These students should be able to interpret an argument, justify a mathematical process, and make comparisons dealing with a wide variety of mathematical tasks.	21	0	0	
They should also be able to perform calculations involving similar figures including right triangle trigonometry.	21	0	0	
They should understand and apply properties of geometric figures and relationships between figures in two and three dimensions.	21	0	0	
Students at this level should select and use appropriate units of measure as they apply formulas to solve problems.	20	1	0	
Students performing at this level should be able to use measures of central tendency and variability of distributions to make decisions and predictions; calculate combinations and permutations to solve problems, and understand the use of the normal distribution to describe real-world situations.	19	1	0	
Students performing at the proficient level should be able to identify, manipulate, graph, and apply linear, quadratic, exponential, and inverse functions (y = k/x);	21	0	0	"inverse variations" rather than "inverse functions"
solve routine and non-routine problems involving functions expressed in algebraic, verbal, tabular, and graphical forms;	21	0	0	
and solve quadratic and rational equations in one variable and solve systems of linear equations.	20	0	1	
Totals	185	2	1	

Table H-3
ALD Statement Ratings

Advanced

Statement	*Did you see test items related to this statement?* Yes	No	Not Sure	If you would modify this statement, please indicate how.
Twelfth-grade students performing at the advanced level should demonstrate in-depth knowledge of the mathematical concepts and procedures represented in the framework.	21	0	0	
They can integrate knowledge to solve complex problems and justify and explain their thinking.	21	0	0	
These students should be able to analyze, make and justify mathematical arguments, and communicate their ideas clearly.	21	0	0	
Advanced level students should be able to describe the intersections of geometric figures in two and three dimensions, and use vectors to represent velocity and direction.	13	3	3	
They should also be able to describe the impact of linear transformations and outliers on measures of central tendency and variability; analyze predictions based on multiple data sets;	16	3	1	
and apply probability and statistical reasoning in more complex problems.	18	2	1	
Students performing at the advanced level should be able to solve or interpret systems of inequalities; and formulate a model for a complex situation (e.g., exponential growth and decay)	14	5	1	
and make inferences or predictions using the mathematical model.	18	1	1	
Totals	142	14	7	

Appendix I

Exemplar Item Ratings

Table I-1
Comparison of 2005 NAEP Mathematics Pilot Study and ALS
Exemplar Item Ratings

Basic

Item	OIB Page # Group A	OIB Page # Group B	Probability of Success	Rating as Exemplar: Very Good Pilot	Very Good **ALS**	OK Pilot	OK **ALS**	Do Not Use Pilot	Do Not Use **ALS**
M1	2	2	0.83	38%	**53%**	43%	**33%**	19%	**13%**
M7	6	8	0.94	24%	**53%**	67%	**33%**	10%	**13%**
M9	7	H-1	0.93	--	**67%**	--	**27%**	--	**7%**
P21_1	8	H-2	0.94	--	**43%**	--	**33%**	--	**23%**
M11	9	H-3	0.92	--	**27%**	--	**27%**	--	**47%**
M12	11	12	0.83	62%	**70%**	29%	**30%**	10%	**0%**
M13	12	13	0.89	48%	**70%**	52%	**17%**	0%	**13%**
M14	13	14	0.95	48%	**53%**	48%	**20%**	5%	**27%**
M17	17	H-4	0.86	--	**43%**	--	**43%**	--	**13%**
M22	23	21	0.80	52%	**70%**	48%	**30%**	0%	**0%**
P3_1	24	22	0.80	62%	**57%**	38%	**23%**	0%	**20%**
P36_1	25	24	0.75	29%	**10%**	43%	**47%**	29%	**43%**
P9_1	29	26	0.76	29%	**23%**	33%	**23%**	38%	**53%**
P6_1	33	30	0.71	62%	**50%**	33%	**33%**	0%	**17%**
P36_2	35	31	0.70	33%	**13%**	48%	**67%**	19%	**20%**
M33	36	33	0.70	38%	**47%**	48%	**53%**	10%	**0%**
P21_2	40	H-5	0.69	--	**27%**	--	**53%**	--	**20%**

Notes: Item handles in yellow fail to meet criteria suggested by ACT based on ratings by ALS panelists (≥ 33% "Very good" and <33% "do not use").
Items in blue are polytomously-scored items that were not eliminated and appear at more than one achievement level.

Table I-2
Comparison of 2005 NAEP Mathematics Pilot Study and ALS
Exemplar Item Ratings

Proficient

Item	OIB Page # Group A	OIB Page # Group B	Probability of Success	Rating as Exemplar: Very Good Pilot	Very Good **ALS**	OK Pilot	OK **ALS**	Do Not Use Pilot	Do Not Use **ALS**
M40	41	40	0.90	71%	**33%**	24%	**53%**	5%	**13%**
P9_2	44	41	0.95	33%	**7%**	38%	**43%**	29%	**50%**
D7	45	43	0.88	38%	**47%**	48%	**17%**	14%	**37%**
M44	47	44	0.95	52%	**40%**	43%	**20%**	5%	**40%**
M46	49	H-6	0.82	--	**37%**	--	**47%**	--	**17%**
M49	51	48	0.91	29%	**40%**	48%	**23%**	24%	**37%**
P3_2	53	51	0.86	33%	**37%**	62%	**30%**	5%	**33%**
P5_1	55	52	0.89	33%	**33%**	33%	**30%**	33%	**37%**
M55	57	54	0.85	33%	**47%**	62%	**50%**	5%	**3%**
P20_1	58	55	0.87	0%	**13%**	33%	**27%**	67%	**60%**
P9_3	59	56	0.90	33%	**10%**	67%	**50%**	0%	**40%**
M56	60	H-7	0.89	--	**57%**	--	**30%**	--	**13%**
M60	62	61	0.86	38%	**33%**	48%	**53%**	14%	**13%**
M66	66	H-8	0.83	--	**70%**	--	**30%**	--	**0%**
P5_2	67	67	0.81	29%	**23%**	43%	**50%**	29%	**27%**
P17_1	68	68	0.82	19%	**30%**	43%	**30%**	5%	**40%**
P6_2	71	70	0.76	62%	**43%**	38%	**37%**	0%	**20%**
M68	72	71	0.78	67%	**73%**	33%	**27%**	0%	**0%**
M69	73	H-9	0.78	--	**67%**	--	**23%**	--	**10%**
M70	75	72	0.75	67%	**30%**	29%	**47%**	5%	**23%**
P21_3	76	H-10	0.80	--	**50%**	--	**43%**	--	**7%**
M73	78	H-11	0.74	--	**40%**	--	**47%**	--	**13%**
M74	80	76	0.81	81%	**53%**	19%	**40%**	0%	**7%**
M79	81	H-12	0.75	--	**63%**	--	**30%**	--	**7%**
M80	82	80	0.74	67%	**57%**	10%	**43%**	24%	**0%**
P19_1	84	H-13	0.74	--	**53%**	--	**37%**	--	**10%**
P15_1	85	81	0.78	48%	**37%**	52%	**50%**	0%	**13%**
P9_4	86	82	0.72	48%	**27%**	29%	**40%**	24%	**33%**
D11	89	87	0.75	38%	**57%**	57%	**40%**	5%	**3%**
D14	96	H-14	0.67	--	**60%**	--	**37%**	--	**3%**
M91	98	H-15	0.68	--	**40%**	--	**53%**	--	**7%**

Notes: Item handles in yellow fail to meet criteria suggested by ACT based on ratings by ALS panelists ($\geq$ 33% “Very good” and <33% “do not use”).

Table I-3
Comparison of 2005 NAEP Mathematics Pilot Study and ALS
Exemplar Item Ratings

Advanced

Item	OIB Page # Group A	OIB Page # Group B	Probability of Success	Rating as Exemplar: Very Good Pilot	Very Good **ALS**	OK Pilot	OK **ALS**	Do Not Use Pilot	Do Not Use **ALS**
M92	99	H-16	0.95	--	**30%**	--	**40%**	--	**30%**
P36_3	100	94	0.79	29%	**20%**	38%	**33%**	33%	**47%**
P15_2	101	95	0.91	43%	**47%**	57%	**47%**	0%	**7%**
M95	103	98	0.90	67%	**47%**	33%	**53%**	0%	**0%**
M94	104	100	0.87	24%	**23%**	19%	**37%**	57%	**40%**
P17_2	106	104	0.85	29%	**23%**	19%	**23%**	52%	**53%**
M97	107	105	0.88	38%	**23%**	48%	**40%**	14%	**37%**
M100	109	H-17	0.89	--	**37%**	--	**37%**	--	**27%**
M102	110	111	0.88	67%	**28%**	24%	**45%**	10%	**28%**
M101	112	112	0.88	43%	**47%**	38%	**33%**	19%	**20%**
M104	114	H-18	0.89	--	**23%**	--	**47%**	--	**30%**
P19_2	115	H-19	0.77	--	**33%**	--	**47%**	--	**20%**
P21_4	116	H-20	0.77	--	**50%**	--	**33%**	--	**17%**
P20_2	117	113	0.77	38%	**17%**	10%	**30%**	52%	**53%**
M107	119	H-21	0.71	--	**37%**	--	**40%**	--	**23%**
M108	121	115	0.80	86%	**43%**	14%	**47%**	0%	**10%**
M111	124	123	0.75	62%	**63%**	19%	**33%**	19%	**3%**
D18	129	129	0.69	48%	**--**	33%	**--**	19%	**--**

Notes: Item handles in yellow fail to meet criteria suggested by ACT based on ratings by ALS panelists (≥ 33% "Very good" and <33% "do not use").
Items in blue are polytomously-scored items that were not eliminated and appear at more than one achievement level.