Developing Achievement Levels on the 2006 National Assessment of Educational Progress in Grade Twelve Economics

Process Report

Presented by ACT, Inc.
June 6, 2007

REDACTED BY GOVERNING BOARD

Developing Achievement Levels on the 2006 National Assessment of Educational Progress in Grade Twelve Economics

Process Report

The work for this report was conducted by ACT, Inc., under contract ED-06-CO-0098 with the National Assessment Governing Board.

Process Report
Table of Contents

APPENDICES

List of Tables

Page

List of Figures

Page

EXECUTIVE SUMMARY

OVERVIEW

This report describes the process and outcomes of a meeting that was held in March 2007 to set achievement levels for the 2006 National Assessment of Educational Progress (NAEP) in grade 12 economics. The meeting was conducted by ACT, Inc., under contract with the National Assessment Governing Board. The contract calls for ACT to conduct Achievement Level Setting (ALS) activities consistent with Board policies and to develop recommendations for setting achievement levels. The actual setting of achievement levels is a policy judgment by the Governing Board, 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 (ALDs), recommended cut scores, and identifies items that may be used to illustrate what students in the achievement levels know and can do (exemplar items).

Processes and conclusions from project activities that preceded the ALS meeting and from two Special Studies are also described in this report. Governing Board 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 conducted a Field Trial to develop a new method, Mapmark with whole booklet feedback, and a Pilot Study to identify which of two bookmark-based standard setting procedures, Mapmark with whole booklet feedback or Mapmark with domains, would be more appropriate for use with economics content.

Additional information about the project may be found in the following two other sources. The Technical Report documents technical advice ACT received in the project and data analysis procedures used throughout the process. The Special Studies Report documents the method and outcomes of two Special Studies conducted to determine if an independent panel would interpret the Achievement Level Descriptions in the same way as they were interpreted by panelists in the ALS.

BACKGROUND

The National Assessment Governing Board has been setting achievement levels for grades and subject areas in the 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 the Governing Board 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 NAEP, also known as the *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 policymakers assessing the nation's progress toward its educational goals.

The 2006 NAEP in grade 12 economics is the first of its kind. The Governing Board has never previously assessed economics performance at any level. The Economics Framework for the 2006 National Assessment was developed and approved by the Governing Board in 2002 for the first administration of the Economics NAEP in spring 2006 (National Assessment Governing Board, 2006).

For the current project, ACT proposed to develop a new standard setting method, Mapmark with whole booklet feedback, and to compare this method to the method ACT used to set achievement levels for the 2005 NAEP in grade 12 mathematics, called *Mapmark with domains*. Both methods are based on the bookmark procedure (Mitzel, Lewis, Patz, & Green, 2001). Bookmark was introduced in 1996 (Lewis, Mitzel, & Green, 1996), and has since become the most widely used standard setting method in state assessments (Council of Chief State School Officers, 2001). ACT's Mapmark methods improve the bookmark process with the use of item maps that illustrate the relative difficulty of all items within the assessment pool by locating each item on a score scale where a given response probability (in this case, 0.67) is met (Masters, Adams, & Loken, 1994). They differ in the type of feedback provided in the middle rounds. In Mapmark with domains, panelists are provided with percent correct scores on subareas of content, or domains, at the cut scores for each achievement level. In Mapmark with whole booklet feedback, panelists are provided with actual test booklets to show examples of student performance on the assessment at the cut score and the middle of each achievement level.

ACT conducted a Field Trial to evaluate the efficacy of the Mapmark with whole booklet feedback method in enabling panelists to set achievement levels, to train staff in implementation, and to gather information on areas for improvement in the method. Mapmark with whole booklet feedback was then compared to the Mapmark with domains method in a Pilot Study. ACT presented the results of these activities, and its recommendations, to the Governing Board's Committee on Standards, Design, and Methodology (COSDAM). COSDAM chose the Mapmark with whole booklet feedback method for the operational ALS meeting. The following points were noted by ACT's Technical Advisory Committee on Standard Setting (TACSS) and by COSDAM.

- Resulting cut scores and achievement level percentages from the two methods were highly similar to one another.

- Panelists' evaluations of the two methods were similar and positive; evaluations of the whole booklet method were slightly more positive than evaluations of the domains method.

- Mapmark with domains is more costly to implement than Mapmark with whole booklet feedback because it requires an additional initial investment to develop content domains.

- Mapmark with whole booklet feedback can serve as a model for states to adopt that is more cost effective than Mapmark with domains, is less burdensome with respect to preparation of materials for feedback, and is similar to the bookmark method most commonly used at the state level.

- Mapmark with whole booklet feedback may be more useful as a model because it can be used across all content areas and provides holistic feedback in a format very familiar to educators.

As recommended by the Governing Board, ACT implemented the Mapmark with whole booklet feedback method in the operational ALS meeting.

ALS Meeting

The ALS meeting lasted four days, March 7-10, 2007 (Wednesday through Saturday). It was conducted at the Westin Hotel in St. Louis. Sessions generally started at 8:00 a.m. and lasted until 5:00 p.m. or 6:00 p.m., except the last day, which adjourned at noon. The ALS agenda is in Appendix A.

The Panelists

Policies of the Governing Board regarding qualifications of panel members and composition of panels were followed. Panelists were selected from a sample of nominees provided by nominators contacted by ACT. ALS nominators were identified by drawing a single sample of public school districts from which nominations of teachers, nonteacher educators, and general public representatives were solicited. Nominators of private school teachers were identified from a sample of private schools drawn separately. An additional random sample of economics teachers was drawn directly from a national database of teachers, and nominations of nonteachers and members of the general public were also sought from all 50 state education associations, 14 economics professional organizations, and 250 colleges and universities. Approximately 200 of the 9,000 nominators contacted submitted a total of 292 nominations. From this pool of nominees, panelists were selected for three studies: the Pilot Study, Special Studies, and the ALS. Efforts were made to ensure that the panelists included teachers (55%), nonteacher educators (15%), and general public (30%), as well as proportional representation by gender, race, and geographic region. A total of 31 panelists were recruited for the ALS (Table 1).

Table 1: Demographics of Panelists Participating in the Achievement Level Setting

Type	Males	Females	Caucasian	African Am.	Am. Indian	Hispanic	**Total N (%)**
Teacher	7	11	16	1		1	18 (58)
Nonteacher	3	1	3	1			4 (13)
General Public	5	4	7	1	1		9 (29)
TOTAL N (%)	15 (48)	16 (52)	26 (84)	3 (10)	1 (3)	1 (3)	31 (100)

Design Factors

Panelists were divided into two rater groups (group A and group B) and six table groups. Groups and tables were design factors in the ALS meeting. 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. Group A and group B worked with different but equivalent and overlapping item pools. Each pool contained about 60% of the items in the 2006 assessment pool. Combined, they represented 100%.

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, et. al., 2001) with the addition of item maps in round 1, and provided whole booklet feedback and consequences data in subsequent rounds. The method is described in detail in the pages that follow.

Orientation

The ALS began with an overview of standard setting, a description of how panelists were selected for participation, and an orientation to the NAEP and the history of the National Assessment Governing Board (NAGB). Following this orientation, the participants took a form of the NAEP exam and scored their own performance. They were then instructed on the Economics Framework and specific elements of the Mapmark with whole booklet feedback method.

Round 1

Round 1 began with an exercise aimed at familiarizing panelists with the items in the assessment and, subsequently the gradient of difficulty in the assessment content. In this exercise, they reviewed the items in their pool in a Constructed Response Ordered Item Book (CROIB) and an Ordered Item Book (OIB), which presented the items in order of difficulty from easiest to hardest. More specifically, items were ordered by the scale value or estimate of student ability that corresponds to a 0.67 probability of correctly answering the item or getting full credit on the item based on the results from the 2006 administration of the assessment. Items are also presented on the item maps by the scale values that correspond to a 0.67 probability of a correct answer when scoring at a given score level (see Appendix B). The RP value of 0.67 is based on an IRT model.

This review was done with the whole group, the table group, and independently, beginning with a whole group review of the common constructed response items in the CROIB and ending with a table group review of all the items, constructed response and multiple choice, in the group's item pool. For each item, panelists were asked to identify the knowledge, skills, and abilities (KSAs) a student must demonstrate in order to correctly answer an item or to reach a given score point. Panelists then reviewed the remaining constructed response items within their table group. For polytomously scored (constructed response) items, panelists identified which additional KSAs are necessary to reach the next score point on the item, and for all items, panelists identified which KSAs are needed to get one item correct above and beyond the KSAs required to get easier items of similar content correct.

Following the KSA review, the Achievement Level Descriptions for economics were reviewed and discussed. Panelists were then instructed in how to place a bookmark representing their understanding of the KSAs a student should have mastered to be just qualified to be in an achievement level. Mastery was defined according to the 0.67 response criterion. Panelists were told that the students would be expected to have about a 67% chance of correctly answering the items at the scale value where the bookmark was placed, a higher likelihood of correctly answering items below that scale value, and a lesser likelihood of correctly answering items above that scale value.

Round 2

Panelists were shown the three cut scores representing the median of all panelists' round 1 bookmark placements (hereafter called *group cut scores*) and the general dispersion of all individual panelists' cut scores. Panelists were asked to locate in their Ordered Item Book (OIB), item maps, and booklet specific materials where their individual cut scores fell in relation to the group cut scores and to consider this differential in reviewing booklets relative to the group cut scores. Panelists then reviewed and discussed student performance in test booklets from three different forms of the assessment. Four booklets were provided at the group cut score and the middle of each achievement level. Panelists evaluated performance with respect to these points and in relation to their own cut scores for each level. Following this review, panelists were instructed to use their OIB and their item maps to help them in selecting a second set of cut scores consistent with their review of the booklets.

Round 3

In addition to the usual feedback after each round, panelists were shown the percentage of students who would fall into each achievement level based on the round 2 group cut scores. Panelists discussed this *consequences data*. They were asked to consider the consequences data as a reality check on their cuts and were given an opportunity to adjust their cut scores if appropriate.

Post-Round Activities

Following round 3, panelists were again given feedback from the previous round in the form of the group cut scores and consequences data based on those cut scores. On a consequences questionnaire, they indicated their reactions to the consequences data, including whether they wished to recommend alternative cut scores or would like to leave the cut scores unchanged. Finally, they provided recommendations concerning the selection of exemplar items for the achievement levels. Panelists were instructed to discuss each potential exemplar item with their table group, but they were to 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 (ALDs) in this task and to rate each item *very good*, *OK*, or *do not use* as an exemplar.

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 2 are representative of the fact that the Mapmark with whole booklet feedback process was well implemented. Average responses from the 1998 and the 2005 ALS processes, all on a 1–5 Likert scale, are shown for comparison. The Mapmark with whole booklet feedback process was viewed at least as positively as previous ALS processes.

Table 2: 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*)	Economics ALS	4.77
	2005 Math*	4.37
	1998 Civics*	4.04
I would describe the *effectiveness* of the Achievement Level Setting method as… (5 = *highly effective*)	Economics ALS	4.42
	2005 Math	4.28
	1998 Civics	3.59
This ALS process provided me an opportunity to use my *best judgment* to recommend cut scores (5 = *to a great extent*)	Economics ALS	4.81
	2005 Math	4.57
	1998 Civics	4.11
The *instructions* on what I was to do during each round were… (5 = *absolutely clear*)	Economics ALS	4.58
	2005 Math	4.17
	1998 Civics	4.18
My *understanding* of the tasks I was to accomplish during each round was… (5 = *totally agree*)	Economics ALS	4.71
	2005 Math	4.27
	1998 Civics	4.11

*Both were grade 12 assessments

On the key process evaluation questions in Table 2, 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 economics 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 economics ALS process performed well in relation to previous ALS processes.

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

- 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
- reasonableness of results when compared to external sources of information.

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 the Governing Board. In setting the achievement levels, the Governing Board 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 earlier in that they were developed before the ALS meeting and outside of the scope of this project. The ALDs were developed by economics experts working with the Governing Board and were provided to ACT for use in this standard setting project. They were used in the Field Trial and the Pilot Study as well as in the ALS meeting. The Achievement Level Descriptions are contained in Appendix C of the full 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. Based on this and on the fact that the ALDs serve as the basis from which the cut scores were established, ACT did not recommend any changes or modifications to the Achievement Level Descriptions.

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 4.71. In addition, no panelist provided a rating of less than 3.

- On a scale of 1 to 5, with 5 being totally adequate, the average response to the statement, "At the time I placed my round __ cut score recommendations, my understanding of the Achievement Level Descriptions" was 4.63, 4.71, and 4.81, respectively, for rounds 1 through 3.

Cut Scores

ACT recommended that the Governing Board adopt the group cut scores from round 3 of the ALS meeting (326 for Basic, 363 for Proficient, and 411 for Advanced on a converted NAEP scale called the *ACT NAEP-like scale*). This recommendation was based partly on the conclusion of ACT's TACSS that the ALS process had procedural validity and produced reliable results across panelist type and group. It is also based on the conclusion that the achievement level percentages associated with the cut scores are likely to be considered reasonable (Figure 1). Also, round 3 cut scores are based on all of the information that ACT recommends be considered by panelists in adopting cut scores, including student performance data.

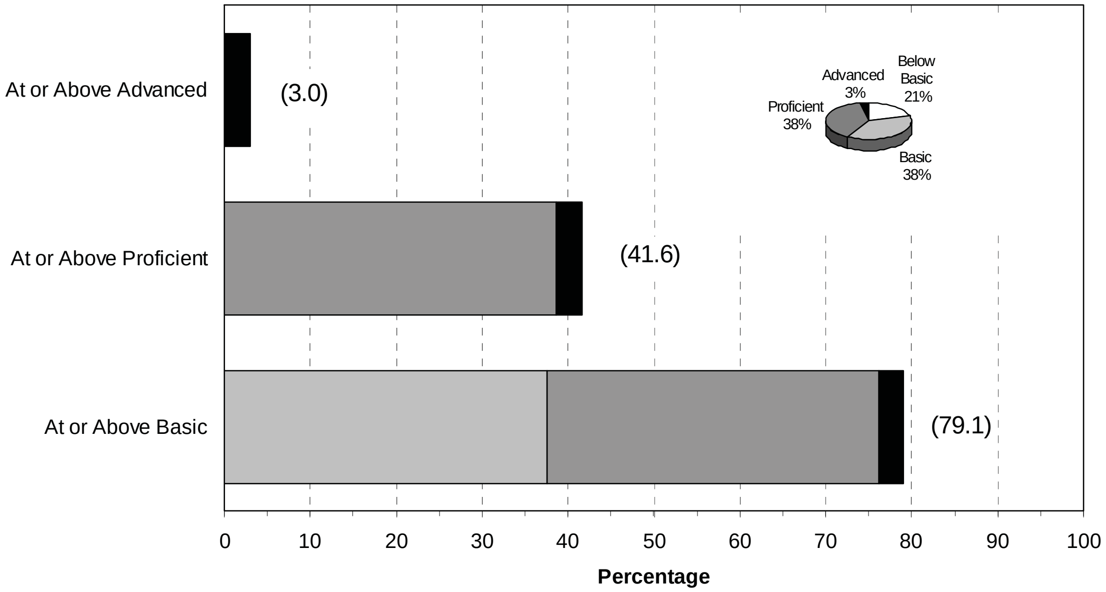


Figure 1. Percent of students at or above each achievement level based on the final ALS cut scores.

There were small but noticeable differences between the results of the ALS and the Pilot Study. In view of these differences, ACT and the Governing Board's Committee on Standards, Design, and Methodology looked for evidence that results from one meeting are more reasonable than results from the other. In addition to random variation, differences in standard setting results may be due to many factors, including: differences in the standard setting methodology and procedure, physical accommodations, facilitators, panelists, observers, interactions among panelists over rounds, and panelist understanding of the differences in the purpose of the meeting. The analysis of procedural and internal consistency data from the ALS suggest that the panelists were well qualified and the method was conducted well, was understood by panelists, and was not unduly impacted by variation in the panelists. In addition, classification studies with an independent panel of economics teachers and nonteacher educators produced classifications of student performance on assessments and items into the achievement levels consistent with the classifications based on the cut scores for the ALS. The conclusion was that the results of the ALS are reasonable.

Exemplar Items

Following round 3 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 3 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 place their bookmarks and determine cut scores. All potential exemplars associated with scale values within an achievement level using this criterion were selected for panelist review. This resulted in a total of 57 items/score points selected as potential exemplars; 10 were associated with the Basic level, 34 with Proficient, and 13 with Advanced. Panelists individually rated the items as *OK*, *very good*, and *do not use* based on the match between item content and the ALDs.

ACT suggested that the Governing Board eliminate from consideration those exemplars which were rated by 20% or more of the panelists as *do not use*. In addition, ACT recommended that classification of items by Special Studies' panelists and ratings of items by content experts, be taken into consideration when selecting exemplar items.

RECOMMENDATIONS

ACT's principal recommendations concern the three outcomes of the Achievement Level Setting Process—Achievement Level Descriptions, cut scores, and exemplar items.

- ACT endorses the Achievement Level Descriptions.
- ACT recommends the cut scores from round 3 of the ALS meeting.
- ACT recommends that the Governing Board use the lists of items and panelists' ratings from the ALS meeting, coupled with other information, in the process of selecting exemplar items.

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

Developing Achievement Levels for the 2006 NAEP in Grade Twelve Economics: 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 to identify "appropriate achievement goals for each grade or age in each subject area to be tested" under the National Assessment. The Governing Board policy specifies three achievement levels—Basic, Proficient, and Advanced—and states that the purpose of these levels 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. ALDs are 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 knowledge and skills required for performance at each level.

As currently specified by the Governing Board policy, there are two stages to the NAEP Achievement Level Setting (ALS) process. In Stage 1, grade- and subject-specific ALDs are developed from general policy definitions. In Stage 2, the ALDs are translated into cut scores and exemplar items to represent the achievement levels are identified. Stage 2 has traditionally been performed in an ALS meeting by a panel of teachers, nonteacher educators, and representatives of the general public. The targeted percentages of these types of panelists are, respectively, 55%, 15%, and 30%. This is in keeping with the Governing Board policy that the development of achievement levels shall be a widely inclusive activity. The Governing Board may 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 the Governing Board. Key criteria in the Governing Board's policy judgment are the validity and reliability of the ALS process and the apparent reasonableness of the results. The Governing Board specifies that the final reports for ALS activities are to serve as the principal means of documenting these criteria for specialists in the field as well as for the general public.

Background on the Current Project

The development of the Economics Framework, assessment, and ALDs, and consequent administration of this assessment to a national sample of 12th grade students in spring 2006 led the Governing Board to issue a procurement to establish grade 12 economics cut scores and to identify exemplar items. The grade 12 economics NAEP is the first of its kind and was administered to a large, nationally representative sample of students. Item statistics and student distribution data for all ALS activities in this project are based on the results from this administration.

This report provides a detailed description of the method and outcomes of a meeting that was held in March 2007 to set achievement levels for the 2006 NAEP in grade 12 economics. It also describes project activities preceding the ALS meeting that were designed to assess the reliability and validity of two different standard setting methods and thereby inform the decision as to which method to implement in the ALS meeting itself. These activities included the development of content domains for the Mapmark with domains standard setting method, a Field Trial of the new standard setting method Mapmark with whole booklet feedback, and a Pilot Study to assess the validity and reliability of both methods. Also summarized are two Special Studies designed to determine if an independent panel interpreted the ALDs in a manner consistent with the ALS panelists’ interpretation.

Both of the proposed methods under consideration for implementation in the ALS were based on the bookmark procedure. The bookmark method was introduced as recently as 1996 (Lewis, et al., 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, et al., 1994) and holistic feedback, such as domain-score feedback (Schulz, et al., 2005) or whole booklet feedback (Loomis & Hanick, 2000). ACT had conducted extensive research on these issues through previous standard setting contracts with the Governing Board (Reckase, 2000), through other NAEP-related projects (Schulz, et al., 2005), and in support of its own assessment programs (Schulz, Kolen, & Nicewander, 1999). The methods in this contract both use the bookmark procedure (Mitzel, et al., 2001) in round 1, and provide holistic 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 representing both item difficulty and student achievement on the score scale (Appendix B).

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, economics education, and experience with the NAEP. (See Appendix D for a list of the TACSS members.) The TACSS met four times over the course of the project and provided input on key components of the project including the design of the methods; the design of the domain development process, Field Trial, and Pilot Study;

the method to use in the ALS meeting; data analysis procedures, and the formulation of conclusions and recommendations presented to the Governing Board.

CONTRACT ACTIVITIES PRIOR TO THE ALS MEETING

Contract activities prior to the ALS meeting fall into three general categories: (a) domain development, (b) the Field Trial, and (c) the Pilot Study. These activities are described in detail in the following sections.

Domain Development

A key feature of the Mapmark with domains method is the development and use of relatively specific subareas of content, or *domains*, that help teachers, nonteacher educators, and members of the general public understand what it is that students at a given level of achievement can or cannot do, and what growth in achievement means in relation to mastery of these content areas (Schulz, et al., 2005). The domains should have the following features:

1. *Clear Definitions*. 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, nonteacher educators and the general public alike, the knowledge, skills, and abilities required by items in the domain.

2. *Coherence*. 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. *More than a Single Item*. The domains should each be comprised of multiple items.

4. *Varied in Difficulty*. A fourth criterion, that the domains differ in difficulty, added to their utility in the Mapmark with domains method in the 2005 Mathematics Achievement Level Setting. This characteristic is not considered essential to the successful use of domains in standard setting, but, in as much as such separation may add to their value, was also a goal in the domain development process.

The outcomes of the domain development process were to be:

1. approximately 20 teacher domains; each with clear definitions and one to two exemplar items;
2. score domains (3-5 per content area) into which the teacher domains were grouped based on similarity in difficulty; and
3. data on the cohesiveness and usefulness of domains for a Mapmark process.

The 2006 Economics Framework was used as the basis from which the domains were built. The framework contains a total of 20 *standards*. Each standard represents a well-defined area of content and is represented by a few lines of relatively jargon-free text and between 1 and 9 more specific areas of content called *benchmarks*. Most of the standards are defined within one of three broad content areas (Market, National, and International Economies), although a few are described in more than one content area for a total of 26 (see NAEP Economics Framework for more detail). Previous experience in mathematics standard setting suggested that standard setting panelists can understand and effectively use up to 23 teacher domains, where each domain is defined exclusively within a content area (ACT, Inc., 2005).

Since the standards were already well-defined and sufficient in number (enough, but not too many for panelists to use in standard setting), the principal question remaining about how they might be used in, or modified for, a standard setting process was whether the content that they represented could be distinguished in terms of instructional timing. When domains are ordered in instructional timing, profiles of performance on the domains can be meaningful in a standard setting process. Achievement levels then can be defined and distinguished from one another with reference to percentage correct scores on the domains.

Domain development was a four step process, summarized in Table 3. Each step was designed to evaluate the domains and items associated with them and to further refine the domains to reflect a natural hierarchy of knowledge, skills, and abilities. This involved the review and reclassification of some items. All activities used all items in the NAEP grade 12 economics pool, both secure items and items scheduled for release.

Table 3: Development of Grade 12 Economics Content Domains

Step	Event/Process	Date/Deadline (2006)	Purpose/Product
1	Domain Development Meeting	September 17-19	First Draft of Domains
2	Refine	By October 6	Second Draft of Domains
3	Domain Item Classification Study	October 13	Evaluate Domain Coherence
4	Refine	By November 16	Finalize Domains and Evaluate

Participants in Domain Development

Five economics content experts were recruited to assist ACT with domain development. One of the five, Stephen Buckles, is the former president of the National Council on Economic Education and was central to the development of the economics NAEP. He was recruited to serve as a content facilitator in the domain development process. Dr. Buckles' role was to answer any content-related questions that arose and to provide content-related examples to clarify task instructions. He facilitated, but did not participate in, the tasks.

The remaining four content experts were all members of the economics NAEP framework committee and had participated in the development of the economics NAEP. They served as panelists in the domain development process.

Before The Domain Development Meeting

Since ACT's plan was to use standards (the National Voluntary Standards incorporated into the framework of the economics assessment) to define the domains as much as possible, it was necessary for both domain development and evaluation purposes to establish a *link* between the test items and the standards with regard to instructional timing and difficulty. The goal was to determine if differences in difficulty amongst potential teacher domains actually reflect meaningful differences in the timing of mastery of content.

Difficulty of the standards was calculated by aggregating item difficulty as represented by the b-value to the standard level. Instructional timing information was acquired through the content expert assessment of the relative timing of mastery of the knowledge, skills, and abilities associated with each standard. The first task was, therefore, to gather data on instructional timing. This task was completed prior to the domain development meeting so as to allow time for eight additional tasks in the meeting itself.

Task 1: Independent Ratings of Benchmarks on Instructional Timing

There are 20 standards in the Economics Framework. Most standards appear within only one content area (Market Economy, National Economy, or International Economy), but a few (e.g., standard 17) appear in more than one. When a standard is counted separately for each content area in which it appears, there are 26 standards in the economics assessment. Each standard is represented by one or more benchmarks. Benchmarks are the most specific unit of content in the Economics Framework. There are a total of 105 benchmarks in the framework with the number per standard ranging from 1 (i.e., standard 17 is represented by one benchmark in each content area) to 9 (standard 20 is found only in the National Economy content area, but is represented by 9 benchmarks in that area).

Instructional timing was defined operationally by a rating scale in which the relative ordering of mastery of concepts was presented in relation to an instructional sequence (see Figure 2). Prior to the domain development meeting, the four content expert panelists were asked to rate the benchmarks in the framework using this rating scale. The goal of this task was to determine where mastery of the majority of knowledge, skills, and abilities (KSAs) in each benchmark occurs in relation to mastery of the KSAs in other benchmarks, or to obtain a rating of instructional timing for each benchmark and, consequently, each standard. The specific instructions used in the benchmark rating task are contained in Appendix E.

Benchmark Rating Scale	
Rating	Meaning
5	The knowledge, skills, and abilities associated with this benchmark are mastered after the vast majority of other benchmarks have been mastered and late in an instructional sequence in economics the goal of which is mastery of all the benchmarks in the NAEP Framework.
4	Mastery of the knowledge, skills, and abilities associated with this benchmark typically follows mastery of the majority of benchmarks that occur earlier in a sequence.
3	The knowledge, skills, and abilities associated with this benchmark are mastered about midway through an instructional sequence in economics.
2	Mastery of the knowledge, skills, and abilities associated with this benchmark typically follows mastery of some earlier benchmarks.
1	The knowledge, skills, and abilities associated with this benchmark are mastered very early in an instructional sequence in economics.
DOES NOT APPLY	The knowledge, skills, and abilities (KSAs) associated with this benchmark are mastered in an economics curriculum in no particular order in relation to other KSAs.

Figure 2. Rating scale used in the benchmark rating task in domain development.

The instructional timing of the standard was then assessed by aggregating the benchmark ratings to the standard level. So as to allow for a comparison between the instructional timing ratings and the difficulty of the standard, benchmark and standard difficulty were calculated by ACT after Task 1 for presentation and discussion in Task 2. Items are classified by benchmark in the framework, so it is possible to assess the difficulty of the benchmark and standard by aggregating the item difficulty to the level of benchmark and standard (within content areas).

The Domain Development Meeting

In the domain development meeting, additional instructional timing data were collected, this time at the item level (see Task 2 described below), and the same four content experts were provided with data summaries. Referring to the data summaries and to item text, the content experts provided input on domain coherence and the meaningfulness of differences in domain difficulty as they relate to instructional timing. They then used this information to make recommendations for combining standards into a smaller number of teacher domains and for reclassifying items to improve domain coherence and meaning. This process was facilitated by ACT staff member Matt Schulz, who was assisted by content facilitator Stephen Buckles.

Task 2: Independent Rating of Items by Instructional Timing

This task was designed to assess whether content experts' expectations of the instructional timing of standards, based on benchmarks in the framework, were actually reflected by the instructional timing of items in the assessment. This would provide information as to how well the items represent the benchmarks and standards as described in the framework, and would allow experts to analyze potential sources of discrepancy between standard timing ratings and difficulty. The rating scale and instructions used in the item rating task are similar to those used in the benchmark rating task and are contained in Appendix F. Briefly, the content experts independently rated each item in the assessment (N = 186) on a rating scale of instructional timing (location in a sequence of mastery) that was similar to the rating scale used to rate the benchmarks. Polytomously-scored items were rated according to the knowledge, skills, and abilities needed to obtain full credit on the item. Results from this task were combined with results from the benchmark rating task, and other statistical information, for use in the remaining tasks as described below.

Results of Tasks 1 and 2 were analyzed to produce estimates of difficulty and instructional timing at the following levels of analysis:

Difficulty:

- Items (item scale value)
- Benchmarks (mean item scale value per benchmark)
- Standards (mean item scale value per standard)

Instructional Timing:

- Items (instructional timing ratings)
- Benchmarks
  - Mean rating of item instructional timing
  - Benchmark ratings
- Standards
  - Mean rating of item instructional timing
  - Mean rating of benchmark instructional timing

The scale value of an item was set to be the scale value at which a student has a 0.67 probabilty of correctly answering the item or of reaching the score level on a polytomously scored item. This is a reasonable measure of the difficulty of the item in a standard setting method, such as bookmark or Mapmark, that uses an item mapping technology and a response probability criterion of 0.67 for constructing the item map or Ordered Item Book.

The correlations between the standard difficulty and the item and benchmark instructional timing variables were positive and strong at 0.63 and 0.66, respectively (Table 4). A positive correlation indicates that the items in the assessment represent the standards in terms of instructional timing. There was also a moderately positive (but slightly smaller due to smaller numbers of items within each benchmark) correlation between the two instructional timing variables associated with the benchmarks.

The correlation between the mean instructional timing of individual items and the items' scale values, or difficulty, was also moderately positive. More difficult items should generally be associated with content and skills that have higher ratings of instructional timing.

Table 4: Correlations between Difficulty and Instructional Timing of Items, Benchmarks and Standards

		Difficulty	
	Items	Benchmark	Standard
Instructional Timing	r (N)	r (N)	r (N)
Items	.44[1] (186)	.51 (95)	.63 (26)
Benchmark	N/A	.48 (95)	.66 (26)

[1] All correlations are significant at the p<.001 level

The content experts were shown the mean benchmark rating, item rating, and the mean scale value (standard difficulty) for each standard (Table 5). These results and the correlations were discussed with the content experts. Areas of discrepancy between instructional timing ratings and difficulty were identified and discussed. Experts were asked to identify possible explanations for such discrepancies and to keep these differences in mind as they consider the coherence of the domains in Task 3.

Task 3: Ratings of Coherence and Difficulty Order

The purpose of this task was to obtain content experts' input as to the coherence of the domains as represented by the standards and the reasonableness of the difficulty ordering of the standards given the instructional timing ratings. At this point, content experts were asked to consider the standards as "teacher domains." Items were presented to content experts, one page per item, organized by order of difficulty (scale value) within domain within content area. Domains within content area were ordered by expected percent correct scores conditional on a level of achievement (scale value) associated with an overall performance of 67% correct on the content area. For example, a scale score of 288 was needed to get 67% of the items in the International Economy content area correct. Expected percent correct scores on domains within the International Economy were computed conditionally on a scale score of 288.

Table 5: Mean Instructional Timing Ratings and Scale Value (Difficulty) for Each Standard, as Presented to the Content Experts for Discussion in Task 2

Content Area	Standard	Mean Benchmark Rating	Mean Item Rating	Mean Scale Value
Market	14	1.5	1.9	229.7
	7	2.4	2.4	242.2
	4	1.6	2.3	243.2
	15	3.1	2.8	247.0
	1	1.4	1.8	254.3
	9	2.6	3.4	255.2
	13	3.3	2.8	255.8
	10	2.3	2.5	273.1
	2	3.0	2.9	277.6
	16	4.0	3.3	278.7
	8	2.8	3.0	285.6
	17	4.3	3.5	299.0
National	12	3.0	3.3	265.6
	15	3.2	3.4	266.8
	16	3.1	2.7	272.3
	3	1.8	1.9	279.1
	17	4.0	3.8	281.7
	11	2.5	3.0	287.7
	18	3.6	3.3	291.6
	19	2.9	3.3	302.1
	20	4.3	3.9	317.0
International	5	3.5	3.6	259.2
	15	3.6	3.5	279.5
	6	3.1	3.4	279.6
	7	3.5	4.2	293.0
	17	4.5	4.5	321.1

The content experts were also given a domain item map to accompany this task (Figure 3). There was one domain item map per content area. Columns on the domain item map corresponded to domains within the content area, with domains ordered from left to right by increasing expected percent correct score conditioned on a level of student achievement associated with getting 67% of all items within the content area correct. The conditional expected percent correct scores were shown at the bottom of the item maps. The item map showed the relative difficulty of each item within a domain and, overall, how much variation there was in item difficulty within and across domains within a content area. This was an important consideration in judging the coherence of a domain because there is generally expected to be less variation in item difficulty within than across domains due to domains representing more specific areas of knowledge, skills, and abilities than the content area as a whole.

International Economy Domains

Domain	22		23	24		25		26
(Standard)	(5)		(15)	(6)		(7)		(17)
Scale								
above								P25_2
396								
393								
390								
387								P22_4
384								
381								
378								
375								
372								
369								
366				M154				
363								
360								
357								
354								
351						P17_2		
348								
345								
342								
339								P25_1
336								
333								
330								M149
327								
324								
321								
318								
315			M143					
312								
309								
306								
303				M133				
300								
297								
294								
291						M112	M117	P22_3
288				M106		M109		
285	M97			M99				
282			M93			P17_1		
279	M90			M89				
276			M84	M85				
273								
270								
267								
264								
261				M55				P22_2
258				M51	M53			
255						M48		
252								
249								
246	M37							
243	M31	M33	M32					
240								
237								P22_1
234								
231								
228								
225				M14				
222								
219								
216								
213								
210								
207								
204								
201								
198								
195								
192								
189								
186								
183								
180								
below								
Expected Percent Correct	81%		72%	70%		64%		52%

Figure 3. Domain item map for the International content area, provided to content experts in Task 3.

The four content experts were asked to take the domains one at a time, in order of their difficulty within content areas, and briefly review all of the items classified into the domain. For each domain, they were to indicate their level of agreement on a scale from 1 (strongly disagree) to 5 (strongly agree) with the statement “The items in this domain represent a reasonably coherent area of content.”

Content experts were then asked to rate if they felt the difficulty of the domains seemed reasonable in comparison to that of other domains within the same content area. Experts were told to consider the domains within a content area in terms of strata of difficulty, similar to how score domains were to be created in a later step (see Task 7 below). They were asked to rate the difficulty order of domains only in relation to domains in a different stratum of difficulty and not to compare two domains at approximately the same level of difficulty.

The content experts were to indicate their level of agreement on a scale from 1 (strongly disagree) to 5 (strongly agree) with the statement “The difficulty ordering of this domain seems reasonable when compared to other domains in the content area.” One important piece of information panelists used to answer this question was the instructional timing of the domains. In those cases where domains were more or less difficult than their instructional timing would strictly predict, content experts were asked to consider other facets of the domain, such as the cognitive category of the items, their context, or even aspects of the domain outside the framework, such as potential differences in the instructional emphasis placed on the domains nationally, in order to judge whether the difficulty order of the domains was reasonable. The content experts’ domain coherence and difficulty reasonableness ratings are provided in Table 6. Any ratings at a mean of 3 or lower were highlighted, and these results were reviewed with the content experts in Task 4.

Table 6: Content Experts' Domain Mean Coherence and Difficulty Reasonableness Ratings

Content Area	Standard	Mean Coherence	Mean Difficulty
Market	14	4.00	4.50
	7	4.50	4.00
	4	4.75	3.75
	15	3.25	3.25
	1	3.50	3.00
	9	4.75	3.50
	13	3.50	4.00
	10	4.75	2.75
	2	4.75	4.00
	16	3.00	4.00
	8	5.00	4.25
	17	3.00	4.00
National	12	4.25	3.25
	15	3.00	3.50
	16	3.50	2.50
	3	3.75	2.75
	17	4.00	3.25
	11	4.75	3.50
	18	4.50	4.25
	19	4.00	4.25
	20	4.25	3.75
International	5	4.00	4.00
	15	4.75	3.75
	6	3.75	3.25
	7	5.00	4.00
	17	5.00	4.25

Task 4: Combining Standards to Reduce the Number of Teacher Domains

The purpose of this task was to obtain the recommendations of content experts for combining standards to reduce the number of teacher domains from the original 26 standards. Some of the standards have very few items (as few as just one for standard 17 in Market Economy), while others may have received low ratings for coherence and reasonableness of difficulty relative to other domains. Input was also sought specifically with regard to domains that corresponded to the same standard, but were in different content areas. Four standards (7, 15, 16, and 17) were represented in more than one content area. For example, standard 17 was represented in three different content areas. Content experts were asked specifically what they recommended be done with these "multiple domain" standards. One option, for example, was to create a single domain using items across content areas, and to put the combined domain into one of the content areas. The four content experts made recommendations for combining standards independently. Then, summary information on content experts' ratings of domain coherence and difficulty order was presented (Table 6) along with their reactions to specific questions about what to do with multiple-domain standards. This information was discussed and, as a result of the discussion, the content experts decided to remove standard 15 from the International content area and combine it with standard 15 in National and to delete standard 17 from the Market content area and combine it with standard 8 in Market. All other standards were left as is. This resulted in 24 domains as shown in Table 7.

Table 7: Domains and Their Corresponding Standards after Task 4

Content Area	Domain	Standard	Notes
Market	1	14	
	2	7	
	3	4	
	4	15	
	5	1	
	6	9	
	7	13	
	8	10	
	9	2	
	10	16	
	11	8	Standards 17 and 8 in Market combined
National	12	12	
	13	15	Standards 15 in National and International combined
	14	16	
	15	3	
	16	17	
	17	11	
	18	18	
	19	19	
	20	20	
International	21	5	
	22	6	
	23	7	
	24	17	

Task 5: Item Reclassifications

The purpose of this task was to obtain content expert input for reclassifying items in order to make the remaining domains more coherent and reasonable in order of difficulty. The domains remaining after Task 5 were used to organize and present the items to content experts. Items were organized, one page per item, in order of difficulty within domain within content area. In an *Item Reclassification Form*, items were listed in the order of their appearance in the experts' materials (Figure 4). The form also showed other information about the items, including their scale value, mean instructional timing rating, and their framework classifications as to standard, context, benchmark, and cognitive category. The experts considered all of this information in recommending whether to reclassify an item into a different domain, or to declassify it. Declassification means that the item, at least for the time being, was not put into any domain. Content experts completed the item reclassification form independently, providing an item with a rating of 0 (do not change classification), 1-26 (reclassify into domain indicated from 1-26), or D (declassify).

Ref Num	Content Area	Standard	Benchmark	Domain	Cognitive Category	Context	Handle	Mean Instructional Timing	Scale Value	Recommendation 0=OK, 1-26=Reclassify D = Declassify
1	Market	14	1.14.3	1	Reasoni	Household	M7	2.3	217	
2	Market	14	1.14.2	1	Knowing	Household	M8	2.3	220	
3	Market	14	1.14.1	1	Knowing	Household	M44	1.3	252	
4	Market	7	1.7.3	2	Reasoni	Household	M2	2.8	191	
5	Market	7	1.7.2	2	Knowing	CxtFree	M11	1.5	222	
6	Market	7	1.7.5	2	Reasoni	Individual	M12	2.3	224	
7	Market	7	1.7.1	2	Knowing	CxtFree	M13	1.8	226	
8	Market	7	1.7.3	2	Knowing	CxtFree	M38	2.5	247	
9	Market	7	1.7.3	2	Reasoni	CxtFree	M40	2.8	248	
10	Market	7	1.7.3	2	Knowing	Household	M61	2.5	262	
11	Market	7	1.7.3	2	Reasoni	Household	M67	3.0	265	
12	Market	7	1.7.3	2	Knowing	CxtFree	M123	2.3	295	
13	Market	4	1.4.2	3	Reasoni	Individual	M5	2.3	210	
14	Market	4	1.4.1	3	Knowing	Individual	M6	1.8	214	
15	Market	4	1.4.2	3	Applyin	Individual	M21	2.3	237	
16	Market	4	1.4.1	3	Reasoni	Household	M24	2.0	238	
17	Market	4	1.4.1	3	Reasoni	Public	M35	2.0	244	
18	Market	4	1.4.1	3	Knowing	Individual	M41	2.8	249	
19	Market	4	1.4.2	3	Reasoni	Individual	M56	2.3	261	
20	Market	4	1.4.1	3	Reasoni	Individual	M66	2.5	264	
21	Market	4	1.4.2	3	Reasoni	Household	M76	2.8	272	

Figure 4. Portion of the item reclassification form used by content experts to determine classification of the items.

Content experts' input from the independent reclassification task was summarized and discussed. The content experts' independent recommendations were shown for each item. If three or more of the experts felt an item should be reclassified into a particular domain, or declassified, it was. When only one expert recommended reclassifying an item, it was not reclassified. At the conclusion of the discussion, only four items had been reclassified, in addition to the five reclassified when the content experts combined standard 17 with standard 8 in Market and standard 15 in International with standard 15 in National.

Task 6: Creating Score Domains

The purpose of this task was to obtain content expert input for further reducing the number of teacher domains. The domains remaining after Task 4, and after reclassifying items in Task 5, were considered to be potential *teacher domains* for purposes of standard setting. These are the content specific groupings of items.

Score domains are groupings of teacher domains of similar difficulty within a content area into a single domain. In the 2005 mathematics NAEP standard setting, the score domains were used to group expected correct scores of similarly difficult teacher domains into a single score. Each teacher domain within a score domain was defined independently. Teacher domain content was not necessarily related to the content of other teacher domains in the same score domain, and the score domains were defined only in terms of the teacher domains that comprised them.

In order to group teacher domains together, expected percent correct scores on the teacher domains within the content area were computed conditionally on a level of student achievement associated with getting 67% correct on all items within the content area. The expected percent correct scores were presented on the item map as they were in Task 3 (Figure 3) and expected percent correct curves were presented in domain score plots (Figure 5). The domain score plots illustrate what percent of a domain students at different levels of achievement on the score scale would be expected to get correct. When the curves are close to one another, this indicates two teacher domains are at a similar level of difficulty. The facilitator suggested that the content experts consider combining some of the teacher domains at the same level of difficulty into a single score domain.

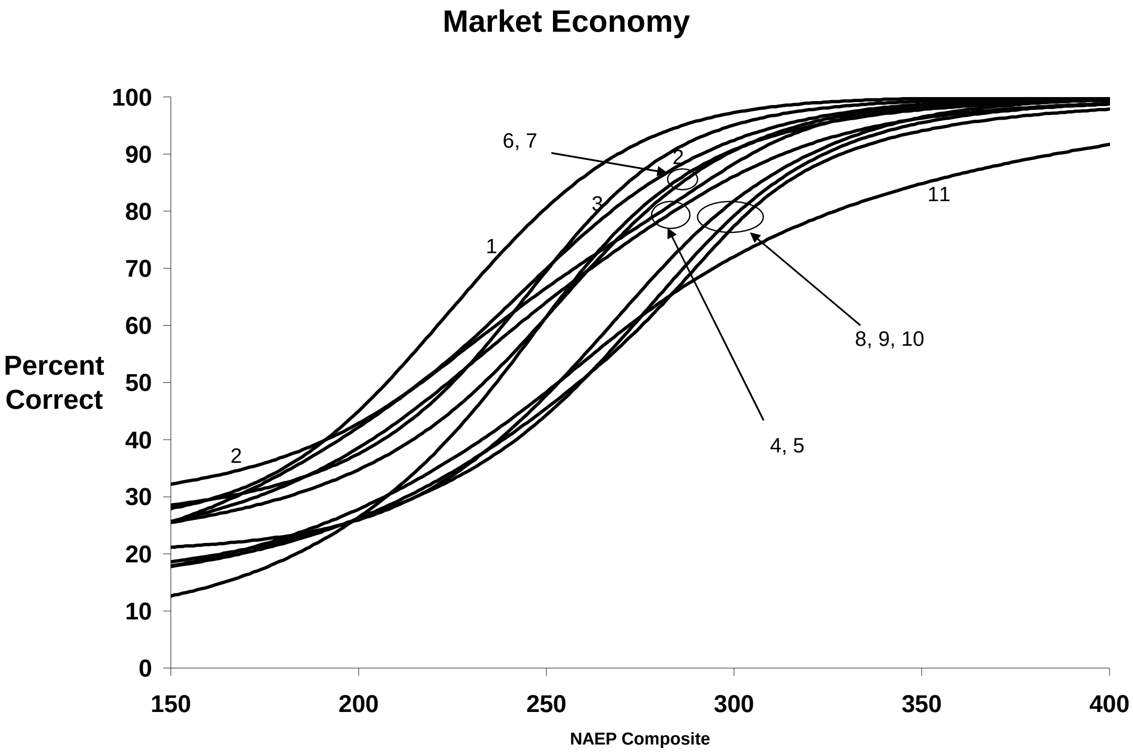


Figure 5. Domain Score Plot for the Market Economy content area. Teacher domains 8, 9 and 10 on this plot are at about the same level of difficulty and might be considered for combination into a score domain.

Once score domains were defined, the experts were asked to independently recommend which teacher domains within a score domain could be combined. Their independent recommendations were summarized and discussed as a group. An important instruction in this task was that teacher domains within the same score domain could potentially be combined into a single teacher domain, but that teacher domains from different score domains could not be combined. The aim was to identify all teacher domains to be defined in Task 7. Upon completion of this task, the teacher domains were grouped into score domains and were titled and numbered within each content area (Table 8).

Table 8: Score and Teacher Domains Resulting from Task 6

Score Domain	Teacher Domain	Standard	Title
1	M1	14	Entrepreneurs
	M2	7	Markets and Equilibrium
	M3	4	Incentives
2	M4	15/13	Income, Firm Investment, Human Capital, and Productivity
	M5	1	Scarcity, Productive Resources, Costs, and Unintended Consequences
	M6	9	Competition
3	M7	10	Economic Institutions
	M8	2	Effective Decision Making
	M9	16	Economic Role of Government
	M10	8/17	Supply, Demand, and Prices
4	N1	12	Interest Rate Determination
	N2	15/Intl 15	Human and Physical Capital
5	N3	16/17	Economic Role of Government
	N4	3	Resource Allocation Methods and Economic Systems
	N5	18	Gross Domestic Product, Prices, and Employment
6	N6	11/19	Money, Unemployment, and Inflation
7	N7	20	Fiscal and Monetary Policy
8	I1	5	No Title
9	I2	6/17	Benefits and Costs of Trade
	I3	7	Exchange Rates

Task 7: Creating Domain Definitions

The purpose of this task was to obtain content experts' input for representing the content of the domains by means of a title, brief narrative, and one or two sample items. For this task, the four content experts worked in pairs. One pair worked on the domains in the Market content area. The other pair worked on domains within the National content area. The pair who finished their content area first began working on the International content area. The content experts were told to write the definitions based on the actual content, or items, within each domain. For domains that consisted of a combination of standards as identified in Task 6, experts developed a single title and narrative. At the conclusion of this task, preliminary titles and definitions had been written for all the domains except for domain I1, the first domain in the International content area. No sample items had been selected for any domains.

Task 8: Evaluating Domain Information for Standard Setting

The purpose of this task was to obtain content experts' input regarding the value added of domains over the three content areas in standard setting. For this task, the final teacher domains, as defined in Task 7, were used. These domains were represented on an item map and at the level of score domains on a Domain Score Chart similar to those that were used in the 2005 grade 12 mathematics ALS project. A hypothetical Proficient cut score was represented on these materials.

To put these materials in context, rounds 1 and 2 of a Mapmark with domains standard setting procedure were described. Content experts were then asked to provide input and recommendations regarding the use of domain information. They were asked whether the domain item map presents useful information over and above the content area item map for understanding what students at the cut score can or cannot do. They were also asked if expected percent correct score information based on the domain item map would be useful in representing areas of strength or weaknesses of students at the cut score. Finally, they were asked for any suggestions to improve the utility of domains. In this discussion, the content experts indicated that the teacher domains were useful but that the score domains were not. It was recommended that ACT eliminate the score domains. Based on this recommendation and later discussion with the TACSS, ACT decided to use only the teacher domains (hereafter called domains) in the Mapmark with domains procedure for economics.

Revising and Refining Domains after the Domain Development Meeting

Task 9: Revising Domains for Use in the Domain Item Classification

The domains produced in the domain development meeting are important in the overall domain development process because they represent the input of content experts. Experience has shown, however, that the domains can be improved through a subsequent process of close collaboration between at least one measurement/domain development expert and at least one content expert. The necessary expertise for this subsequent process was represented by a *domain development team* comprised of three people: one content expert from the domain development meeting, the content facilitator from the domain development meeting, and ACT's NAEP Assistant Project Director. This team reviewed the domains from the domain development meeting, further refined definitions to reflect the item content, and selected 1 to 2 example items per domain to represent the domain. In addition, the team identified any domains that still seemed to be lacking in coherence and/or specificity as to content and reclassified or declassified items and redefined domains accordingly. The goal was to remain as faithful as possible to the domains created in the domain development meeting and to make modifications only as required to enhance the coherence of the domains themselves and the clarity of their definitions.

At the conclusion of the refine and revision process, there were 10 domains in the Market content area, 9 in National, and 3 in International. Each domain had a clear title and definition (Appendix G) and between 1 and 2 exemplar items. In addition, of the 186 items classified during the domain development meeting, the domain development team reclassified 42 (23%). Eight of the 42 reclassified items were selected as exemplar items to illustrate the domain and 3 of the 42 were declassified because they did not seem to clearly fit into any domains. These final domains were used in Task 10, the Domain Item Classification Study, for validation.

Domain Item Classification Study

Task 10: Domain Item Classification Study

A Domain Item Classification Study was conducted to evaluate the coherence and usefulness of the domains resulting from Task 9 and the clarity of their associated definitions. If teachers tend to agree on which domains items belong to, and they feel the domains are useful and that the domain definitions are clear, the domains are likely to be useful for the purposes intended in this project.

Participants

Five teachers of high school economics and two members of the general public with workforce experience in economics participated in the meeting. The panelists were recruited from school districts and businesses surrounding ACT corporate headquarters in Iowa City. All teachers were from different schools.

Process

The meeting was held in Iowa City, IA on October 13, 2006 and was facilitated by ACT's NAEP Assistant Project Director Matt Schulz. Observing were ACT staff Christina Peterson, Nancy Petersen, and Jim Sconing, NAGB Assistant Director of Psychometrics Susan Loomis, and ACT TACSS member Bob Forsyth. An agenda for the meeting is provided in Appendix H. At the beginning of the meeting, panelists were provided with some basic information about NAEP, and then were oriented to the item classification task through a practice exercise. In this exercise, panelists were told to familiarize themselves with the domains in the Market content area by reading the definitions and looking at the sample items. Then, they independently classified eight items in the Market content area into the domains in that same content area. After independently classifying these items, they shared their classifications with the panel and discussed their reasons behind those classifications. Participants were told that there was no "correct" classification, but that the purpose of the exercise was to help them get a sense of how their thinking compared with that of the other panelists.

Following the practice exercise, the teachers classified all 132 remaining items not used as exemplars into the domains, starting with the Market content area and moving through the National and International content areas.

The basic procedure for domain item classifications was as follows: Panelists were seated at tables large enough and far enough apart so that the domain definitions for the domains within the content area could be laid out side by side. They were provided with the domain definitions for one content area at a time and were instructed to spread these out at the top of the table. They were then provided with the items in that content area, which were printed one per page. The panelists classified the items taking one economics content area at a time, by sorting the items into piles corresponding to the domain definitions. No item statistics were used in this process. An “unclassified” category was available at all times for panelists to use if they felt an item did not fit into any of the available domains within an area. The order in which content areas were processed was the same for all panelists but items were in a different random order for each panelist.

When panelists were finished classifying the items within a given content area, they reviewed their item classifications one domain at a time, reclassifying if needed. At this time they also responded to rating scale questions for each domain (Appendix I), indicating their level of agreement with each of the following statements: The domain represents a distinct and well defined area of content; the domain definition was helpful for classifying the items; overall the items I assigned to this domain were clearly related; a student's score on this domain (a percent correct score) would be useful to a teacher in identifying a student's strengths and weakness in economics.

Results

Panelist ratings of the domains as distinct and well defined, and as useful for identifying students’ strengths and weaknesses were high across all domains. On a scale from 1 (strongly disagree) to 5 (strongly agree), panelist mean ratings were close to or above 4 on both scales for all domains (Table 9).

In addition, internal and external consistency of the panelist classifications were assessed. Internal consistency is the degree to which all and/or the majority of the panelists in the study were in agreement with each other on the domain into which items were classified. The TACSS had determined that if there was a majority agreement among panelists for at least two-thirds of the items, then internal consistency was sufficient to suggest that the domains would be useful in the standard setting. There was majority agreement among panelists for 94% of the items overall and for between 88% and 100% of the items within each content area. These results indicated slightly greater agreement than results for a similar study in the 2005 mathematics contract where the majority of the panelists in the study were in agreement with each other on the domain into which items were classified for 88% of the items overall and for between 85% and 90% of the items within each content area (Figure 6).

Table 9: Mean Panelist Level of Agreement with Domain Statements

"The domain represents a distinct and well defined area of content" and
"A student's score on this domain (a percent correct score) would be useful to a teacher in identifying a student's strengths and weakness in economics"
(1 = Strongly Disagree, 5 = Strongly Agree)

	Domain is Distinct	Domain is Useful
Domain	Mean Score	
Market Economy		
1	4.29	3.86
2	4.43	4.57
3	4.14	4.57
4	4.29	4.57
5	4.14	4.71
6	4.71	4.00
7	4.11	4.71
8	4.29	4.71
9	4.00	4.86
10	4.14	4.57
National Economy		
1	4.57	4.57
2	4.43	4.29
3	4.57	4.57
4	4.00	4.43
5	3.86	4.29
6	4.57	4.71
7	4.71	4.86
8	4.29	4.71
9	4.57	4.86
International Economy		
1	4.71	4.71
2	5.00	4.71
3	4.29	4.14

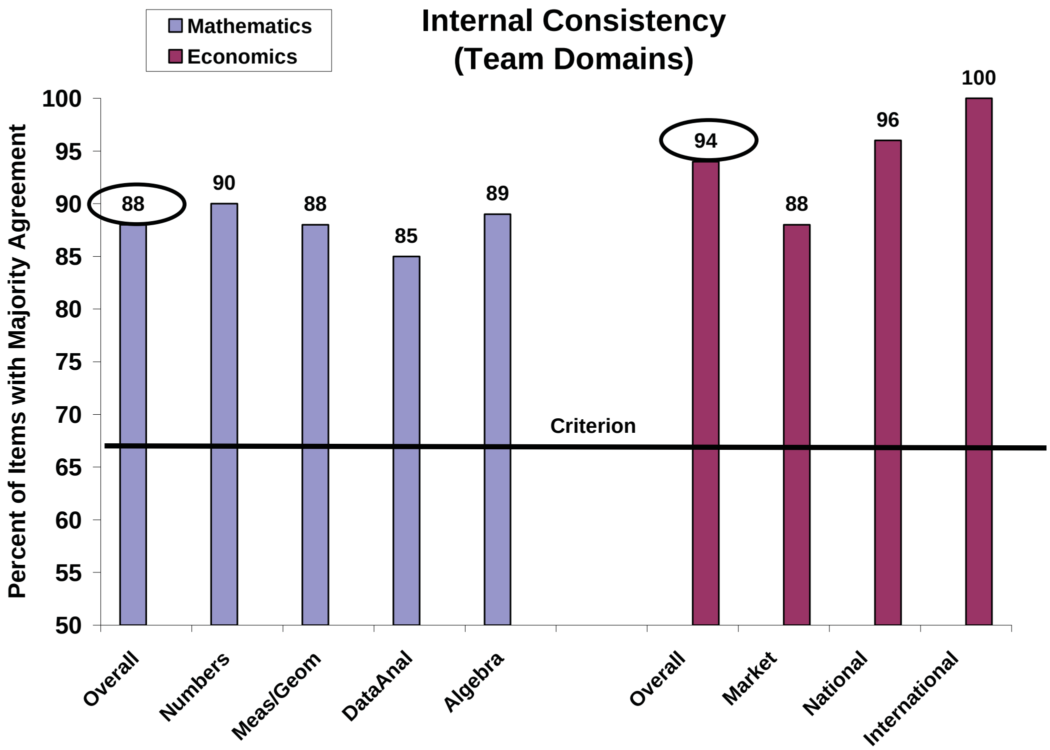


Figure 6. Summary of internal consistency in domain item classification for economics and mathematics.

To compute an index of external consistency, the panelists' final classification of each item was compared to each item's classification made by the domain development team just prior to the Domain Item Classification meeting. The panelists' classification of an item was defined as the domain selected for the item by the plurality of panelists. In the event of a tie, the panelists' classification was the more difficult domain within the content area. Prior to the Domain Item Classification Study, ACT and the TACSS had determined that if there was agreement between the teacher classification and the team classification for at least two-thirds of the items for which there was a plurality in the domain item classification results, then external consistency was sufficient to suggest that the domains would be useful in the standard setting. There was agreement between the panelists' classification and the team classification for 89% of the items overall. The consistency indices were also considered in comparison to similar data from the 2005 mathematics project and the results indicated slightly greater agreement for the economics domains (see Figure 7).

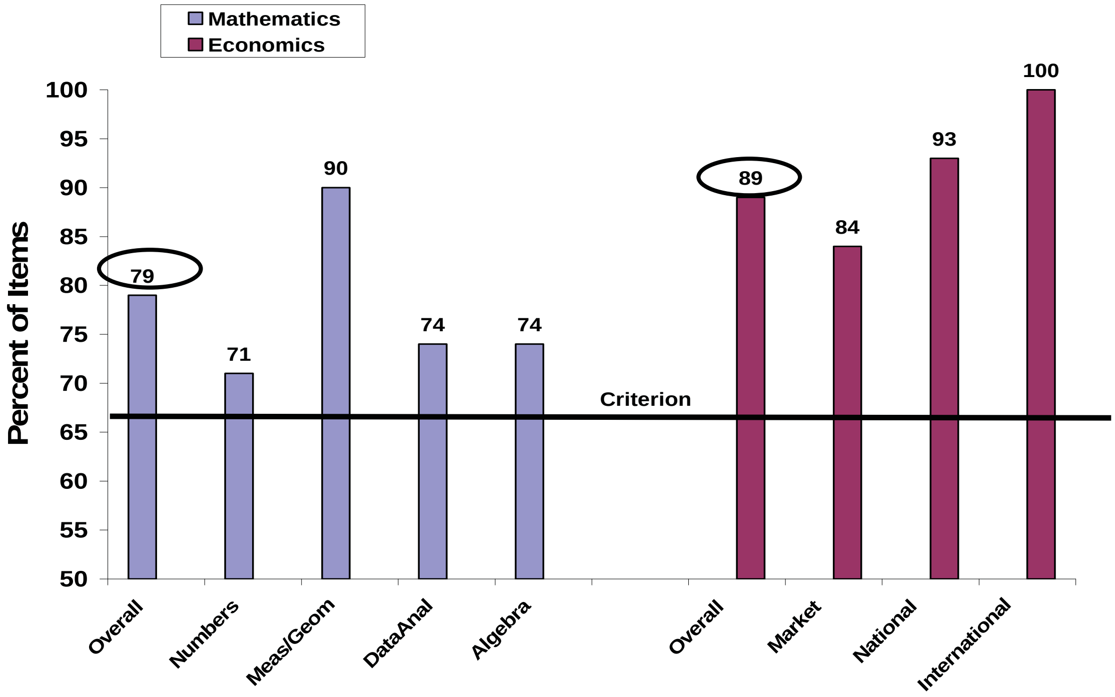


Figure 7. Summary of external consistency for economics and mathematics.

Finalize Domains

Task 11: Finalize Domains

Finally, in Task 11, the domain development team of two content experts and one ACT staff member reviewed the results from the Domain Item Classification meeting to finalize the domains. In particular, they reviewed the 15 items for which the Domain Item Classification panelists disagreed with the team classification. Seven of those items were reclassified into the panelist classification. Seven of those items remained in the team classification. And one item was reclassified entirely. Where necessary, the domain definitions were also minimally updated to reflect these changes.

Table 10 shows the titles of the economics domains that were ultimately used in the Mapmark with domains standard setting activities. A total of 22 domains were defined. The number of domains per economics content area ranged from three (in International Economy) to ten (in Market Economy). Complete definitions for each domain are provided in Appendix G. Although each definition was accompanied by one or two sample items, these items had not been released at the time of this process report and so are, therefore, not included in this report.

In the Pilot Study, Mapmark with domains 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 2006 assessment that were classified into the same domain, but also to the sample items.

Table 10: Titles of Teacher Domains by Content Area for Grade 12 Economics

Market Economy

Domain	Title
M1	Entrepreneurs
M2	Incentives
M3	Markets and Equilibrium
M4	Productivity, Income, and Capital
M5	Scarcity and Opportunity Cost
M6	Economic Institutions
M7	Competition
M8	Economic Role of Government
M9	Interaction of Supply, Demand, and Prices
M10	Additional Costs and Additional Benefits in Decision Making

National Economy

N1	Money, Loans, and Interest Rates
N2	Economic Growth
N3	Resource Allocation
N4	Government Programs and Taxes
N5	Spending, Income, and Related National Measures
N6	Real Interest Rates
N7	Inflation and Unemployment
N8	Money Supply
N9	Fiscal and Monetary Policy

International Economy

I1	Benefits and Costs of Trade
I2	Exchange Rates
I3	Tariffs

Figure 8 shows the expected percent correct curves for domains representing the National content area. The curves are based on items in the 2006 assessment, and illustrate that the domains do differ somewhat in difficulty, with some overlap.

The results of the Domain Development process and the Domain Item Classification Study were presented to the TACSS. The TACSS found that the domains were well developed and the recommendation was made that the domains be used in the standard setting process in the Pilot Study.

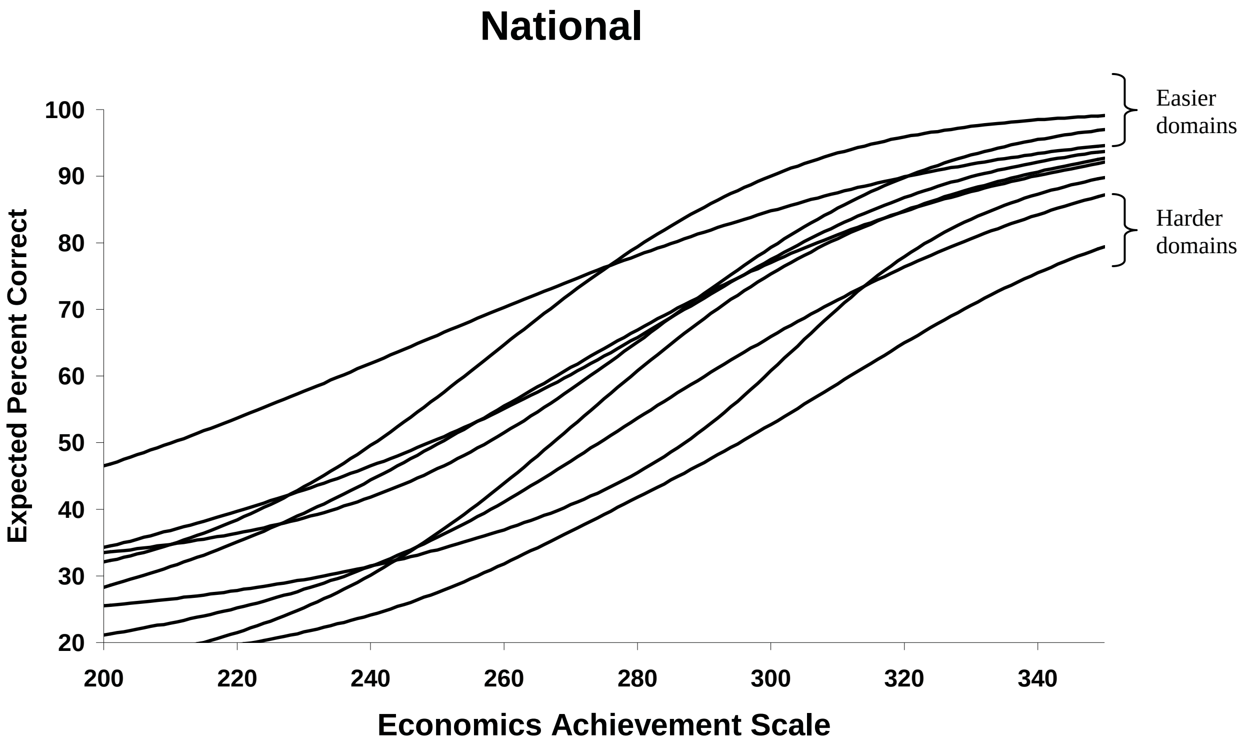


Figure 8. Expected percent correct curves for domains in the National content area.

Field Trial

On October 20-21, 2006, the first two rounds of the Mapmark with whole booklet feedback procedure were conducted with nine panelists at ACT headquarters in Iowa City, Iowa (see Appendix J for the Field Trial agenda). The panelists were recruited from within a 60-mile radius of Iowa City. There were four teachers, two nonteacher educators, and three members of the general public. There were six men and three women. The purpose of the Field Trial was to evaluate details of ACT's second recommended standard setting procedure for use in the Pilot Study, Mapmark with whole booklet feedback, and to allow ACT staff to gain experience with this method.

In order to conduct and evaluate essential aspects of the Mapmark procedure in two days, instead of the usual 3-day process, a number of activities were shortened. These included:

- administration of half of a test form to the panelists instead of a full test form;
- use of only half of the items on the test (five of the ten blocks);
- an abbreviated introduction and overview;
- delivery of feedback in round 3, but no setting of cut scores in round 3.

The Mapmark with whole booklet feedback procedure is described in detail in the ALS section of this document and a summary of the Field Trial is provided here. Observing the Field Trial were ACT staff Nancy Petersen, Jim Sconing, and Matt Schulz and NAGB Assistant Director of Psychometrics Susan Loomis. The Field Trial began with an overview of NAEP which was followed by an overview of the standard setting project, the Field Trial, and a description of the Mapmark with whole booklet feedback method and materials. Following this orientation, there were three rounds of the procedure.

Round 1

In round 1, which is the same in the Mapmark with domains and the whole booklet feedback procedures, panelists began with an exercise aimed at familiarizing them with the items in the assessment and, subsequently, the gradient of difficulty in the assessment content. In this exercise, they reviewed the items in order of difficulty from easiest to hardest. For each item, panelists were asked to identify the knowledge, skills, and abilities (KSAs) a student must demonstrate in order to correctly answer an item or to reach a given score point. For polytomously scored (constructed response) items, panelists identified what additional KSAs were necessary to reach the next score point on the item, and for all items, panelists identified what KSAs were needed to get one item correct above and beyond the KSAs required to get easier items of similar content correct.

Following the KSA review, panelists were provided with an overview of the Achievement Level Descriptions for economics (Appendix C) and a brief discussion before setting their round 1 cut scores by placing bookmarks in their Ordered Item Book.

Round 2

In round 2, panelists reviewed and discussed student performance on three forms of the assessment at the borderline and the middle of each achievement level. Panelists were asked to consider: (a) where their cut scores fell in relation to each booklet, and (b) if they felt that the booklets represented borderline or solid performance for the given achievement level. Following this review, a second set of cut scores were set.

Round 3

In round 3, panelists were provided with consequences data, indicating the percent of students within and at-or-above each achievement level as established by the cut score. Unlike the Pilot Study and ALS, the Field Trial panelists were not asked to set a third round of cut scores.

Process Evaluation

At the end of each round and each day, panelists were provided with an evaluation form in order to assess the areas of strength and areas for further development in the method. Evaluations were largely positive. On the final evaluation, panelists’ overall perception of the efficacy of the method in yielding reasonable cut scores were comparable to the panelists’ response to similar questions after the mathematics 2005 Pilot Study (Figure 9).

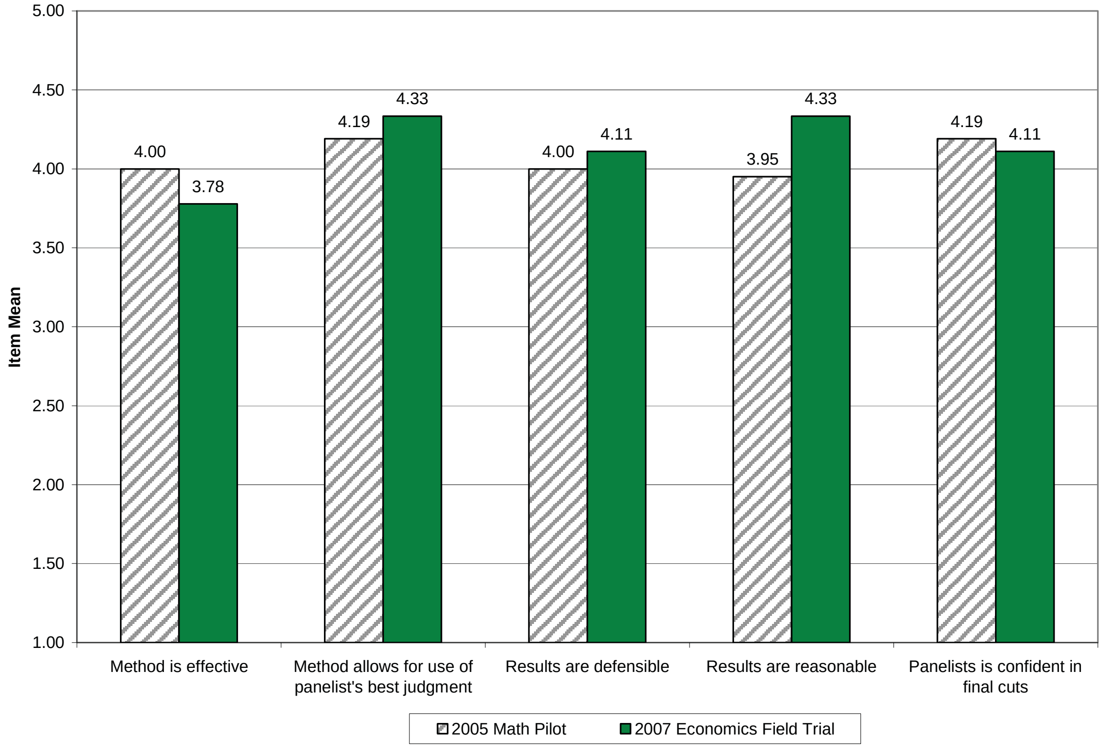


Figure 9. Mean ratings of economics Field Trial and 2005 math Pilot Study on key outcome questions.

In addition, process evaluation results indicated that whole booklet materials in the Mapmark process were effective and understood by panelists. Although all materials in the method received high ratings for their utility (mean ratings for materials were all above 4.00 on a scale from 1 = *not at all helpful* to 5 = *very helpful*), booklet-specific materials received the highest ratings (Table 11). In a debriefing, panelists indicated that the use of whole booklet examples clarified the meaning of performance at a given cut score in terms familiar to panelists outside of the standard setting process. Panelists were also comfortable combining the item-level information in their OIBs with the holistic information in the whole booklets for selecting a cut score.

Table 11: Mean Ratings of Materials for Mapmark with Whole Booklet Feedback

The _____ was/were:

(5 = Very Helpful, 3 = Somewhat Helpful, 1 = Not at all Helpful)

	Average Rating	
Material	2005 Math Study	Economics Field Trial
Ordered Item Book	4.76	4.56
Item Maps	4.24	4.44
Cut score dispersion chart	n/a	4.44
Consequences data	4.07	4.00
Item score table*	n/a	4.89
Booklet score chart*	n/a	4.75
Booklet score plot*	n/a	4.56

*These materials are specific to the Mapmark with whole booklet feedback method

A few suggestions were also made for improving the process. These included providing greater clarity on how the results of the KSA review in round 1 were to be used and providing more information in the round 1 cut score instructions on the definition of the *lower borderline* of performance. Adaptations to the method were made accordingly for implementation in the Pilot Study.

Results of the domain development process and the Field Trial were reviewed by ACT's internal Technical Advisory Team (TAT) and external Technical Advisory Committee on Standard Setting (TACSS). It was recommended that Mapmark with domains and Mapmark with whole booklet feedback be the two methods implemented in the Pilot Study.

Pilot Study

In the Pilot Study, conducted on December 6-9, 2006 at the Westin Hotel in St. Louis, Missouri, two fully operational standard setting methods, Mapmark with domains and Mapmark with whole booklet feedback, were implemented and subsequently compared. The Mapmark with domains method was fundamentally similar to the method ACT used to set achievement levels for the 2005 NAEP in grade 12 mathematics and is described below. The essential elements of the Mapmark with whole booklet feedback procedure are described in the ALS meeting process section of this report. Those elements in the Pilot Study Mapmark with whole booklet feedback procedure that differed from the ALS will be described below.

Observers and Participating Panelists

Thirty-three panelists participated in the Pilot Study (16 in Mapmark with whole booklet feedback and 17 in Mapmark with domains). The panelists for the Pilot Study, ALS, and Special Studies were recruited from the same national sample of nominees. The process of acquiring the sample of nominees and of recruiting panelists for participation in each meeting is described in the ALS section of this report. Efforts were made to ensure that the sample of nominees included teachers (55%), nonteacher educators (15%), and members of the general public (30%), as well as proportional representation by gender, race, and geographic region. Because responses from nonteacher educators and members of the general public have historically been poor, a proportionally large number of each of these types was included in the nominee sample. A total of 33 panelists participated (Table 12).

Table 12: Demographics of Panelists Participating in the Pilot Study

Type	Males	Females	Caucasian	African Am.	Asian	Hispanic	Other	Midwest	Northeast	South	West	**Total**
Teacher	10	13	19		2	1	1	5	5	8	5	23
Nonteacher	4	1	4		1			3	1		1	5
General Public	3	2	3	2				1		4		5
Total	17	16	26	2	3	1	1	9	6	12	6	33

Observing the Pilot Study were ACT staff Nancy Petersen and Jim Sconing, NAGB Assistant Director of Psychometrics Susan Loomis, and ACT TACSS members Barbara Dodd from the University of Texas-Austin and Mary Pitoniak from Educational Testing Service.

Advance Materials

Before the Pilot Study, all panelists were mailed materials that contained background information on setting achievement levels and on the purpose of the meeting. Participants in both methods were sent:

- National Assessment Governing Board brochure
- NAEP 2006 Economics Framework
- NAEP 2006 Economics Achievement Level Descriptions
- Hotel diagram and directions to the meeting
- Confidentiality agreement
- Press release form

In addition, panelists received two materials that were specific to the method in which they would participate. These were:

- Briefing Booklet
- Preliminary Agenda

The Mapmark with whole booklet feedback Briefing Booklet used in the Pilot Study differed from that used in the ALS. In the Pilot Study, this booklet provided a highly detailed and technical description of every step in the method. Nearly half of the Pilot Study panelists indicated that they felt this information was not useful, and some commented that the level of detail made the Booklet confusing. The Booklet was modified for the ALS to provide a higher-level overview of the method with less technical detail.

Orientation and Group Division

The Pilot Study began with all 33 panelists together. An overview of standard setting and a description of how panelists were selected for participation were provided by Project Director Christina Hamme Peterson. An orientation to the NAEP and the history of the National Assessment Governing Board was provided by NAGB Assistant Director of Psychometrics Susan Loomis. Christina Hamme Peterson also explained the reason for the Pilot Study and, so as to avoid cross-method contamination, asked panelists not to discuss procedures with panelists in the other method. Panelists were informed that the purpose of the Pilot Study was to evaluate two standard setting methods, and they were told that it was important to avoid discussing the methods and results with panelists from the other method group. Following this orientation, the participants were divided into two groups, one for each method, and went into separate, assigned rooms.

Within each method, the panelists were further divided into two rater groups, called group A and group B. Each group was seated at two tables of approximately four people, tables 1 and 2, and tables 3 and 4, respectively. The distribution of panelists across methods, groups, and tables within methods is given in Table 13. Efforts were made to ensure panelist diversity in rater groups, tables, and methods. The rater grouping allowed for the division of the assessment item pool (186 total items) into two pools of comparable size (about 113 items) and difficulty. Panelists from groups A and B looked at 41 items in common and between 72 and 73 unique items each. Assignment of items to groups was done largely on the basis of item blocks, however, the assignment of items to pools via blocks was modified slightly to accommodate the domains. 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 within each domain. This reassignment did not change the equivalence of the pools in other respects, and was not essential for the Mapmark with whole booklet feedback method, but was retained in the ALS. The item pools were equivalent with regard to: (a) mean and variation of item difficulty, (b) representation of content areas, and (c) percent of items of each type. Exact characteristics of the items in the pools are provided in Table 19 in the ALS section of this report.

Table 13: Distribution of Panelists across Methods and Tables

Mapmark with Whole Booklet Feedback

Table	Teacher	Nonteacher	GP	Caucasian	African Am.	Asian	Hispanic	Other	Male	Female	WE	MW	SO	NE
1	3	1	0	3	0	1	0	0	2	2	0	1	2	1
2	2	1	1	3	1	0	0	0	2	2	1	1	1	1
3	3	0	1	3	0	0	1	0	2	2	1	1	1	1
4	3	0	1	3	0	1	0	0	2	2	0	2	2	0
TOTAL	11	2	3	12	1	2	1	0	8	8	2	5	6	3

Mapmark with Domains

Table	Teacher	Nonteacher	GP	Caucasian	African Am.	Asian	Hispanic	Other	Male	Female	WE	MW	SO	NE
1	3	0	1	3	0	0	0	1	2	2	1	1	2	0
2	3	1	0	4	0	0	0	0	2	2	1	1	1	1
3	3	1	0	3	0	1	0	0	2	2	1	1	1	1
4	3	1	1	4	1	0	0	0	3	2	1	1	2	1
TOTAL	12	3	2	14	1	1	0	1	9	8	4	4	6	3

Mapmark with domains is a four-round method and Mapmark with whole booklet feedback is a three-round method (see Table 14). The two methods are identical through the first round and in the final round of the process, and they differ in the type of holistic feedback provided in the middle rounds. Each method group met separately after the general orientations on the first day. Process evaluation questionnaires were administered throughout the process. Agendas for both methods are provided in Appendix K.

Table 14: Summary of Activities in Each Round by Method

Round	Mapmark with whole booklet feedback	Mapmark with Domains
1	**KSA Review:** Review of items in order of difficulty and the gradient of knowledge, skills and abilities necessary to correctly answer each item	
2	**Holistic Feedback:** Review of student performance on actual booklets across the achievement scale	**Holistic Feedback:** Review of student performance on subareas of content, or domains, in the form of expected percent correct on each domain for students at the borderline of each achievement level
3	**Consequences Data:** Review of consequences data (proportion of students performing at or above and within each achievement level, based on the cut scores)	**Holistic Feedback:** Discussion of student performance on domains, and of how well that performance reflects the Achievement Level Descriptions
4		**Consequences Data:** Review of consequences data (proportion of students performing at or above and within each achievement level, based on the cut scores)

Mapmark with Domains

The Mapmark with domains method was facilitated by a process facilitator, Assistant Project Director Matt Schulz, and a content facilitator, retired economics teacher and NAEP Economics Framework committee member, Joy Joyce. As in the Mapmark with whole booklet feedback method, Mapmark with domains began with the administration of a form of the NAEP exam. Panelists scored their own performance on the NAEP and then an orientation to method-specific elements was provided.

Round 1

The first round of Mapmark with domains is identical to the first round of Mapmark with whole booklet feedback, as described in the ALS section of this report. At the conclusion of the first round, panelists had established cut scores for all three achievement levels.

Feedback from Round 1

Feedback from round 1 consisted of: (a) group cut scores (median), (b) cut score distribution, (c) rater location, and (d) domain scores. In addition to providing the numerical values of cut scores, feedback was shown on item maps and domain score charts to focus panelists' attention on the intended, criterion-referenced meaning of cut scores. Just as in the Mapmark with whole booklet feedback method, the group cut scores and a panelist's bookmarked items were marked on the Primary Item Map.

Before panelists were shown domain score feedback, they were given a presentation on how and why the domains were defined. The presentation included a brief overview of the domain development process and described the intended attributes of the domains (see Domain Development section of this report).

Expected percent correct curves based on a select group of domains within the National and Market content areas were shown to illustrate that domains varied in difficulty and to show panelists where the expected domain scores in the Percent Correct Table (PCT) came from. Figure 10 shows the curves for these domains as illustrated in the Pilot Study presentation. Vertical lines in this plot represent the round 1 cut scores. The dashed, horizontal line represents a 67% criterion for mastery of the domains.

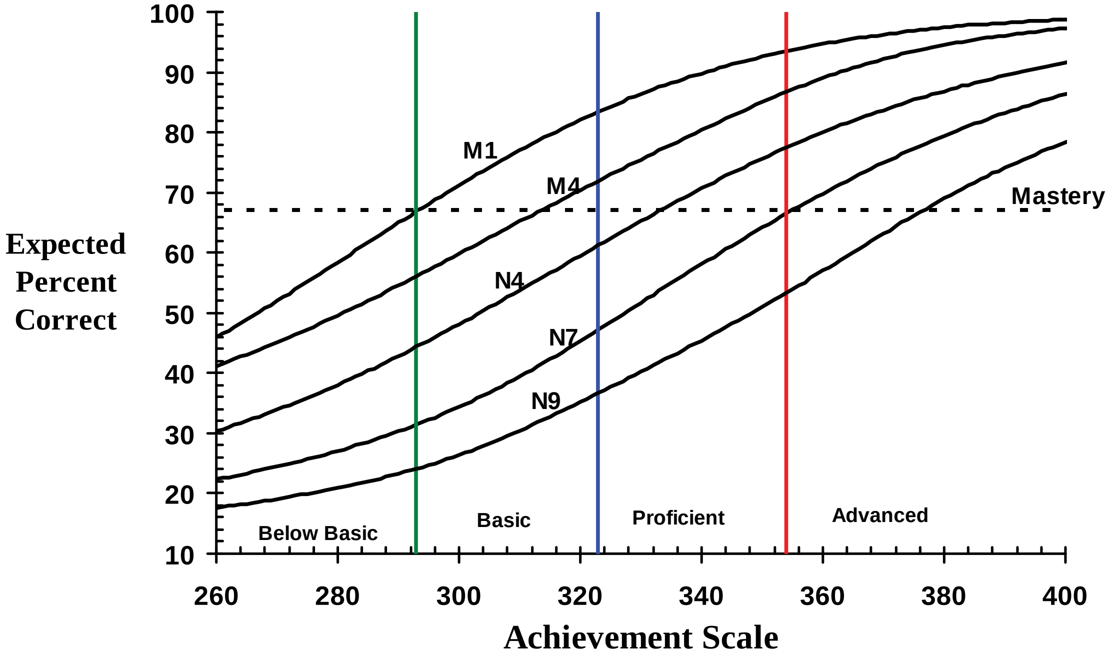


Figure 10. Percent correct curves for select domains in the National and Market content areas, with vertical lines showing location of round 1 cut scores and a horizontal line representing a 67% criterion for mastery.

A Percent Correct Table (PCT) was used to show the expected percent correct scores corresponding to the cut scores. The PCT for round 1 cut scores is shown in Figure 11. This table shows the domain titles and, for each domain, the expected percent correct scores conditional on the lower boundary of the Basic, Proficient, and Advanced achievement levels, as defined by the group 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 cut scores, based on the domain scores in the PCT.
The highest, lowest, and closest to 67% domain scores for the Proficient cut score in the PCT were circled (see Figure 11) to draw panelists’ attention to the fact that in one of their

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

Content Area	Domain	Expected Percent Correct at Lower Borderline of…		
		Basic	Proficient	Advanced
Market	M1. Entrepreneurs	58%	83%	97%
	M2. Incentives	53%	85%	99%
	M3. Markets and Equilibruim	57%	82%	97%
	M4. Productivity, Income, and Capital	49%	71%	94%
	M5. Scarcity and Opportunity Cost	49%	70%	93%
	M6. Competition	40%	74%	96%
	M7. Economic Institutions	39%	66%	95%
	M8. Internation of Supply, Demand, and Prices	38%	59%	83%
	M9. Economic Role of Government	33%	55%	91%
	M10. Additional Costs and Benefits in Decision Making	31%	56%	92%
National	N1. Money, Loans, and Interest Rates	56%	74%	91%
	N2. Spending, Income, and Related National Measures	41%	71%	97%
	N3. Resource Allocation	42%	59%	89%
	N4. Economics Growth and Productivity	38%	60%	87%
	N5. Government Programs and Taxes	38%	57%	92%
	N6. Real Interest Rates	24%	51%	87%
	N7. Inflation and Unemployment	27%	46%	79%
	N8. Money Supply	29%	40%	82%
	N9. Fiscal and Monetary Policy	21%	36%	68%
International	I1. Benefits and Costs of Trade	39%	66%	89%
	I2. Exchange Rates	35%	55%	79%
	I3. Tariffs	24%	43%	70%

Figure 11. Percent Correct Table with the highest, lowest, and closest to 67% circled in the Proficient achievement level.

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 domain for every scale score within a range that goes from 10 points below the "low" cut score to 10 points above the "high" cut score from the previous round.

Figure 12 shows the DSC for the Proficient achievement level with the location of Panelist 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.

	Scale	Market Economy										National Economy									International		
	Score	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	N1	N2	N3	N4	N5	N6	N7	N8	N9	I1	I2	I3
	368	96	98	96	92	90	94	93	80	87	89	89	95	85	83	88	83	74	74	62	86	75	66
	367	96	98	96	91	90	94	93	80	86	88	89	95	84	83	87	82	74	74	61	86	75	65
	366	96	98	95	91	90	94	92	79	86	88	88	94	84	82	87	82	73	73	61	85	75	65
	365	96	98	95	91	90	93	92	79	85	87	88	94	84	82	86	82	73	72	60	85	74	65
	364	95	98	95	90	89	93	92	79	85	87	88	94	83	82	86	81	72	71	60	85	74	64
	363	95	98	95	90	89	93	91	78	84	86	88	94	83	81	85	81	71	70	59	84	73	64
	362	95	98	95	90	89	93	91	78	84	86	87	93	82	81	85	80	71	69	58	84	73	63
High	361	95	97	95	89	88	92	91	78	83	85	87	93	82	80	84	80	70	68	58	84	73	63
	360	95	97	94	89	88	92	90	77	82	85	87	93	81	80	84	79	70	67	57	83	72	62
	359	95	97	94	89	88	92	90	77	82	84	87	92	81	80	83	79	69	66	56	83	72	62
	358	94	97	94	88	87	92	90	77	81	84	86	92	80	79	82	78	69	66	56	83	71	61
	357	94	97	94	88	87	91	89	76	80	83	86	92	80	79	82	78	68	65	55	82	71	61
	356	94	97	94	88	87	91	89	76	80	83	86	91	79	78	81	77	68	64	55	82	71	60
	355	94	97	93	87	86	91	88	75	79	82	85	91	79	78	81	77	67	63	54	82	70	60
	354	94	96	93	87	86	91	88	75	78	82	85	90	78	78	80	76	67	62	53	81	70	59
	353	93	96	93	86	85	90	87	75	78	81	85	90	78	77	79	75	66	61	53	81	69	59
	352	93	96	93	86	85	90	87	74	77	80	85	90	77	77	79	75	65	60	52	81	69	58
	351	93	96	92	85	85	90	86	74	76	80	84	89	76	76	78	74	65	59	52	80	68	58
	350	93	96	92	85	84	89	86	73	75	79	84	89	76	76	77	73	64	58	51	80	68	57
	349	92	96	92	85	84	89	85	73	75	79	84	88	75	75	77	73	64	57	50	79	68	57
	348	92	95	92	84	83	89	85	73	74	78	83	88	75	75	76	72	63	56	50	79	67	56
	347	92	95	91	84	83	88	84	72	73	77	83	87	74	74	75	72	62	55	49	79	67	56
	346	92	95	91	83	82	88	84	72	72	76	83	87	74	74	75	71	62	55	49	78	66	55
	345	91	95	91	83	82	88	83	71	72	76	82	86	73	73	74	70	61	54	48	78	66	55
	344	91	95	91	82	82	87	83	71	71	75	82	86	72	73	73	69	61	53	48	77	65	54
	343	91	94	90	82	81	87	82	70	70	74	82	85	72	72	72	69	60	52	47	77	65	54
	342	90	94	90	81	81	86	81	70	69	74	81	85	71	72	72	68	59	51	47	76	65	53
	341	90	94	90	81	80	86	81	70	68	73	81	84	71	71	71	67	59	51	46	76	64	53
	340	90	93	89	80	80	85	80	69	68	72	81	84	70	71	70	66	58	50	45	76	64	52
	339	90	93	89	80	79	85	79	69	67	71	80	83	69	70	70	66	58	49	45	75	63	52
	338	89	93	89	79	79	84	79	68	66	70	80	83	69	70	69	65	57	49	44	75	63	51
	337	89	93	88	79	78	84	78	68	65	70	80	82	68	69	68	64	56	48	44	74	62	51
	336	89	92	88	78	78	84	77	67	65	69	79	81	68	69	67	63	56	47	43	74	62	50
	335	88	92	88	78	77	83	77	67	64	68	79	81	67	68	67	62	55	47	43	73	61	50
	334	88	91	87	78	77	82	76	66	63	67	79	80	66	68	66	62	54	46	42	73	61	49
	333	88	91	87	77	76	82	75	66	63	66	78	79	66	67	65	61	54	46	42	72	60	49
	332	87	91	87	77	76	81	74	65	62	66	78	79	65	66	64	60	53	45	41	72	60	48
	331	87	90	86	76	75	81	74	65	61	65	77	78	65	66	64	59	52	44	41	71	59	48
	330	86	90	86	76	75	80	73	64	61	64	77	77	64	65	63	58	52	44	40	71	59	47
	329	86	89	85	75	74	80	72	64	60	63	77	77	64	65	62	57	51	43	40	70	58	47
	328	86	89	85	75	74	79	71	63	59	62	76	76	63	64	62	57	50	43	39	70	58	46
	327	85	88	84	74	73	78	71	63	59	61	76	75	62	64	61	56	50	42	39	69	58	46
	326	85	88	84	74	73	78	70	62	58	61	76	75	62	63	60	55	49	42	38	68	57	45
	325	84	87	84	73	72	77	69	62	57	60	75	74	61	62	59	54	49	42	38	68	57	45
	324	84	87	83	72	72	76	68	61	57	59	75	73	61	62	59	53	48	41	37	67	56	44
	323	83	86	83	72	71	76	67	60	56	58	74	72	60	61	58	52	47	41	37	67	56	44
	322	83	86	82	71	71	75	67	60	56	57	74	72	60	61	57	51	47	40	36	66	55	44
Median--->	321	83	85	82	71	70	74	66	59	55	56	74	71	59	60	57	51	46	40	36	66	55	43
	320	82	85	81	70	70	74	65	59	55	56	73	70	59	60	56	50	45	39	35	65	54	43
	319	82	84	81	70	69	73	64	58	54	55	73	69	58	59	55	49	45	39	35	64	54	42
	318	81	83	80	69	69	72	64	58	53	54	72	69	58	58	55	48	44	39	34	64	53	42
	317	81	83	80	69	68	71	63	57	53	53	72	68	57	58	54	47	44	38	34	63	53	41
	316	80	82	79	68	68	71	62	57	52	52	72	67	57	57	53	46	43	38	33	63	52	41
	315	80	81	79	68	67	70	61	56	52	52	71	66	56	57	53	46	42	38	33	62	52	40
	314	79	81	78	67	67	69	60	55	51	51	71	66	56	56	52	45	42	37	32	61	51	40
	313	79	80	78	67	66	68	60	55	51	50	70	65	55	56	52	44	41	37	32	61	51	39
	312	78	79	77	66	66	67	59	54	50	49	70	64	55	55	51	43	41	37	31	60	50	39
	311	78	78	76	66	65	66	58	54	50	49	70	63	54	54	50	42	40	36	31	59	50	38
	310	77	78	76	65	65	66	58	53	49	48	69	62	54	54	50	42	39	36	31	59	49	38
	309	76	77	75	65	64	65	57	53	49	47	69	62	53	53	49	41	39	36	30	58	49	37
	308	76	76	75	64	64	64	56	52	48	46	68	61	53	53	49	40	38	35	30	58	48	37
	307	75	75	74	64	63	63	55	52	48	46	68	60	52	52	48	39	38	35	29	57	48	36
	306	75	75	74	63	63	62	55	51	47	45	67	59	52	52	48	39	37	35	29	56	47	36
	305	74	74	73	63	62	61	54	50	46	44	67	58	51	51	47	38	37	35	28	56	47	35
	304	74	73	72	62	62	60	53	50	46	44	67	58	51	50	47	37	36	34	28	55	46	35
	303	73	72	72	61	61	60	53	49	45	43	66	57	51	50	46	36	36	34	28	54	46	34
	302	72	71	71	61	61	59	52	49	45	42	66	56	50	49	46	36	35	34	27	54	46	34
	301	72	70	71	60	60	58	51	48	44	42	65	55	50	49	45	35	35	33	27	53	45	33
	300	71	70	70	60	60	57	51	48	44	41	65	55	49	48	45	34	34	33	26	52	45	33
	299	71	69	69	59	59	56	50	47	43	40	64	54	49	48	44	34	34	33	26	52	44	33
Panelist X	298	70	68	69	59	59	55	49	47	43	40	64	53	48	47	44	33	34	33	26	51	44	32
	297	69	67	68	58	58	54	49	46	42	39	64	52	48	47	43	33	33	32	25	50	43	32
	296	69	66	67	58	58	54	48	46	42	39	63	52	48	46	43	32	33	32	25	50	43	31
	295	68	65	67	57	57	53	48	45	41	38	63	51	47	45	43	31	32	32	25	49	42	31
	294	68	65	66	57	57	52	47	45	41	37	62	50	47	45	42	31	32	32	24	48	42	30
	293	67	64	66	56	56	51	46	44	40	37	62	50	47	44	42	30	31	32	24	48	41	30
	292	66	63	65	56	56	50	46	44	40	36	62	49	46	44	42	30	31	31	24	47	41	29
	291	66	62	64	55	55	49	45	43	39	36	61	48	46	43	41	29	31	31	24	46	40	29
	290	65	61	64	55	55	49	45	43	39	35	61	48	45	43	41	29	30	31	23	46	40	29
	289	64	60	63	54	54	48	44	42	38	35	60	47	45	42	41	28	30	31	23	45	40	28
	288	64	60	62	54	54	47	44	42	38	35	60	46	45	42	40	28	30	30	23	44	39	28
	287	63	59	62	53	53	46	43	41	37	34	59	46	44	41	40	27	29	30	23	44	39	27
	286	62	58	61	53	53	45	43	41	37	34	59	45	44	41	40	27	29	30	22	43	38	27
	285	62	57	61	52	52	45	42	40	36	33	59	45	44	40	39	26	29	30	22	43	38	26
	284	61	57	60	52	52	44	42	40	36	33	58	44	43	40	39	26	28	30	22	42	37	26
	283	60	56	59	51	51	43	41	39	35	32	58	43	43	39	39	25	28	29	22	41	37	26
	282	60	55	59	51	51	42	41	39	35	32	57	43	43	39	38	25	28	29	21	41	37	25
	281	59	54	58	50	50	42	40	38	34	32	57	42	42	39	38	24	27	29	21	40	36	25
	280	58	54	58	50	50	41	40	38	34	31	57	42	42	38	38	24	27	29	21	40	36	24
Low	279	58	53	57	49	49	40	39	38	33	31	56	41	42	38	38	24	27	29	21	39	35	24
	278	57	52	56	49	49	40	39	37	33	31	56	41	41	37	38	23	27	29	21	38	35	24
	277	57	51	56	48	48	39	39	37	32	30	55	40	41	37	37	23	26	28	20	38	35	23
	276	56	51	55	48	48	38	38	36	32	30	55	40	41	36	37	23	26	28	20	37	34	23
	275	55	50	55	47	47	38	38	36	31	30	55	39	40	36	37	22	26	28	20	37	34	23
	274	55	50	54	47	47	37	37	35	31	29	54	39	40	36	37	22	25	28	20	36	33	22
	273	54	49	54	46	46	36	37	35	31	29	54	38	40	35	36	22	25	28	20	36	33	22

Figure 12. 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 circling their own cut score on the DSC, panelists were able to see how much difference there was between their cut score and the group cut score 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.

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 domain. 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 content area, from left to right on the Domain Item Map.

Figure 13 shows a section of the Domain Task 1 form for group A. The complete form was three pages, one for each content area, and included all 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.

Market Economy Domain	Item Handle	I see how this item is like other items in its domain. (Check ✓)		
		Yes	Not Sure	No
M1) Entrepreneurs	M7			
	M44			
	M88			
M2) Incentives	M5			
	M6			
	M21			
	M41			
	M56			
M3) Markets and Equilibrium	M12			
	M38			
	M40			
	M123			

Figure 13. Section of Domain Task 1 form for group A.

Panelists responded to the question, "I see how this item is like other items in its domain," for each item in their pool that was classified into a 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. Within the content area, domains were ordered by difficulty as indicated by the average p-value of the items within the domain. Items were ordered by difficulty within domain within content area. Before considering the items within a given domain, panelists read the domain definition and looked at the sample items.

Materials for Domain Task 1 included a Domain Ordered Item Book (DOIB). The DOIB contained the 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 content area. Figure 14 shows the Domain Item Map for the International Economy content area. Panelists observed the trend of increasing difficulty in the domains as one goes from left to right in the DIM. Facilitators also drew panelists' attention to the variability of item difficulty within the domains. This variability means that no single item is a very reliable indication of the difficulty of a more general skill.

Scale	International Economy Domains					
	I1			I2		I3
above						P25_2
447						
444						
441						P22_4
438						
435						
432						
429						
426						
423						
420	M154					
417						
414						
411						
408						
405				P17_2		
402						
399						
396						
393						
390						P25_1
387						
384						M149
381						
378						
375						
372						
369						
366						
363						
360						M136
357						
354	M133					
351						
348						
345				M117		P22_3
342	M106			M109	M112	
339						
336	M99			P17_1		M97
333	M89	P5_3	M90			
330	M85					
327						
324						
321						
318						
315	M55					P22_2
312						
309	M53					
306				M48		
303	P5_2					
300	P5_1	M37				
297	M33					
294	M31					
291						P22_1
288						
285						
282						
279	M14					
276						
273						
270						
267						
264						
261						
258						
255						
252						
249						
246						
243						
240						
237						
234						
231						
Border Adv.:	89%			79%		70%
Border Prof.:	66%			55%		43%
Border Basic:	39%			35%		24%

Figure 14. Domain Item Map for International Economy content area.

As panelists worked through the items within a 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. [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 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%.

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 cut score should be higher, lower, or were OK as a standard of lower borderline performance for a given achievement level. Figure 15 shows the form that was used to collect panelists' judgments about domain scores associated with the round 1 cut score for Proficient. Similar forms were used for the other achievement levels.

Content Area	**Domain**		I think the percentage correct score at the **lower borderline of PROFICIENT** should be**...** (check the appropriate cell)		
		EPC	**Lower**	**OK**	**Higher**
Market	M1. Entrepreneurs	83%			
	M2. Incentives	85%			
	M3. Markets and Equilibruim	82%			
	M4. Productivity, Income, and Capital	71%			
	M5. Scarcity and Opportunity Cost	70%			
	M6. Competition	74%			
	M7. Economic Institutions	66%			
	M8. Internation of Supply, Demand, and Prices	59%			
	M9. Economic Role of Government	55%			
	M10. Additional Costs and Benefits in Decision Making	56%			
National	N1. Money, Loans, and Interest Rates	74%			
	N2. Spending, Income, and Related National Measures	71%			
	N3. Resource Allocation	59%			
	N4. Economics Growth and Productivity	60%			
	N5. Government Programs and Taxes	57%			
	N6. Real Interest Rates	51%			
	N7. Inflation and Unemployment	46%			
	N8. Money Supply	40%			
	N9. Fiscal and Monetary Policy	36%			
International	I1. Benefits and Costs of Trade	66%			
	I2. Exchange Rates	55%			
	I3. Tariffs	43%			

Figure 15. Domain Task 2 form for the Proficient achievement level.

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.

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 cut score.

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 group cut score. If all of the checks were in the *higher* column, one would probably want to select a cut score higher than the round 1 group cut score.

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 group cut scores. Panelists could see the change in the group cut score 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 group cut scores. 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 12 only this time to show the location of the round 2 group cut score, 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 round 2 group cut scores. They placed flags on these pages.

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 chose 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.

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.

Round 4

Round 4 of the Mapmark with domains method is identical to round 3 of the Mapmark with whole booklet feedback method described in the ALS section of this report.

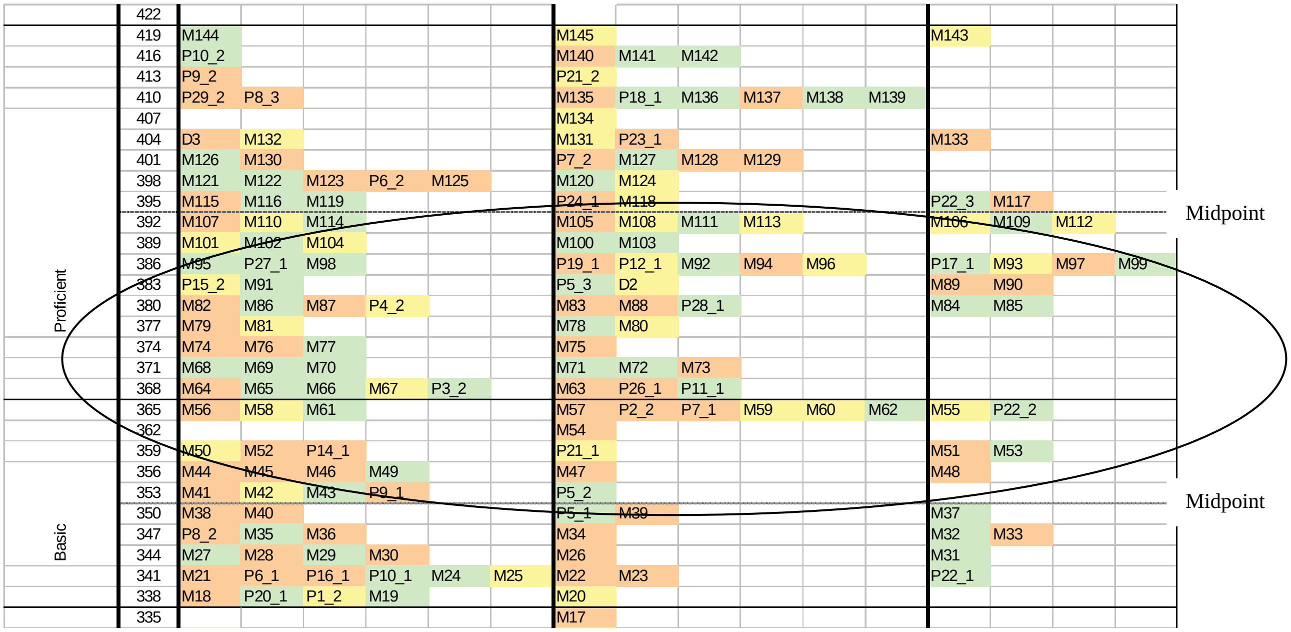
Consequences Questionnaire and Exemplar Item Rating, Both Methods

At the conclusion of both methods, panelists were asked to complete a consequences questionnaire, a copy of which is included in Appendix L. 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. In the Pilot Study, panelists were not provided with any additional consequences information to help them determine what changes to their cut scores would mean in terms of the proportion of students scoring at or above their new recommended cut scores. At the suggestion of ACT's Technical Advisory Committee on Standard Setting (TACSS), a chart called the Cut Score Proportion Chart was provided in the ALS to allow panelists to see the relative impact of changing from one cut score to another if they wanted to raise or lower the cut score. This is described in the ALS section of this report, and is the only difference between the consequences questionnaire procedure used in the Pilot Study and the procedure used in the ALS.

In addition, panelists were asked to rate exemplar items for their suitability as examples of what students know and can do at each achievement level. Potential exemplars were drawn from blocks of the assessment that were selected for eventual release to the public. These were blocks 1, 2, and 4.

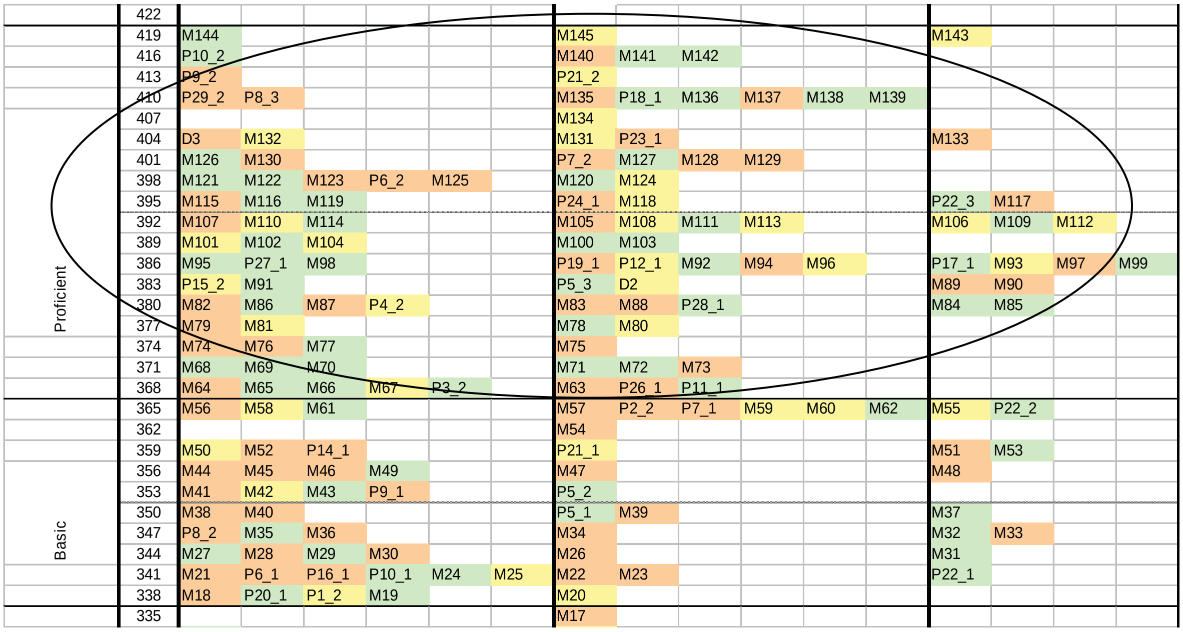
Exemplar items were identified for review by panelists from released items such that a student at the midpoint of the achievement level would have at least a 67% chance of answering that item correctly. Items were identified from the midpoint of the achievement level of interest to the midpoint of the achievement level below (see Figure 16). For each method, the exemplars were drawn on the basis of the group cut scores resulting from the final round of that method.



	422																	
	419	M144						M145						M143				
	416	P10_2						M140	M141	M142								
	413	P9_2						P21_2										
	410	P29_2	P8_3					M135	P18_1	M136	M137	M138	M139					
Proficient	407							M134										
	404	D3	M132					M131	P23_1					M133				
	401	M126	M130					P7_2	M127	M128	M129							
	398	M121	M122	M123	P6_2	M125		M120	M124									
	395	M115	M116	M119				P24_1	M118					P22_3	M117			
	392	M107	M110	M114				M105	M108	M111	M113			M106	M109	M112		Midpoint
	389	M101	M102	M104				M100	M103									
	386	M95	P27_1	M98				P19_1	P12_1	M92	M94	M96		P17_1	M93	M97	M99	
	383	P15_2	M91					P5_3	D2					M89	M90			
	380	M82	M86	M87	P4_2			M83	M88	P28_1				M84	M85			
	377	M79	M81					M78	M80									
	374	M74	M76	M77				M75										
	371	M68	M69	M70				M71	M72	M73								
	368	M64	M65	M66	M67	P3_2		M63	P26_1	P11_1								
	365	M56	M58	M61				M57	P2_2	P7_1	M59	M60	M62	M55	P22_2			
	362							M54										
	359	M50	M52	P14_1				P21_1						M51	M53			
Basic	356	M44	M45	M46	M49			M47						M48				
	353	M41	M42	M43	P9_1			P5_2										
	350	M38	M40					P5_1	M39					M37				Midpoint
	347	P8_2	M35	M36				M34						M32	M33			
	344	M27	M28	M29	M30			M26						M31				
	341	M21	P6_1	P16_1	P10_1	M24	M25	M22	M23					P22_1				
	338	M18	P20_1	P1_2	M19			M20										
	335							M17										

Figure 16. Example of range used in the Pilot Study for selection of potential exemplar items for the Proficient level.

After the Pilot Study, COSDAM determined that potential exemplar items would be identified only from within the achievement level, such that a student at the top of that level would have at least a 67% chance of answering the item correctly (see Figure 17). This was the method of selecting exemplars used in the ALS.



	422																
	419	M144						M145						M143			
	416	P10_2						M140	M141	M142							
	413	P9_2						P21_2									
	410	P29_2	P8_3					M135	P18_1	M136	M137	M138	M139				
Proficient	407							M134									
	404	D3	M132					M131	P23_1					M133			
	401	M126	M130					P7_2	M127	M128	M129						
	398	M121	M122	M123	P6_2	M125		M120	M124								
	395	M115	M116	M119				P24_1	M118					P22_3	M117		
	392	M107	M110	M114				M105	M108	M111	M113			M106	M109	M112	
	389	M101	M102	M104				M100	M103								
	386	M95	P27_1	M98				P19_1	P12_1	M92	M94	M96		P17_1	M93	M97	M99
	383	P15_2	M91					P5_3	D2					M89	M90		
	380	M82	M86	M87	P4_2			M83	M88	P28_1				M84	M85		
	377	M79	M81					M78	M80								
	374	M74	M76	M77				M75									
	371	M68	M69	M70				M71	M72	M73							
	368	M64	M65	M66	M67	P3_2		M63	P26_1	P11_1							
	365	M56	M58	M61				M57	P2_2	P7_1	M59	M60	M62	M55	P22_2		
	362							M54									
	359	M50	M52	P14_1				P21_1						M51	M53		
Basic	356	M44	M45	M46	M49			M47						M48			
	353	M41	M42	M43	P9_1			P5_2									
	350	M38	M40					P5_1	M39					M37			
	347	P8_2	M35	M36				M34						M32	M33		
	344	M27	M28	M29	M30			M26						M31			
	341	M21	P6_1	P16_1	P10_1	M24	M25	M22	M23					P22_1			
	338	M18	P20_1	P1_2	M19			M20									
	335							M17									

Figure 17. Example of range used in the ALS for selection of potential exemplar items for the Proficient level.

Process Evaluation

At the end of each round and day, evaluations were collected from all panelists in the two methods. The evaluations were designed to ascertain any areas of confusion in the tasks and materials and to allow for comparison between the two methods. Figure 18 shows average ratings on a 1-5 scale (1 = Not at all, 5 = To a great extent) on summary questions. Included are results from both Pilot Study methods and the 2005 mathematics ALS method (Mapmark with domains) for comparison. Differences between the methods were not substantially different.

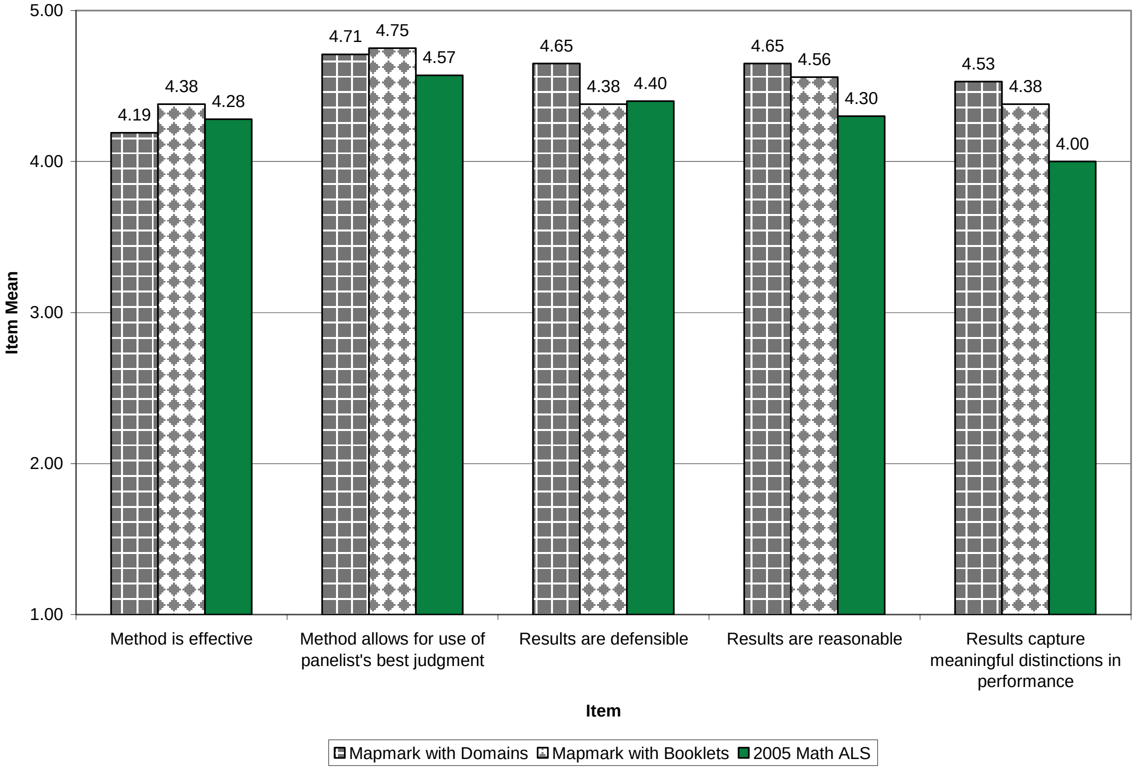


Figure 18. Average responses to summary process questions by method.

Statistical analyses of cut scores showed that both methods had acceptable reliability. Both item pool effects and table effects were modest and had virtually disappeared by the final round. Estimates of the standard error of cut scores at the first round ranged from 4 to 5 points and at the final round ranged from 1 to 3 points for both methods. Unfortunately, ACT can recommend no single, sure method for estimating the standard error of the final 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 cut score from the previous round. In past standard setting projects, panelist cut scores after round 1 seem to have been influenced by the group cut score and cut score distribution. There is a regression to the round 1 group cut score (ACT, Inc., 2005). Estimates of the standard error of the final cut score do not account for a fundamental regression to the median of previous rounds, motivated by panelists' desire for conformity, as well as for the effects of criterion-referenced feedback. For this reason, estimates of the standard error at the final round tend to be smaller and are more likely to underestimate differences between replications of a method using the same item pools but different groups of panelists. In addition, cut scores established in rounds 2 and 3 are based on the baseline established in the first round, and the group cut scores do not tend to vary substantially from the previous round. For this reason, an understanding of the differences between cut scores is most informed by an analysis of results from round 1.

Cut scores from the two methods were also highly similar to one another. On a 300-point scale, ranging from 203 to 503, the Basic cut scores from the two methods differed by just three points (335 for domains vs. 338 for booklets), the Proficient cut scores differed by three points (371 for domains vs. 368 for booklets), and the Advanced cut scores differed by five points (426 for domains vs. 421 for booklets). These differences were not statistically significant at p<.05. After viewing the achievement level percentages, a majority of panelists in both methods indicated on their consequences questionnaire that the cut scores should not be changed. Table 15 shows the corresponding achievement level percentages.

Table 15: Grade 12 Economics Achievement Level Percentages by Method

Method	Below Basic	Basic	Proficient	Advanced
Mapmark with domains	27%	40%	32%	1%
Mapmark with booklets	30%	34%	35%	1%

ACT presented the results of the Pilot Study to the Governing Board Committee on Standards, Design, and Methodology (COSDAM) on February 2, 2007. ACT and its TACSS indicated a slight preference for the Mapmark with whole booklet feedback method based on the following points.

- Resulting cut scores and achievement level percentages from the two methods were highly similar to one another.

- Both methods have good evidence of procedural validity and the achievement level percentages produced by either method are likely to be viewed as reasonable.

- Mapmark with domains is more costly to implement than Mapmark with whole booklet feedback because it requires an additional initial investment to develop content domains.

- Mapmark with whole booklet feedback can serve as a model for states to adopt that is more cost effective than Mapmark with domains and is similar to the most commonly used procedure at the state level: Bookmark. Also, as it does not require domains, it can be used across all content areas, provides holistic feedback in a format very familiar to educators, and can easily be adapted for use across multiple grades at the same time.

COSDAM chose to have ACT implement the Mapmark with whole booklet feedback method for the operational ALS.

THE ACHIEVEMENT LEVEL SETTING PROCESS

Achievement level setting in NAEP refers to the overall process through which cut scores and 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 recruitment of panelists and mailing of advance materials.

Panelist Selection

ACT implemented the same basic design for selecting panelists to set achievement levels that ACT had used for the 2005 ALS process. This design was used for the Special Studies, Pilot Study, and the ALS. Primary requirements based on NAGB policy 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 three steps: Selection of school districts, identification of nominators, and recruitment of panelists. Nominations of panelists were requested from a sample of school districts, teachers, state education associations, colleges/universities, and businesses and professional associations. Panelists were selected for recruitment from the sample of nominees. The same sample of nominees was used for the Pilot Study, Special Studies, and ALS, and the same methods of selection were used for all three studies. 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 unit of sampling. One sample of districts was drawn to identify nominators of teachers, nonteacher educators, and the general public. The stratified random sample was drawn from the Market Data Retrieval (MDR) database of school districts. The one sample provided nominators for each of the three studies; the Pilot Study, the Special Studies, and the ALS. 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 4,494 public districts and 798 private schools were sampled. Please see Table 16 for the distribution of school districts sampled.

Table 16: Distribution of School Districts Sampled

	Public Districts	Private Districts	Total
Nominators for all three types	4,494 (85%)	798 (15%)	5,292

Identification of Nominators

The sample of public school districts was used to identify nominations of teacher, nonteacher educators, and general public representatives. Nominators of private school teachers were identified from a sample of private schools drawn separately.

ACT's experience in recruiting panelists over the years has shown that it is becoming increasingly more difficult to enlist volunteers for standard setting. In anticipation of these difficulties, not only were larger samples drawn, a random sample of economics teachers was also identified from the Market Data Retrieval (MDR) database for direct contact. A total of 9,756 individuals were contacted and asked to serve as nominators . See Table 17 for the distribution of nominators.

Table 17: Distribution of Potential Nominators Contacted

	Public School Districts	Private School Districts	Teachers	State Education Associations	Colleges/ Universities	Economic/ Business Associations	Total
Nominators for all three types	4,494 (46.1%)	798 (8.2%)	4,150 (42.5%)	50 (0.5%)	250 (2.6%)	14 (0.1%)	9,756

Persons holding a specific title or position, such as the following, were contacted and asked to serve as nominators or panelists:

- district superintendents
- principals or heads of private schools
- classroom teachers
- nonclassroom educators (e.g., principals, district curriculum coordinators)
- state curriculum or assessment directors
- deans of colleges and universities (two-year and four-year; public and private)
- heads or members of economic and business associations/organizations

Nominators could submit 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 single pool of nominees was acquired for all the standard setting meetings—Pilot Study, Special Studies, and the ALS. The Pilot Study sample was drawn from the nominees available at the time of sampling. ACT continued to accept nominations throughout the Pilot Study phase. Individuals that were contacted to participate in the Pilot Study that were unable to attend were returned to the nominee pool for possible selection for the Special Studies or ALS meetings. A total of 292 candidates were nominated to serve as potential panelists.

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

- 55% of the members representing 12th grade/high school economics 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
- appropriate percentage (based on census population) of the members representing each of the four NAEP regions

Thirty panelists were required for the ALS panel. Seventy-one 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 (Table 18). A list of the panelists who participated in the ALS is presented in Appendix M.

Table 18: Panelists Participating in the ALS

Type	Males	Females	Caucasian	African Am.	Am. Indian	Hispanic	Midwest	Northeast	South	West	**Total N (%)**
Teacher	7	11	16	1		1		3	12	3	18 (58)
Nonteacher	3	1	3	1			1		2	1	4 (13)
General Public	5	4	7	1	1		5		2	2	9 (29)
TOTAL N (%)	15 (48)	16 (52)	26 (84)	3 (10)	1 (3)	1 (3)	6 (19)	3 (10)	16 (52)	6 (19)	31 (103)

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 February 2, 2007. 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:

- National Assessment Governing Board Brochure
- NAEP 2006 Economics Framework
- NAEP 2006 Economics Achievement Level Descriptions
- Confidentiality Agreement
- Press Release Form

The second mailing was sent on February 22, 2007. The cover letter contained detailed instructions related to travel arrangements and accommodations. Enclosures were:

- Briefing Booklet
- Preliminary Agenda
- Hotel diagram and directions to the meeting

A Briefing Booklet was first used by ACT for the 1994 ALS process. It includes a description of the goals and objectives for the process and a brief description of each step in the process; and it defines key terms used in the standard setting meeting. For the economics ALS, this briefing booklet was modified from its original format in response to some panelists' comments on the Pilot Study evaluation forms, indicating that the book was dense and confusing. These revisions were designed to provide a broader and less technical overview of the standard setting procedure, to be used to orient the panelists to the process. A copy of the ALS Briefing Booklet is provided in Appendix N.

The ALS Meeting

The ALS meeting lasted four days, March 7-10, 2007 (Wednesday through Saturday). It was conducted at the Westin Hotel in St. Louis. Sessions generally started at 8:00 a.m. or 8:30 a.m. and lasted until 4:30 p.m. or 5:30 p.m., except the last day, which adjourned at noon. The agenda is shown in Appendix A.

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 groups were divided to be as equivalent as possible. 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 2006 economics assessment pool were used in the ALS meeting. There were a total of 186 items representing 225 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: (a) multiple choice, (b) dichotomously-scored (one-point constructed response), and (c) polytomously scored (more than 1-point constructed response). The numbers of items by type were 154, 3, and 29, respectively.

The item pool was divided into equivalent, but overlapping, pools for groups A and B. Equivalence was evaluated with regard to: (a) mean and variation of item difficulty, (b) representation of content areas, and (c) percent of items of each type.

The equivalence criteria were met by assigning blocks of items to the two item pools. Blocks are sets of approximately 18 items created for purposes of test form construction to require approximately 25 minutes of student response time. Each student test booklet contains two blocks of cognitive assessment items and some additional background information to complete. The 2006 assessment consists of ten blocks—1 through 10. Six blocks were assigned to each pool. The pools had two blocks in common. The common blocks, blocks 2 and 4, are scheduled to be released to the public after the assessment. (Block 1 is also scheduled for release, but the item pools could not be balanced by assigning blocks to pools with three common blocks.)

The item pool division used in the ALS is identical to that used for both Mapmark methods in the Pilot Study. For the Pilot Study, the assignment of items to pools via blocks was modified slightly to accommodate the domains. 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 within each domain. This reassignment did not change the equivalence of the pools in other respects, as can be seen in Table 19, and was not essential for the Mapmark with whole booklet feedback method, but was retained in the ALS. Table 19 summarizes the item pool for each group and overall with regard to the key characteristics listed above.

Table 19: Item Difficulty Statistics and Number of Items by Subscale and Type Within Group and Overall

	No.	Percent by			Percent by Item			Item Difficulty				
	of	Subscale[a]			Type[b]			(Scale values at RP[c] of 0.67)				
Group	Items	Mkt	Ntl	Int	MC	DI	Poly	Points	Mean[d]	SD	Min	Max
A	113	45	42	12	67	1	31	137	329	43	235	454
B	114	47	38	15	70	1	28	137	331	43	235	454
Pool	186	46	40	14	68	1	30	225	331	43	235	454

[a] Economics content areas: Mkt = Market Economy, Ntl = National Economy, Int = International Economy
[b] MC = Multiple choice; DI = Dichotomously scored constructed response; Poly = Polytomously scored constructed response
[c] RP = Response Probability (of getting the item correct or earning the score point or higher)
[d] Scale values are on a transformed NAEP scale

Facilitation, Observers, and Room Setup

The NAEP ALS Project Director, Christina Hamme Peterson, served as the primary facilitator for the meeting. Rae Jean Goodman has been involved in every phase of the development of the economics NAEP and she served as the content facilitator. Both facilitators had participated in the Field Trial and 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). For day one, the entrance to the room was the “back,” at which observers were seated (see Figure 19). Due to panelist difficulties in hearing comments by other panelists at the opposite side of the room, the room was rearranged for days 2-4, with the observers at the side of the room and the panelist tables in the middle (Figure 20).

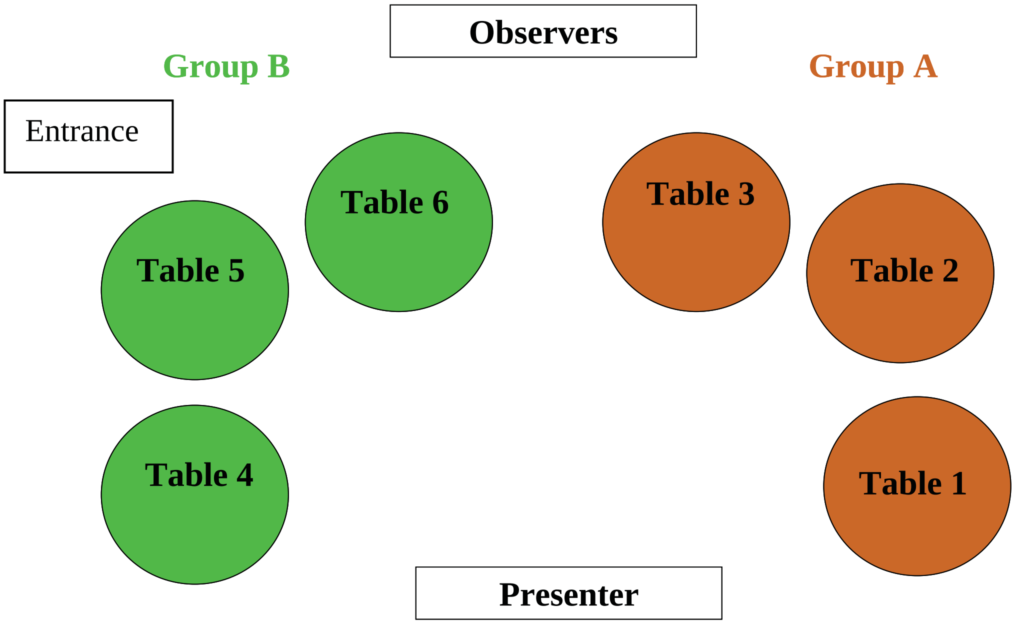


Figure 19. Room and table setup for Day 1.

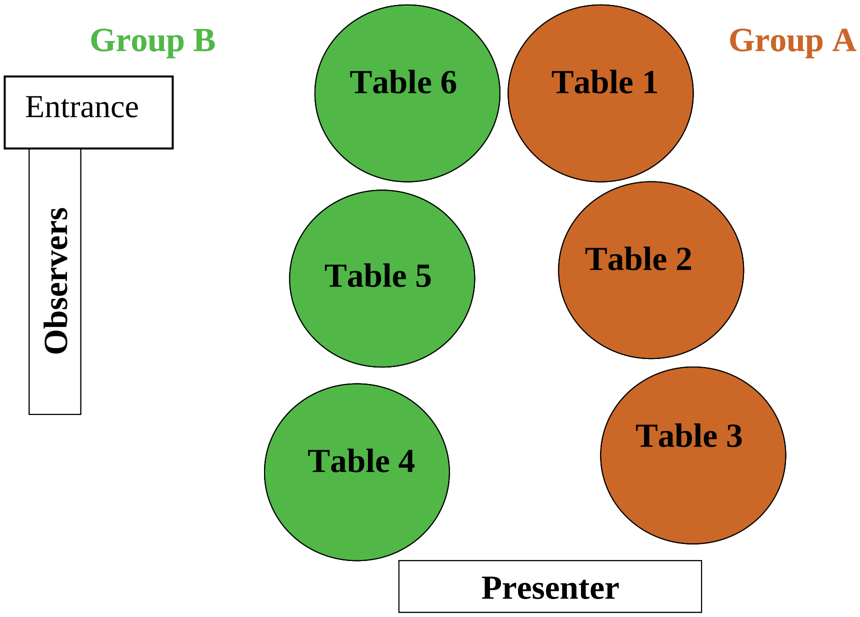


Figure 20. Room and table setup for Days 2-4.

A total of eight observers were present at various times: Susan Loomis, Assistant Director of Psychometrics at the National Assessment Governing Board; John Easton, member of the Governing Board Committee on Standards, Design, and Methodology (COSDAM); Nancy Petersen, Distinguished Research Scientist at ACT and a member of ACT's Technical Advisory Team (TAT); George Englehard, professor at Emory University and a member of ACT's Technical Advisory Committee on Standard Setting (TACSS); Nancy

Mead and Brent Sandene, two members of the NAEP report production team at Educational Testing Service; Andrew Kolstad, senior research associate at the National Center of Education Statistics; and Jade Caines, a graduate student at Emory University specializing in standard setting.

Orientation

General Orientation

In a brief welcome and introduction session, panelists were introduced to meeting coordinators, process and content facilitators, and to the observers. The role of observers was explained and panelists were asked to limit their interactions with observers to matters not directly related to the process.

Following the welcome and introduction, the Governing Board's Assistant Director of Psychometrics provided panelists with background information on NAEP and the Governing Board. This session covered the history, organizational structure, procedures, and key policies of the NAEP as well as the purpose of setting achievement levels. Information about the NAEP economics assessment was also presented.

Once oriented to the NAEP, panelists learned about the ALS process. This introduction explained how the ALS meeting fits into the overall NAEP ALS 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 a general description of the training and tasks involved in ALS.

Taking a Form of the NAEP

In the final session of the morning, panelists took a form of the NAEP. The test form administered was comprised of two blocks scheduled to be released: blocks 2 and 4. The panelists took the test under test-administration conditions similar to standard 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, but that they were to use the experience to gain some additional insight into what students experience taking the test. This was also an opportunity for panelists to become familiar with assessment items and scoring rubrics for items in two blocks included in the item pool of each group.

Orientation to the ALS Method and Materials

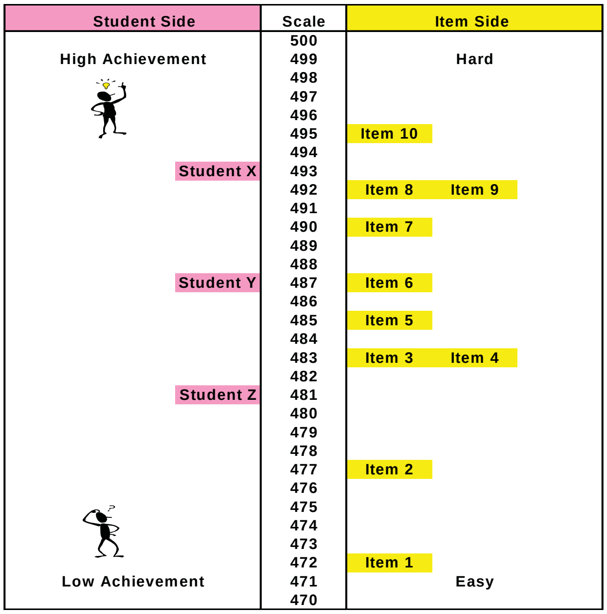
Method-specific aspects of the ALS meeting began after lunch on Day 1 with an orientation to the Mapmark method. The purpose of this orientation was to give panelists a general overview of the process, explain features of the economics assessment, and to introduce them to several of the key materials they would be using: the Ordered Item Book, Item Map, and student booklets.

A basic overview of Mapmark with whole booklet feedback was first provided. Panelists were told that Mapmark with whole booklet feedback is a three round method, with each round designed to provide information and feedback to set cut scores initially and then to evaluate and possibly change those cut scores in successive rounds. In the first round, panelists were told they would become very familiar with the items in the assessment and would identify the knowledge, skills, and abilities that they believed a student must have in order to answer each item correctly.

Panelists were then told that in the second round they would review scored assessments, actual examples of students' completed test forms illustrating different levels of student performance on the achievement scale. During this review, they would ask themselves how well each example illustrated the achievement levels as described in the Achievement Level Descriptions and consequently determine how well the cut scores reflected the ALDs.

In the final round they would be presented with consequences data. This would illustrate the percentage of students who took the economics NAEP in spring 2006 who scored at each achievement level. Panelists were told that they would use this information to gain a sense of the outcomes based on the cut scores and to allow them to determine if their cut scores should remain as they are, or move up or down according to the results.

In order to accomplish these tasks, panelists were instructed in how to use key materials. They were first introduced to the item map. Figure 21 shows a slide of a simplified item map. This slide was used to explain the general principle 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 *landmarks*, i.e., test items.



Student Side	Scale	Item Side	
	500		
High Achievement	499	Hard	
	498		
	497		
	496		
	495	Item 10	
	494		
Student X	493		
	492	Item 8	Item 9
	491		
	490	Item 7	
	489		
	488		
Student Y	487	Item 6	
	486		
	485	Item 5	
	484		
	483	Item 3	Item 4
	482		
Student Z	481		
	480		
	479		
	478		
	477	Item 2	
	476		
	475		
	474		
	473		
	472	Item 1	
Low Achievement	471	Easy	
	470		

Figure 21. A simplified item map as spatially representative of a journey from low to high achievement.

Figure 22 shows a slide used to explain the role of the response probability criterion (RP criterion) in determining the location of item landmarks on the map. This slide explained the location of Item 5 in Figure 21 as a function of the probability of a correct answer on that item at a given score point on the assessment. Items were mapped to the assessment scale values based on an RP criterion of 0.67. In other words, an item was mapped to the scale value at which a student has a 0.67 (or 67%) chance of answering the item correctly. This RP criterion was used to define *mastery* and panelists were instructed to consider a 2-in-3 chance as meaning mastery of the relevant content reflected in the item. Introducing this concept early is important in helping panelists to understand this criterion and to take it into account in their bookmark placements.

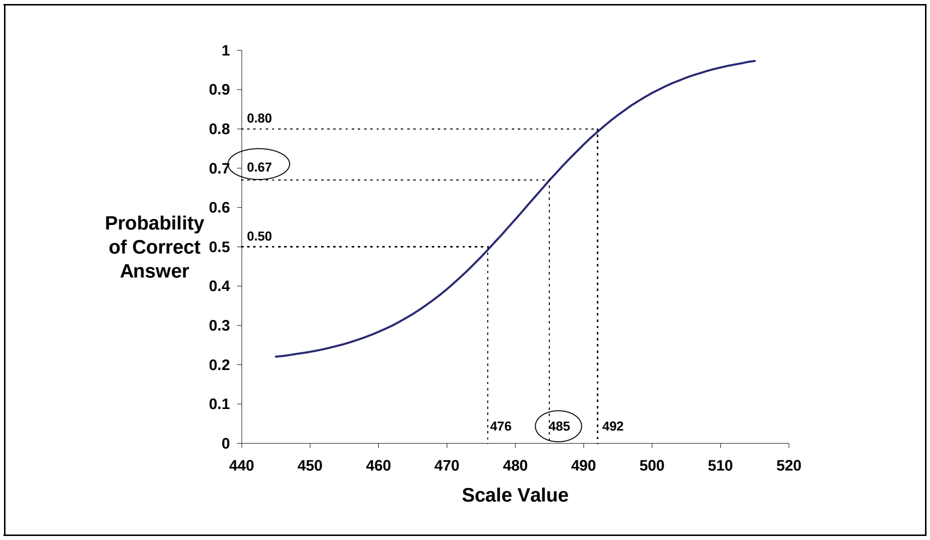


Figure 22. The relationship of the RP criterion to an item's scale value.

Panelists were then shown the Primary Item Map (Appendix B), on which columns correspond to the content areas of the assessment. The item map illustrated the distribution of all of the assessment items on the achievement scale, mapped from easiest to hardest. Panelists were shown how this map would allow them to compare differences in difficulty between items by identifying the distance between those items on the map. The meaning of colors and other information in the Primary Item Map will be explained in association with Figure 24.

Panelists were then oriented to the Ordered Item Book (OIB), which accompanied the Primary Item Map. The OIB contained all of the items with which the panelists would be working in order of their difficulty, beginning with the easiest. Figure 23 was used to illustrate this concept.

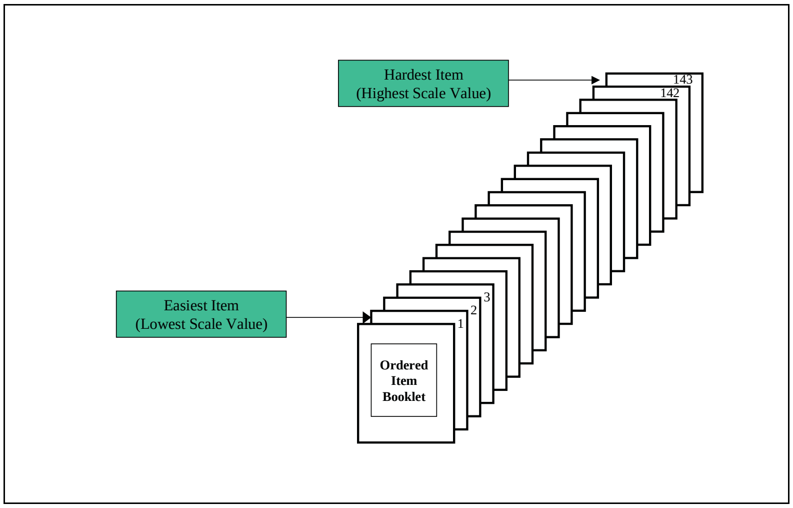


Figure 23. Illustration of how items are ordered by difficulty in the Ordered Item Book (OIB).

Panelists were told that they would be performing a bookmark task in round 1 using only their OIB and Item Map, but that they would use examples of student performance on forms of the economics assessment to inform their judgments in round 2.

Panelists were told about the three different types of items in NAEP (multiple choice, dichotomous constructed response, and polytomous constructed response) and were shown how these types of items were represented in their Item Map and OIB. Figure 24 shows a section of the actual Primary Item Map. Items were represented on item maps by a handle—a unique identifier—consisting of a character followed by a number. The character indicated item type (P = polytomously scored constructed response, 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 occur in multiple places in the item map, one place for each possible score level. The easiness rank of the item was based on the difficulty of the last, or highest, score level.

Scale	Content Area														
	Market					National						International			
above	P27_2	P29_3				P26_3	P28_4					P25_2			
497						P24_2									
494															
491						P23_2						P22_4			
488															
485															
482															
479															
476						P21_3	P28_3								
473															
470												M154			
467															
464															
461	P20_2					P19_2									
458						P18_2									
455						M153						P17_2			
452						M152									
449															
446						M151									
443	P16_2														
440												P25_1			
437	P14_2	P15_3													
434	P13_2					M150						M149			
431						M148	P26_2								
428						P12_2	P28_2	M146	M147						
425						P11_2									
422															
419	M144					M145						M143			
416	P10_2					M140	M141	M142							
413	P9_2					P21_2									
410	P29_2	P8_3				M135	P18_1	M136	M137	M138	M139				
407						M134									
404	D3	M132				M131	P23_1					M133			
401	M126	M130				P7_2	M127	M128	M129						
398	M121	M122	M123	P6_2	M125	M120	M124								
395	M115	M116	M119			P24_1	M118					P22_3	M117		
392	M107	M110	M114			M105	M108	M111	M113			M106	M109	M112	
389	M101	M102	M104			M100	M103								
386	M95	P27_1	M98			P19_1	P12_1	M92	M94	M96		P17_1	M93	M97	M99
383	P15_2	M91				P5_3	D2					M89	M90		

Figure 24. Primary Item Map on which score levels for polytomously scored item P12 (P12_1 and P12_2) are marked by circles.

Circles on the map in Figure 24 show the score locations of a two-point polytomously-scored item, P12. It can be seen that P12 is an item in the National Economy content area, that the scale value of the first score point, P12_1, is in the map score interval whose midpoint is 386, and that the scale value of the second score point, P12_2, is in the interval whose midpoint is 428. Score intervals of the scale value 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 P12 was in both item pools. Items in both pools are *common* items.

Figure 25 shows the location of the score points of item P12 in the group A and group B OIBs and indicates the information contained in the OIB for each score point. 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 P12 was located on page 78 and the second score point was located on page 120. There were 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 changed.

		Page in OIB	
Score	Handle	Group A	Group B
0			
1	P12_1	78	73
2	P12_2	120	119

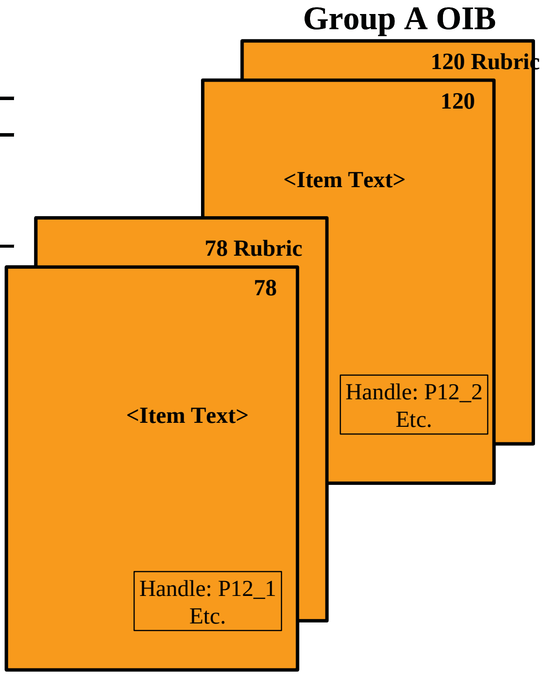


Figure 25. Information showing location and materials for Item P12 in OIB.

On the OIB page that contained the item’s text, there was a framed box, as shown in Figure 25. 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),
- content area classification in the 2006 Economics Framework,
- standard classification in the 2006 Economics Framework,
- answer key,
- identification code, and
- block and sequence number.

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 and which blocks were assigned to which group.

Round 1: 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 2006 Economics Framework prior to the meeting. To reinforce this learning, the framework presentation provided a clear, comprehensive account of the content and organization of the Economics Framework. The framework presentation lasted 45 minutes and was given by the 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 both the content and process facilitators, in which panelists were trained in the process of identifying KSAs required by constructed response items. They began with a dichotomously scored item 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 KSA review 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 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 and dichotomously-scored (constructed response) items in a group's item pool. Items were listed in order of difficulty by the last score point.

Figure 26 illustrates the contents of the CROIB. Unlike the OIB, all the information about a constructed response 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.

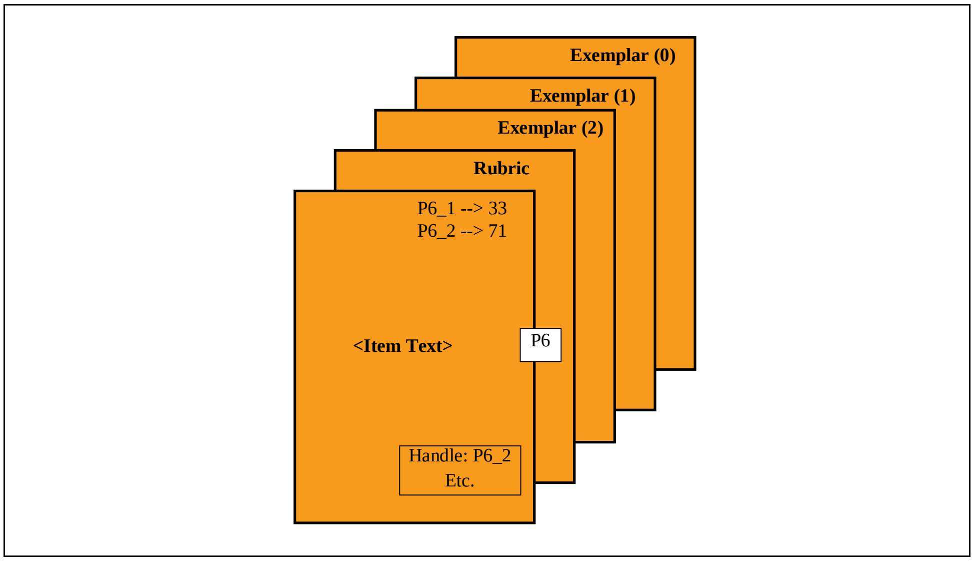


Figure 26. Slide illustrating contents of the Constructed Response OIB.

Panelists used large yellow post-its to record their notes on the KSAs. They were told that their notes were for their own use. They used one post-it for each score point. When panelists were finished with an item, they placed their notes in the Note template. This was a stapled set of 11x17 pages with outlines for accommodating six post-its per page. Within each post-it outline was an item handle and OIB page number identifying the post-it that was to be placed there. At the beginning of KSA Activity 3, post-its were moved from the Note template to the corresponding OIB page in the OIB. As noted earlier, the OIB contained all items, including the constructed response items. Figure 27 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. This allowed panelists to see their constructed response item notes in the context of all of the items in 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.

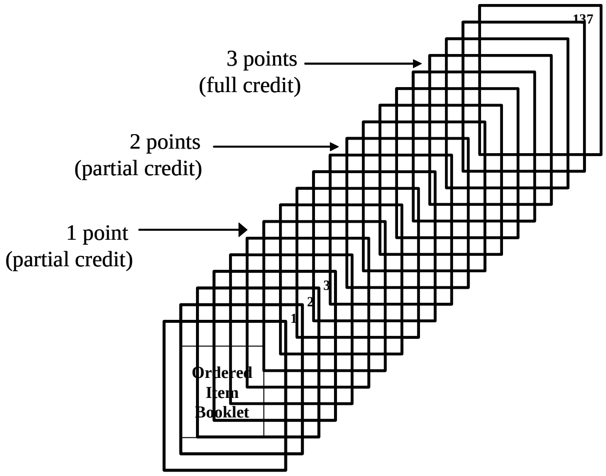
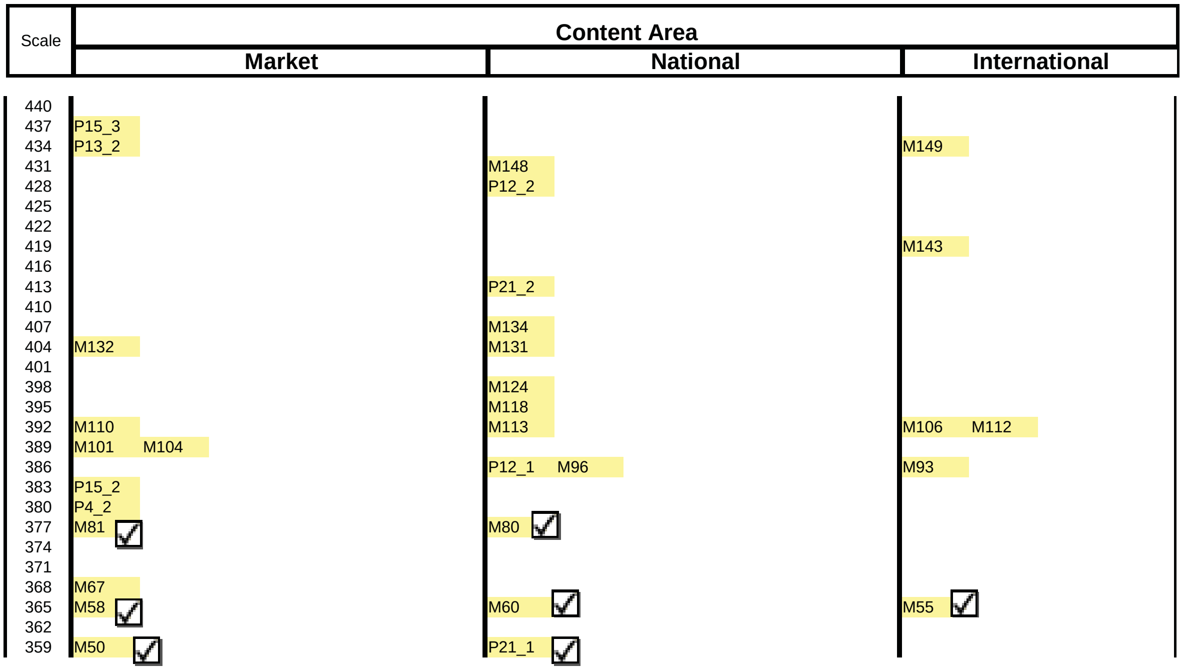


Figure 27. Score levels of a polytomously scored item are treated as separate items and appear at different places in the OIB.

Panelists checked items off on their Primary Item Map as they progressed through the OIB. Figure 28 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 content areas.



Scale	Content Area		
	Market	National	International
440			
437	P15_3		
434	P13_2		M149
431		M148	
428		P12_2	
425			
422			
419			M143
416			
413		P21_2	
410			
407		M134	
404	M132	M131	
401			
398		M124	
395		M118	
392	M110	M113	M106 M112
389	M101 M104		
386		P12_1 M96	M93
383	P15_2		
380	P4_2		
377	M81 ☑	M80 ☑	
374			
371			
368	M67		
365	M58 ☑	M60 ☑	M55 ☑
362			
359	M50 ☑	P21_1 ☑	

Figure 28. Simplified item map illustrating results of item check-off procedure as a panelist progresses through OIB 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 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 Governing Board policy definitions. Panelists were asked to share with the group a description, in their own words, of the KSAs that appeared to be required by each achievement level, and of additional KSAs that appeared to be required by a higher achievement level compared to a lower achievement level (e.g., Proficient vs. Basic).

Placing the Bookmarks

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

- The ALD should be thought of as representing a *range* of performance on the achievement scale.
- The panelist's job is to decide what the lower *borderline* of that range should be.

Panelists were told to think of the lower borderline in terms of a student who was *just qualified* to be in the achievement level and to decide for themselves what *just qualified* 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. Based on findings in the 2005 math Mapmark process (ACT Inc., 2005), 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 29 illustrates a bookmark placement.

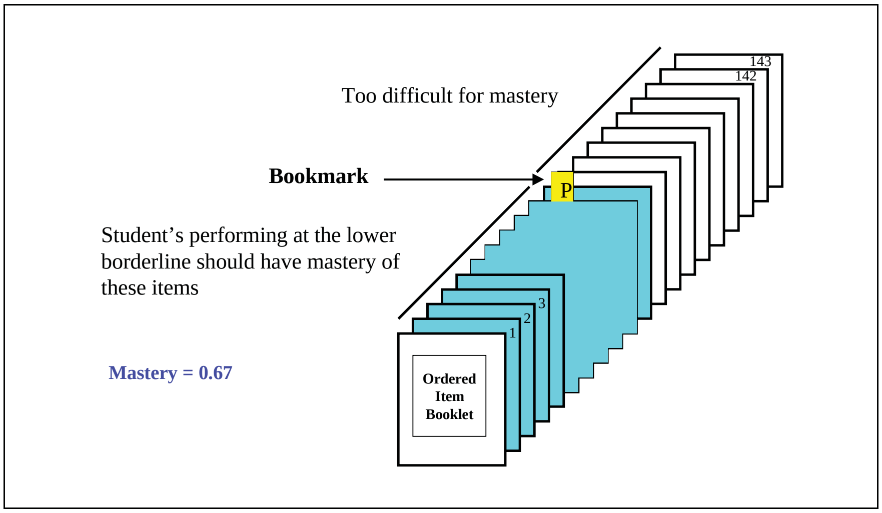
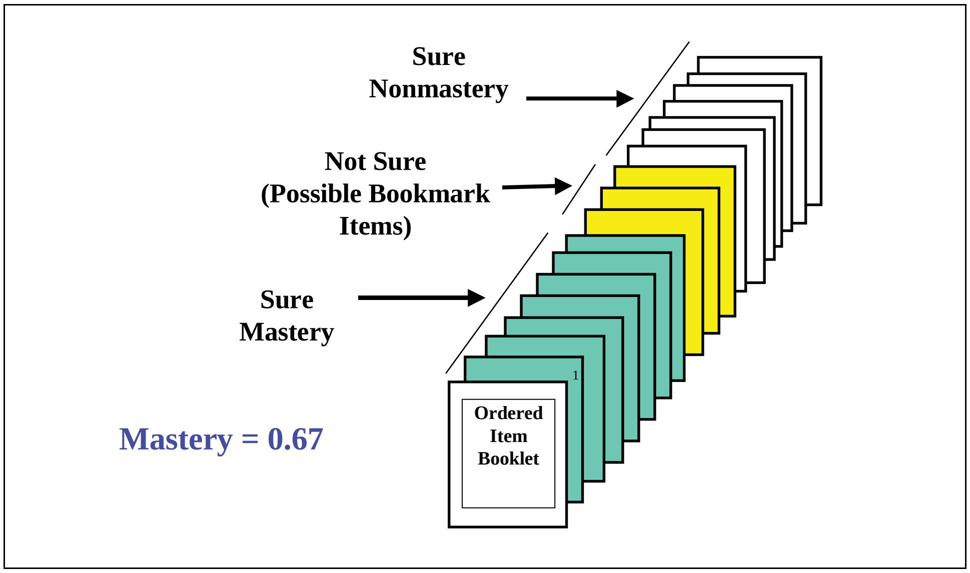

143
142
Too difficult for mastery
Bookmark
P
Student's performing at the lower borderline should have mastery of these items
3
2
1
Mastery = 0.67
Ordered Item Booklet

Figure 29. Bookmark placement task simplified.

Once panelists had this basic idea, the facilitator explained to panelists that it was possible for them to be unsure of where to place their bookmarks because: (a) they may not have felt there was a noticeable or meaningful difference between adjacent items in terms of difficulty, and (b) 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 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 mastered. 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 nonmastery* items. Figure 30 shows a slide that was used to illustrate this concept for panelists.



Sure Nonmastery

Not Sure (Possible Bookmark Items)

Sure Mastery

Mastery = 0.67

Ordered Item Booklet

1

Figure 30. 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. Panelists were instructed to place their bookmarks independently, without discussion with their group. 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. Finally, 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. The scale value corresponding to the bookmarked page was written beneath the bookmarked page number on the panelist’s form. The group cut score was computed for each achievement level.

Round 2: Whole Booklet Feedback

<u>Feedback from Round 1</u>

Feedback from round 1 consisted of: (a) group cut scores, (b) cut score dispersion, and (c) rater-locations relative to the group cut scores. In addition to providing the numerical values of cut scores, feedback was shown on item maps and Ordered Item Books to focus panelists' attention on the intended, criterion-referenced meaning of the round 1 cut scores. Figure 31 shows the cut score distribution chart provided as feedback from round 1. This chart was used to illustrate the location of all panelists' round 1 cut scores for each achievement level, the overlap (if any) between cut scores for achievement levels, and the highest and lowest cut scores by level.

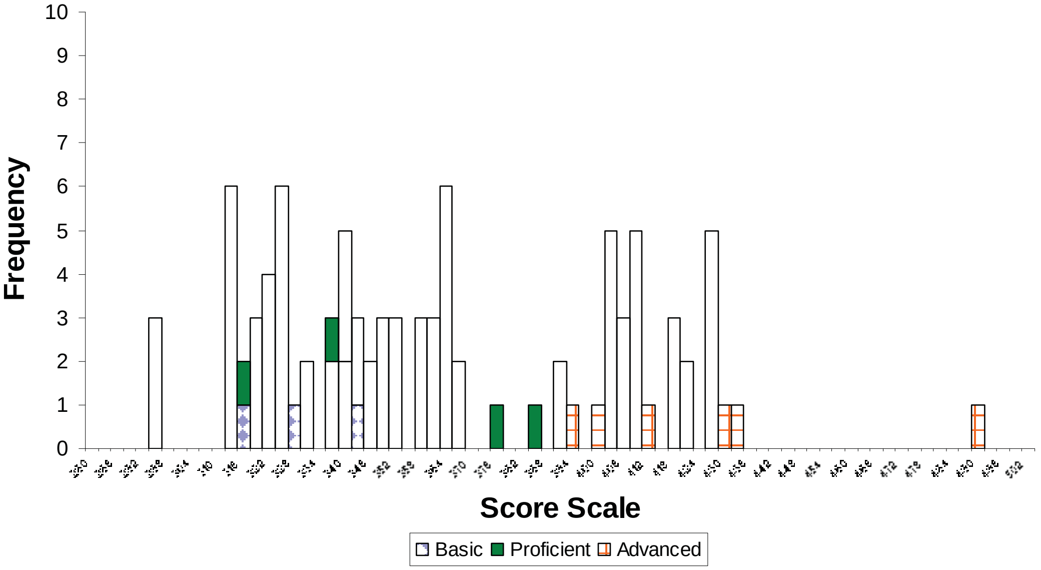


Figure 31. Cut score dispersion chart showing the distribution of cut scores by level.

Figure 32 shows how the group cut scores were then marked on the Primary Item Map. Panelists were instructed to draw the group cut score lines on their maps in the interval containing the cut score. Because they had circled their round 1 bookmarked items, they could compare the group cut score to their own.

Scale	Content Area															
	Market						National						International			
419	M144						M145						M143			
416	P10_2						M140	M141	M142							
413	P9_2						P21_2									
410	P29_2	P8_3					M135	P18_1	M136	M137	M138	M139				
407							M134									
404	D3	M132					M131	P23_1					M133			
401	M126	M130					P7_2	M127	M128	M129						
398	M121	M122	M123	P6_2	M125		M120	M124								
395	M115	M116	M119				P24_1	M118					P22_3	M117		
392	M107	M110	M114				M105	M108	M111	M113			M106	M109	M112	
389	M101	M102	M104				M100	M103								
386	M95	P27_1	M98				P19_1	P12_1	M92	M94	M96		P17_1	M93	M97	M99
383	P15_2	M91					P5_3	D2					M89	M90		
380	M82	M86	M87	P4_2			M83	M88	P28_1				M84	M85		
377	M79	M81					M78	M80								
374	M74	M76	M77				M75									
371	M68	M69	M70				M71	M72	M73							
368	M64	M65	M66	M67	P3_2		M63	P26_1	P11_1							
365	M56	M58	M61				M57	P2_2	P7_1	M59	M60	M62	M55	P22_2		
362							M54									
359	M50	M52	P14_1				P21_1						M51	M53		
356	M44	M45	M46	M49			M47						M48			
353	M41	M42	M43	P9_1			P5_2									
350	M38	M40					P5_1	M39					M37			
347	P8_2	M35	M36				M34						M32	M33		
344	M27	M28	M29	M30			M26						M31			
341	M21	P6_1	P16_1	P10_1	M24	M25	M22	M23					P22_1			
338	M18	P20_1	P1_2	M19			M20									
335							M17									
332	M16						M15									
329	M13												M14			
326	P15_1	M9	D1	M11	M12		M10									
323	M8	P4_1														
320	M7	P3_1														
317	M6															

Figure 32. Primary Item Map showing round 1 group cut scores (horizontal lines) and the location of Panelist X's bookmarked items (circled).

Panelists were also instructed to bookmark the group cut scores in the OIBs. For each achievement level, they were instructed to identify the items that fell between their cut scores and the group's and to determine what these items represented in terms of differences in performance between the two definitions of borderline, as shown in Figure 33. They were instructed to keep in mind where their cut scores fell in relation to the group's, because examples of student performance would be provided at the group cut score and not the individual panelist's cut score.

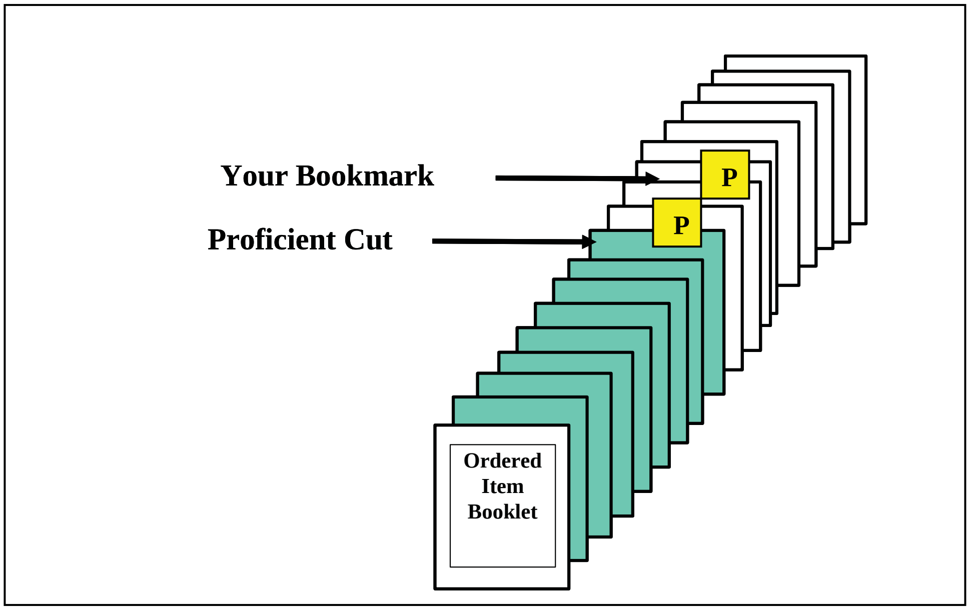


Figure 33. Slide demonstrating the comparison of the group cut score and Panelist X's bookmarked item in the Ordered Item Book.

Whole Booklet Feedback

Panelists were told that their round 2 cut score recommendations would be based on judgments of whether performance exhibited in student booklets scoring at the borderline of each of the achievement levels was too low, OK, or too high for the borderline of that level. Ten booklets on each of three forms would be provided, with each group (A and B) reviewing two forms for a total of 20 booklets per group. The booklets were distributed across the achievement scale, with one booklet on each form scoring at the middle of each achievement level, and two at the cut score. For each form, the expected number of points for each achievement scale value was plotted on the Booklet Score Plot, and the booklets were indicated on the plot at their scale value (Figure 34). These were used to provide a visual illustration of the location of each booklet relative to the cut scores and the achievement scale.

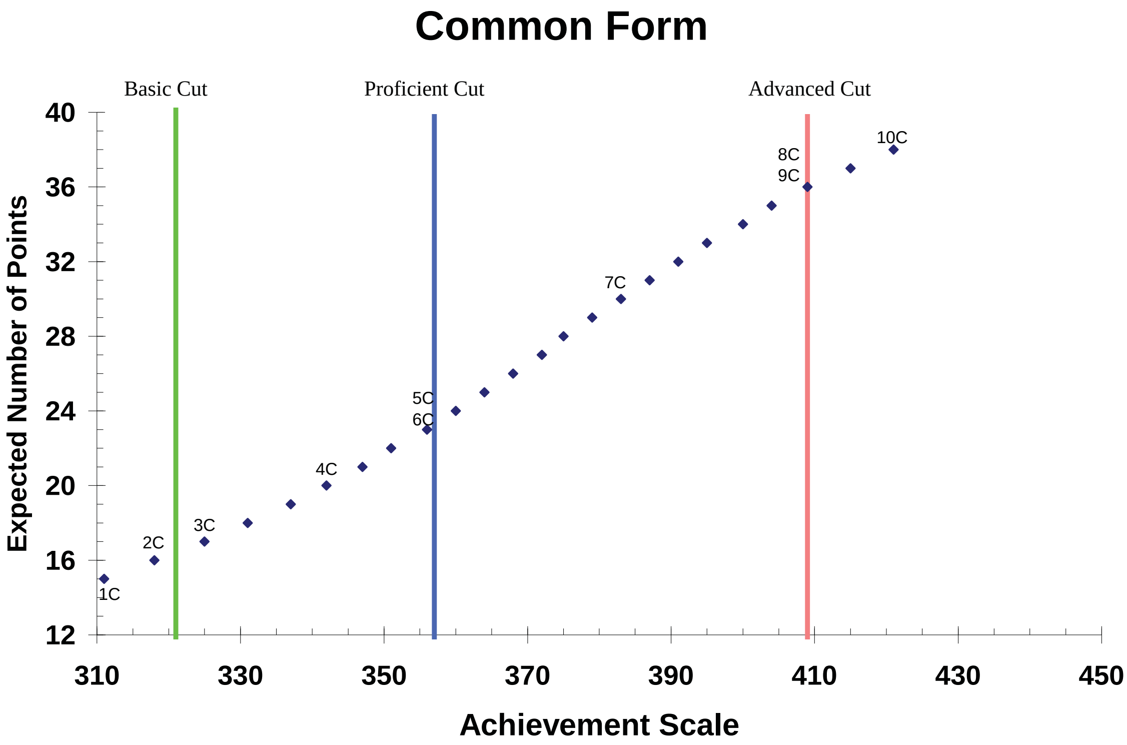


Figure 34. Booklet Score Plot for the common form, showing the round 1 cut scores and the score of each booklet (1C through 10C) on the achievement scale.

Before panelists began their independent review of the student booklets, they were led through a whole group exercise to familiarize them with the Booklet Score Charts (BSC), Item Score Table (IST), and booklet item maps, and to help them begin to understand the relationship between general performance on a form of the test and expected performance on individual test items.

The Booklet Score Charts were specific to each group and were provided for each achievement level. These charts mapped the expected number of points on the common and group-specific forms to the achievement scale within a range from 10 points below the *low* cut score for the achievement level to 10 points above the *high* cut score from the previous round. The booklets were then indicated at the location of their expected number of points. Panelists were asked to circle their cut scores on the Booklet Score Chart and to take note of where their cut scores fell in relation to the booklets they would be reviewing (see example in Figure 35).

		Common Form		Group B Only Form	
	Scale	Booklet	Expected No. of Points	Booklet	Expected No. of Points
	397		33.5		33.0
	396		.		.
	395		33.0		.
	394		.		32.5
	393		32.5		.
	392		.		32.0
	391		32.0		.
	390		.		.
	389		31.5		31.5
	388		.		.
	387		31.0		31.0
High	386		.		.
	385		30.5		30.5
	384		.		.
	383	7C	30.0		.
	382		.	7B	30.0
	381		29.5		.
	380		.		29.5
	379		29.0		.
	378		.		29.0
	377		28.5		.
	376		.		.
	375		28.0		28.5
	374		.		.
	373		27.5		28.0
	372		27.0		.
	371		.		27.5
	370		26.5		.
	369		.		27.0
	368		26.0		.
	367		.		.
	366		25.5		26.5
	365		.		.
	364		25.0		26.0
	363		.		.
	362		24.5		25.5
	361		.		.
	360		24.0	6B	25.0
	359		.		.
	358		23.5		24.5
Prof Cut--->	357		.		.
	356	5C, 6C	23.0		.
	355		.	5B	24.0
	354		22.5		.
	353		.		23.5
	352		.		.
	351		22.0		23.0
	350		.		.
	349		21.5		22.5
	348		.		.
	347		21.0		22.0
	346		.		.
	345		.		.
	344		20.5		21.5
	343		.		.
	342	4C	20.0		21.0
	341		.		.
	340		.		20.5
	339		19.5		.
	338		.		.
	337		19.0	4B	20.0
	336		.		.
	335		.		19.5
	334		18.5		.
	333		.		19.0
	332		.		.
	331		18.0		.
	330		.		18.5
	329		.		.
	328		17.5		18.0
	327		.		.
	326		.		.
	325	3C	17.0		17.5
	324		.		.
	323		.		.
	322		16.5	2B, 3B	17.0
	321		.		.
	320		.		.
	319		.		16.5
	318	2C	16.0		.
	317		.		16.0
	316		.		.
Low	315		15.5		.
	314		.		15.5
	313		.		.
	312		.		.
	311	1C	15.0		.
	310		.	1B	15.0
	309		.		.
	308		.		.
	307		14.5		14.5
	306		.		.
	305		.		.
	304		.		14.0

***Figure 35. Proficient Booklet Score Chart for group B showing the median, high, and low Proficient cut scores and the location of Panelist X’s round 1 cut*.**

For each test form, the Item Score Tables provided the score a student received (0 = incorrect, 1 = correct) for every score point on each student booklet. The items and score points were ordered from easiest to hardest, bottom to top, and the student booklets were ordered from lowest to highest scoring left to right. Figure 36 illustrates the Item Score Table for Form C, the common form. Panelists could use the IST to see, at a glance, the response patterns of students across the range of the achievement scale. For example, in Figure 36 panelists could see that in one of the borderline Proficient booklets, booklet 5C, the student received credit for about 50% of the total points and correctly answered many of the easy items and fewer of the hard items.

In the whole group exercise, the panelists reviewed the Booklet Score Charts and Item Score Tables in relation to the two student booklets at the Proficient cut score on the common form (booklets 5C and 6C in Figure 34). Using the Item Score Table, panelists were told to observe the response patterns of the two student booklets near the Proficient cut score (5C and 6C) and to note that:

- The students answered different items correctly and incorrectly, but the overall proportion of items answered correctly was the same.

- Differences in correct and incorrect answers may be due to variance in student mastery across content areas or standards.

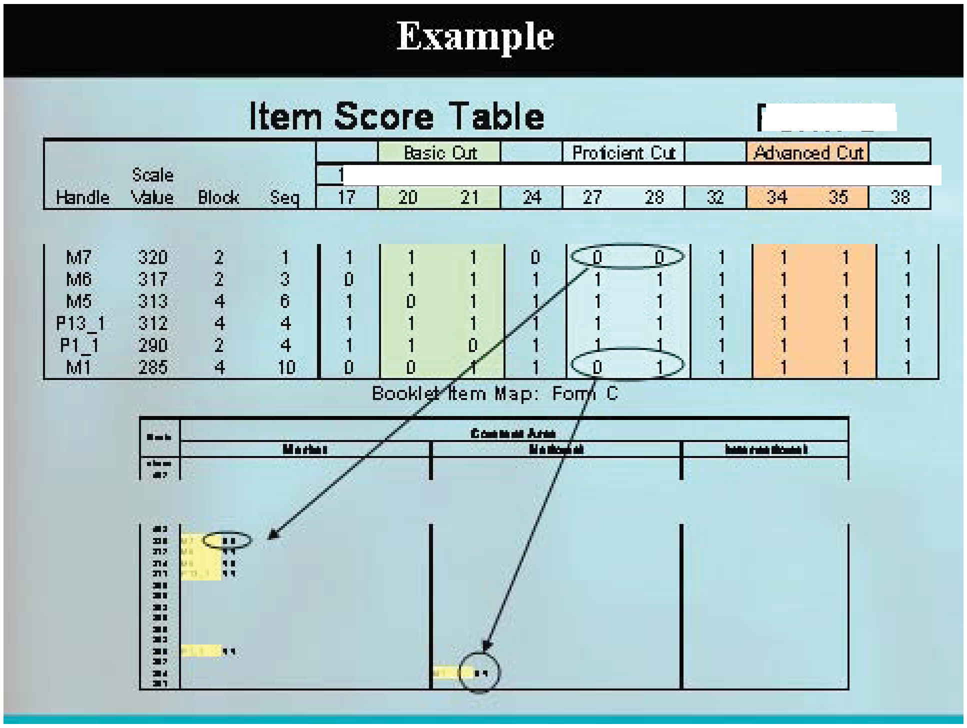
- Students did not get all items below the Proficient cut score correct and all above incorrect, but the probability of a correct response ***increased*** the farther ***below*** the cut score an item was and ***decreased*** the farther ***above*** the cut score an item was.

Item Score Table Form C

					Basic Cut			Proficient Cut			Advanced Cut		
	Scale			1C	2C	3C	4C	5C	6C	7C	8C	9C	10C
Handle	Value	Section	Seq	15	16	17	20	23	23	30	36	36	38
P21_3		1	9	0	0	0	0	0	0	1	0	0	0
P15_3		2	9	0	0	1	1	0	0	1	1	0	1
P13_2		2	4	0	0	0	0	0	0	0	1	0	0
M149		2	18	0	0	1	1	0	0	0	1	1	1
M148		1	17	1	1	1	0	0	1	1	1	0	1
P12_2		2	13	0	0	0	0	0	0	0	1	1	0
M143		2	3	0	0	1	0	0	0	0	1	0	0
P21_2		1	9	0	0	0	1	0	0	1	1	1	1
M134		2	12	0	1	0	0	0	0	0	0	1	1
M131		1	16	0	0	0	0	1	0	0	0	1	1
M132		2	17	0	1	0	0	1	1	0	0	0	0
M124		2	16	1	0	0	0	1	0	0	1	1	1
M118		2	7	0	0	1	1	0	0	0	1	1	1
M112		1	2	0	1	0	0	1	1	0	1	0	0
M113		1	15	0	0	0	0	1	0	1	0	1	1
M110		2	15	1	1	0	0	1	1	0	1	1	1
M106		2	14	0	0	1	0	0	0	1	1	1	1
M104		2	2	0	0	0	0	0	1	1	0	1	1
M101		1	10	1	0	0	0	0	0	0	0	1	1
M96		1	14	0	1	0	0	0	0	1	0	1	1
M93		1	8	0	0	0	0	1	1	1	1	0	1
P12_1		2	13	1	0	0	0	0	0	0	1	1	1
P15_2		2	9	0	0	1	1	0	0	1	1	1	1
P4_2		1	13	1	1	1	0	0	1	1	1	1	1
M81		2	1	0	0	0	0	1	1	0	1	1	1
M80		2	5	0	1	1	0	1	1	1	1	1	1
M67		1	11	1	0	0	1	0	0	1	1	1	1
M58		1	6	1	0	0	0	1	1	1	1	1	1
M60		2	8	1	0	0	0	0	1	1	1	1	1
M55		1	7	1	0	0	1	1	1	1	1	1	1
P21_1		1	9	0	0	0	1	0	0	1	1	1	1
M50		1	5	1	0	0	1	0	0	1	1	1	1
M42		1	12	0	0	0	1	1	0	1	1	1	1
M25		2	11	0	0	0	1	1	1	1	1	1	1
M20		1	18	0	0	1	0	1	1	1	1	1	1
P1_2		1	4	1	0	1	1	1	1	1	1	1	1
P15_1		2	9	0	1	1	1	1	1	1	1	1	1
P4_1		1	13	1	1	1	1	1	1	1	1	1	1
M7		1	1	1	1	1	1	1	1	1	1	1	1
M6		1	3	0	1	0	1	1	1	1	1	1	1
M5		2	6	0	1	1	1	1	1	1	1	1	1
P13_1		2	4	0	1	1	1	1	1	1	1	1	1
P1_1		1	4	1	1	1	1	1	1	1	1	1	1
M1		2	10	0	1	0	1	1	1	1	1	1	1

Figure 36. Item Score Table for Form C, with the left most column listing the items from hardest at the top to easiest at the bottom and the booklets provided from lowest to highest scoring (1C to 10C).

Panelists were instructed to transfer the item scores for booklets 5C and 6C from the Item Score Tables onto the common form item map, as shown in Figure 37. The common form item map was an item map that included only the items on the common form. This transferring task was designed to illustrate the differences in difficulty between items on the common form that students answered, or failed to answer, correctly. Panelists were asked to consider: How much more difficult is one item that the students both got wrong from another item that both students got right? How do these relate to the group Proficient cut score? To the panelist's Proficient cut score?



Example

Item Score Table

					Basic Cut			Proficient Cut			Advanced Cut		
	Scale			1									
Handle	Value	Block	Seq	17	20	21	24	27	28	32	34	35	38
M7	320	2	1	1	1	1	0	0	0	1	1	1	1
M6	317	2	3	0	1	1	1	1	1	1	1	1	1
M5	313	4	6	1	0	1	1	1	1	1	1	1	1
P13_1	312	4	4	1	1	1	1	1	1	1	1	1	1
P1_1	290	2	4	1	1	0	1	1	1	1	1	1	1
M1	285	4	10	0	0	1	1	0	1	1	1	1	1

Booklet Item Map: Form C

Figure 37. Slide instructing the panelist how to transfer the item scores for booklets 5C and 6C from the Item Score Table onto the common form item map.

Once they were able to understand and interpret the information provided in the Item Score Table and common item map, panelists were given the opportunity to independently review booklets 5C and 6C. They were instructed to take note of where their cut scores fall in relation to the scores on these booklets, and to consider if performance represented by the booklets was too high, too low, or just right for the lower borderline of Proficient. A brief discussion was held following this review, in which panelists shared their perceptions of the level of performance exhibited in the booklets as related to the performance described in the Proficient achievement level description. The purpose of the discussion was to help panelists begin the process of gaining a shared understanding of the meaning of borderline performance for the Proficient achievement level.

Following this discussion, panelists began an independent booklet review of all 20 booklets provided to their group. They were told to review at least two booklets at the borderline of each achievement level and one booklet in the middle of each level and one Below Basic level. During this review, they were to consider:

- How performance at the group cut score differed from performance at the middle of an achievement level.
- How students at their round 1 cut score were performing in relation to students at the group cut score.
- If performance at the group cut score was higher, lower, or just right for the lower borderline of the achievement level, given the Achievement Level Description.

At the conclusion of the independent review, panelists discussed with each other the above questions and shared their reactions to the performance exhibited in the booklets. They were told that their task was to share their thoughts, but not to convince one another, and that the purpose of the discussion was to give each of them further information and insight to incorporate into their round 2 cut score recommendations.

Round 2 Cut Score Recommendations

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 Cut Score Recommendation Form. Panelists were instructed to circle the scale value they chose for their round 2 cut score recommendation on their Booklet Score Chart and to move their round 1 bookmark in their OIB to the last item in their OIB with the scale value less than or equal to their recommended cut score.

Specific instructions were provided to aid them in the selection of their round 2 cut scores. They were instructed to select a range of scale scores within which they were deliberating. This range might encompass, for example, the panelist's own cut score at the low end and a booklet that they felt represented borderline performance at the high end. Once they had identified the range, they were to locate the high and low points of this range in their Ordered Item Book and Booklet Score Charts and to consider: (a) the KSAs of items at-or-below potential cut scores in the OIB, and (b) the performance associated with potential cut scores in the booklets indicated on the Booklet Score Chart.

In considering booklets, panelists were also reminded of a number of technical considerations. They were told that there are 50 different forms of the economics assessment and each form has approximately 45 total points. Because the achievement scale represents a much larger range than 45 points, there are some achievement scale values for which there are not corresponding point values on the forms panelists are reviewing. These scale values may correspond to point values on different forms, however, and so panelists can, and should, consider interpolating between raw score points on any given form when adjusting cut scores.

Round 3: Consequences Data

Feedback from Round 2

Feedback from round 2 was presented using the same materials and formats that were used to present feedback after round 1. Feedback from round 2 consisted of: (1) group cut scores, (2) cut score dispersion, and (3) rater-locations relative to the group cut scores. Panelists were given a new Primary Item Map, Booklet Score Chart, and Booklet Score Plots. A table of the group cut scores from rounds 1 and 2 was presented to show panelists how the cut scores had changed over rounds and the current group cut scores.

Consequences Data and Discussion

The percent of students in each achievement level and the percentage at or above each achievement level were reported to panelists as consequences data, based on the distribution of student performances relative to the round 2 cut scores. The percentage of students Below Basic was also included. The consequences data were based on the round 2 group cut scores. Figure 38 shows the consequences data that were given to panelists in round 3. 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.

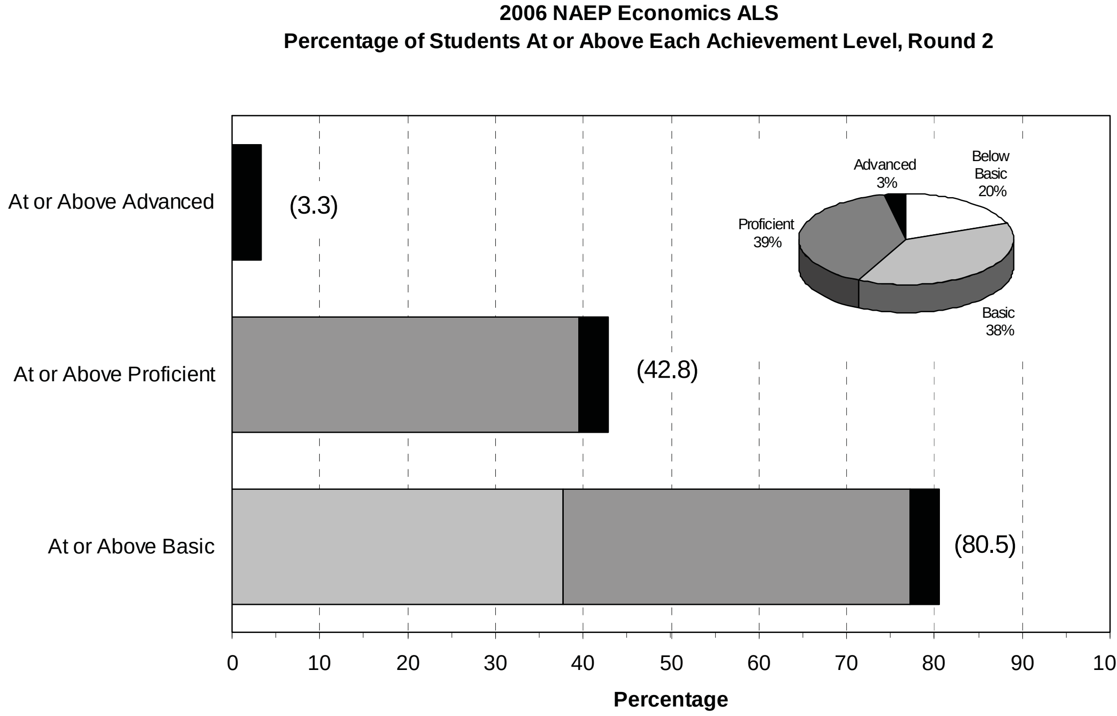


Figure 38. Consequences data presented to panelists in round 3.

The consequences data were discussed prior to panelists making their round 3 cut score recommendations. As a lead-in to the discussion, panelists were told that the data came from the 2006 administration of the grade 12 economics NAEP. The sample was nationally representative, and panelists were told to keep in mind that student performance was influenced by student motivation and by the amount of time available. But regardless of what students can do as illustrated by the consequences data, 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. These included: How do you feel about these cut scores now that you have seen the consequences data? Are you surprised by the percentages? Are these consequences about what you expected for a nationally representative sample of 12th grade students? How are your expectations influenced by your own experiences? What allowance, if any, should be made for motivation? For the timed conditions of the test?

Round 3 Cut Score Recommendations

The purpose of round 3 cut score recommendations was to allow panelists to adjust their cut score recommendations based on feedback after round 2, including the consequences data. Panelists were instructed to work independently, study the feedback from round 2, reflect on the discussion of the consequences data, and to determine if they felt their round 2 cut score recommendations needed to be changed. If they determined to change any of their recommendations, they were instructed to consult their Ordered Item Book and item map to determine if the new cut scores they were considering were consistent with performance described in the Achievement Level Descriptions. Panelists then recorded their cut score recommendations as they did in round 2.

Post-Round 3 Activities

Feedback from Round 3

Feedback from round 3 was given in the usual fashion except that panelists did not complete rater location tasks, identifying where their cut scores fell in relation to the final group cut. Panelists were given a new Primary Item Map with the final cut scores derived from round 3 and a new Cut Score Dispersion Chart. They were instructed to remove their bookmarks from their Ordered Item Book and to discard those bookmarks. They were then to move the group bookmarks to the final cut scores. This was to emphasize that the round 3 cut scores were the final cuts. The feedback also included consequences data based on the round 3 group cut scores. This was presented in the format shown in Figure 38.

Panelists were told that the round 3 group cut scores would be reported to the Governing Board as one of the key outcomes of the ALS meeting. It was very important that panelists understood the level of performance exhibited by students at the cut scores, which is the purpose of the feedback, and that they 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 the Governing Board with information about panelists' reactions to the final consequences data. A copy of the consequences questionnaire is included in Appendix L. 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. At the suggestion of ACT's Technical Advisory Committee on Standard Setting (TACSS) at the conclusion of the Pilot Study, the Cut Score Proportion Chart (Figure 39) was provided to allow panelists to see the relative impact of changing from one cut score to another if they wanted to raise or lower the cut score. This chart provided the percentage of students scoring at or above every fifth score value on the NAEP-like assessment scale. The final cut scores were marked on the chart. Panelists were instructed to use this information to help them decide what final cuts they would recommend to ensure that the consequences data accurately reflected the proportion of students at each of the three achievement levels.

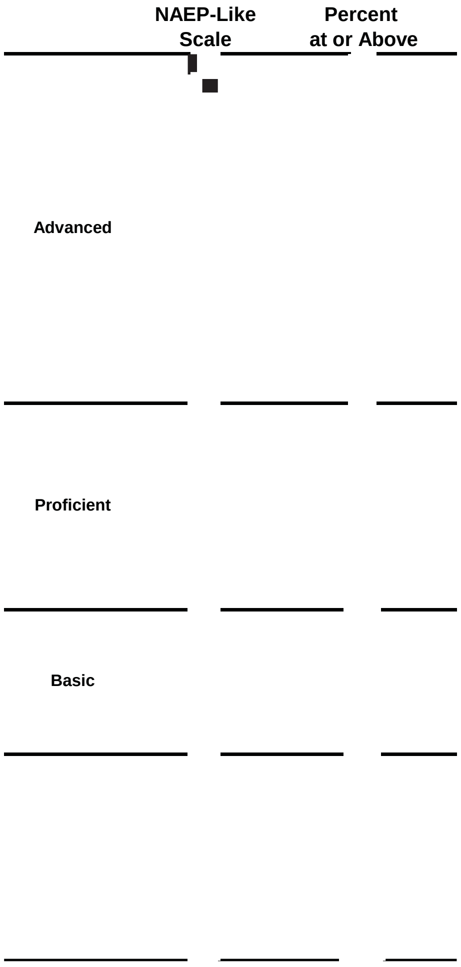


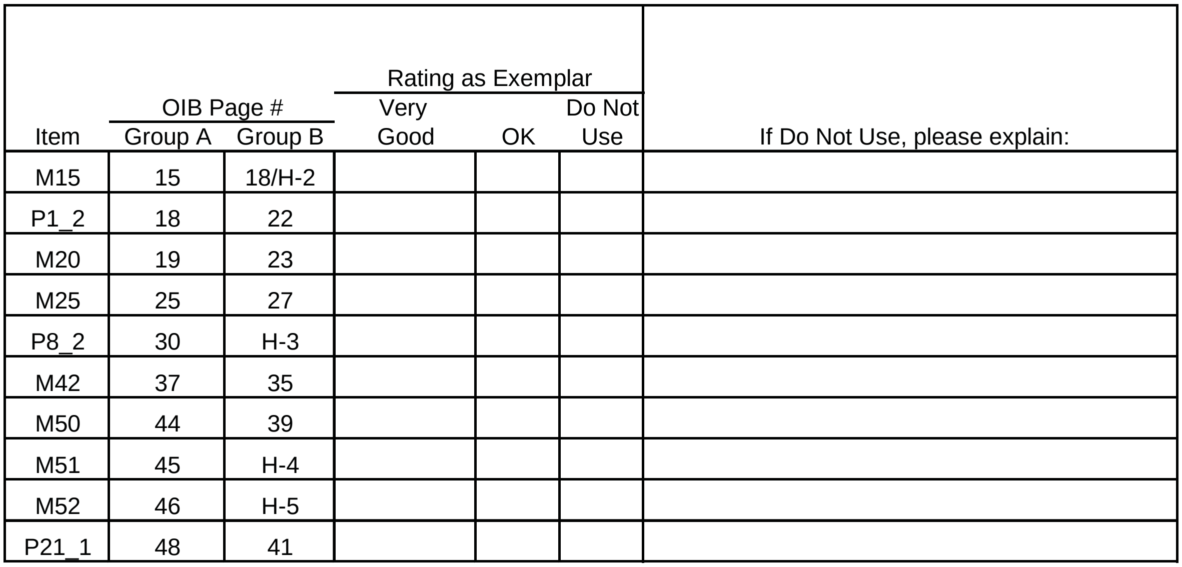
Figure 39. Cut Score Proportion Chart illustrating the percent of students scoring at or above every fifth score level.

Ratings of Exemplar Items

The purpose of the exemplar item rating task was to provide the Governing Board with information concerning the suitability of items for illustrating what students in the achievement levels know and can do. Potential exemplars (or examples) were drawn from blocks of the assessment that were selected for eventual release to the public. These were blocks 1, 2, and 4. 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 the Governing Board with this input.

An item was selected as a potential exemplar for an achievement level if it was mapped to that achievement level and not to a lower or higher level (see Figure 17). This criterion produced reasonable-sized pools of items for potential use as exemplars.

Figure 40 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, and identified the items by handle and the OIB page number where they could be found. Since Block 1 was not in the group B item pool, group B was given a special handout for these items and the page number of Block 1 items in this handout was indicated on the Exemplar Item Rating form.



			Rating as Exemplar			
	OIB Page #		Very		Do Not	
Item	Group A	Group B	Good	OK	Use	If Do Not Use, please explain:
M15	15	18/H-2				
P1_2	18	22				
M20	19	23				
M25	25	27				
P8_2	30	H-3				
M42	37	35				
M50	44	39				
M51	45	H-4				
M52	46	H-5				
P21_1	48	41				

Figure 40. 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 in part on what is called *procedural validity*. Procedural validity is provided in the form of evidence that the procedures were carried out as intended, and were understood by the panelists. At the end of each round and each day, panelists were provided with an evaluation form designed to assess their understanding of instructions, tasks, and materials. There were a total of five questionnaires administered 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. These evaluations were reviewed at the end of each day and any sources of confusion were identified for clarification with individual panelists or the group as a whole. The process evaluation questionnaires are presented in their entirety in Appendix O. Along with the questions, the appendix shows the frequency of responses per Likert-scale category, and the average response.

In order to allow for comparison of procedural data from the ALS with methods used in previous NAEP standard setting meetings, an effort was made to ensure that the evaluation questions were largely the same as questions used to evaluate NAEP ALS methods in the past. Strong support for procedural validity would be demonstrated by consistent mean (average) responses on most items at or above 4.0 on a 1-5 scale. In general, panelist evaluations of the ALS were comparable to or better than evaluations from the 1998 civics and the 2005 mathematics standard setting projects. Based on these results, it seems reasonable to conclude that, in panelists' perceptions, the quality of the ALS process for economics equals or exceeds the quality of other methods used to establish cut scores for the NAEP in other subjects.

Evaluation of the Method Outcomes

The Mapmark ALS process compared well with methods ACT used in past standard setting work for the Governing Board. Key evaluation questions on the last process evaluation questionnaire addressed panelists' overall perception of the effectiveness of the ALS method, whether the process afforded them the opportunity to use their best judgment, whether the process yielded reasonable and defensible cut scores that represented meaningful distinctions between achievement levels, and whether they were confident in their final cut scores. Responses were on a Likert scale of 1-5, with 5 the highest level of agreement.

Figure 41 shows the mean ratings of Mapmark with whole booklet feedback and previous grade 12 subject ALS methods on these key overall process evaluation questions. The ALS method used in the 2005 mathematics contract was Mapmark with domains and the method used in the 1998 civics contract was a modified-Angoff method. 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 with whole booklet feedback method compared favorably with the averages for the other two methods.

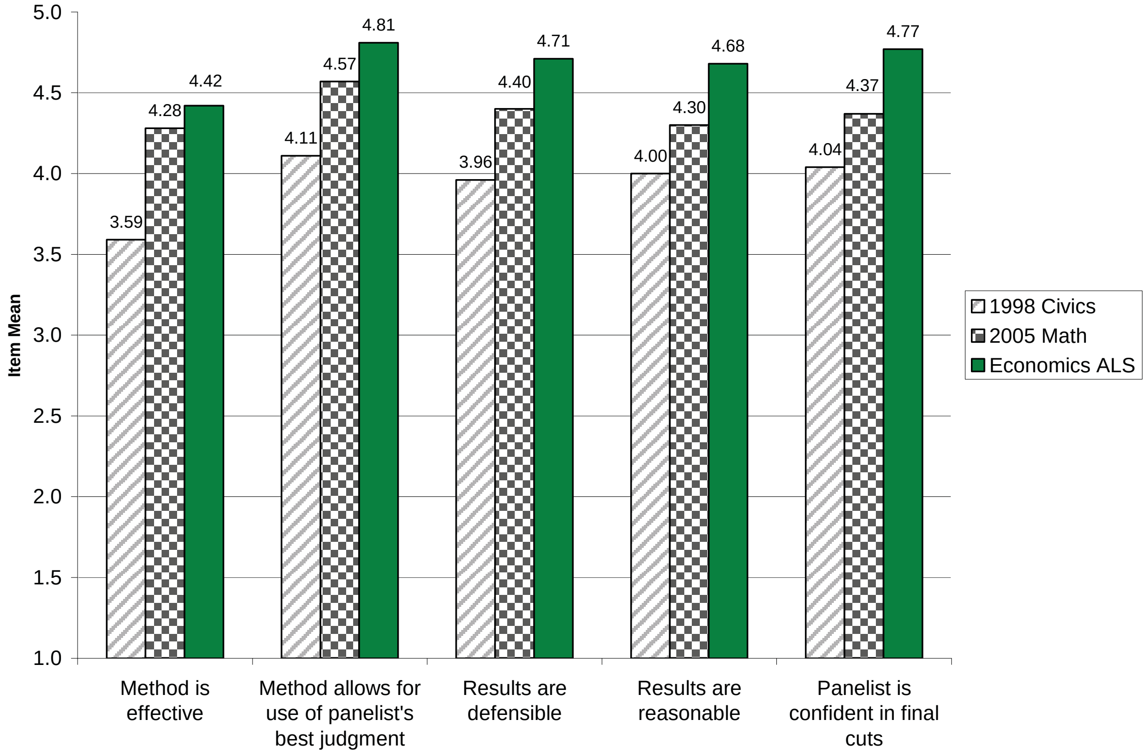


Figure 41. Mean ratings of economics ALS and previous ALS methods on key process outcome questions.

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 31 panelists who completed the last process evaluation questionnaire, 23 responded *definitely*, 6 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 the Governing Board.

Clarity of Instructions and Presentations

Table 20 shows average ratings on questions pertaining to clarity of instructions. The ratings are all above 4.0 and are consistently higher than ratings for a similar method used in the 2005 mathematics ALS. Only the instructions for placing the bookmark are lower, but remain above 4.0.

Table 21 shows average ratings pertaining to clarity of presentations on certain topics addressed the first day of the ALS meeting. The presentations were consistently rated as clear and were comparable to the results from the 2005 mathematics ALS.

Table 20: 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 2005 Math	 Economics ALS
1	1-23	Whole group KSA review	3.84	4.17
1	2-5	Independent OIB review	4.35	4.48
1	2-11	Table discussion of the OIB	4.17	4.32
1	2-29	Placing the bookmarks	4.45	4.10
2	3-6	Borderline Proficient exercise	-	4.29
2	3-27	Recommending round 2 cut scores	4.16	4.61
3	4-8	Using the consequences data	4.47	4.68
3	4-15	Recommending final cut scores	4.53	4.77
Post	5-5	Completing the consequences questionnaire	4.52	4.74
Post	5-17	Exemplar rating task	4.04	4.68

Table 21: Clarity of Topic Presentation

The explanation/overview/presentation of the ____ was

(5 = *absolutely clear*; 3 = *somewhat clear*; 1 = *not at all clear*)

Round	Question Location	Activity	Average Rating 2005 Math	 Economics ALS
Pre	1-4	NAEP in general	4.26	4.53
Pre	1-5	Development of the Economics NAEP	4.37	4.50
Pre	1-6	Major organizations involved and the roles of each	4.23	4.50
Pre	1-15	Method to be followed in this meeting	3.79	4.13
Pre	1-16	How an item map is constructed	3.77	4.23
Pre	1-18	Information in the Ordered Item Book	4.07	4.42
Pre	1-20	Economics Framework	4.00	4.29

At the conclusion of the process, panelists were also asked to rate instructions and their understanding of tasks for the entire process. They were asked to indicate the degree to which they felt the instructions on what they were to do during each round were clear (1 = *not at all clear* to 5 = *absolutely clear*) and the adequacy of their understanding of the tasks they were to complete (1 = *totally inadequate* to 5 = *totally adequate*). Figure 42 shows the mean ratings of Mapmark with whole booklet feedback and previous ALS methods on process evaluation questions on the clarity of instructions and panelist understanding of the tasks.

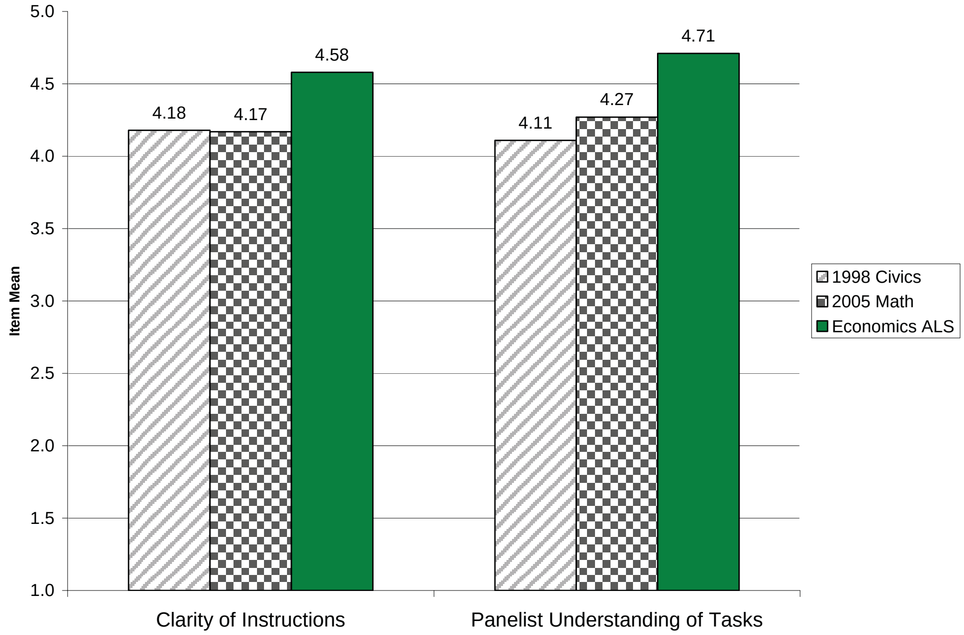


***Figure 42.* Mean ratings of economics ALS and previous ALS methods on clarity of instructions and panelist understanding of task.**

Understanding of Concepts and Feedback

Understanding of concepts and feedback depends on the clarity of presentations and instructions, which the previous section shows was good. It can be seen in Table 22 that panelists had a good understanding of concepts in the ALS process. In particular, understanding of concepts unique to the Mapmark process, such as the concept of how to use item maps and of the information in Booklet Score Charts and plots and Item Score Tables was high, as indicated by average ratings above 4.0 in Table 22.

Table 22: Understanding of Concepts
I understand/understood …
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Round	Question Location	Activity	Average Rating 2005 Math	 Economics ALS
Pre	1-10	The difference between criterion and norm-referenced standards	4.63	4.19
1	2-3	The score levels of polytomous items	4.10	4.61
1	2-6	How to use my item mapwith the Ordered Item Book	4.42	4.81
1	2-30	How to use the ALDs to choose my bookmarks	4.17	4.45
2	3-19	The information in the Booklet Score Chart	-	4.52
2	3-20	The information in the booklet score plot	-	4.45
2	3-21	The information in the Item Score Tables	-	4.61
2	3-24	The difference between borderline and typical performance within an achievement level	4.52	4.65
Post	5-21	The purpose of this meeting	4.80	4.90

Panelists had good understanding of the feedback they were given. As shown in Table 23, average ratings of understanding of general types of feedback such as the group cut scores (round ___ median cut scores), rater location feedback, and consequences data were well above 4.0 after round 1 and remained high with each round.

Table 23: Understanding of Feedback
I understand/understood the round ___ …
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Feedback	Round 1	 2	 3
Median cut scores	4.77	4.81	4.77
What students at the round ___ cut scores can do	4.52	4.74	4.81
Rater location feedback (where panelist cut scores were in relation to median)	4.77	4.81	-
Cut score dispersion chart	4.65	4.84	-
Consequences data	-	4.90	4.80

Understanding the Achievement Level Descriptions and Borderline Performance

Panelist understanding of both the Achievement Level Descriptions and the concept of performance at the lower borderline at each achievement level was also assessed. These are two concepts critical to the process of identifying cut scores. As expected, their understanding of these two critical concepts increased across rounds.

At the conclusion of round 1, panelists were asked to rate their understanding of the Achievement Level Descriptions for each level (Basic, Proficient, and Advanced). Panelist responses were used to assess if clarification was needed during the meeting for any one level. After the meeting, the mean rating was calculated across all three levels for round 1. For rounds 2 and 3, panelists were asked to rate their understanding of the Achievement Level Descriptions for all levels combined. Table 24 shows the mean ratings by round for all levels combined of panelist understanding of the Achievement Level Descriptions.

Table 24: Understanding of Achievement Level Descriptions

My understanding of the Achievement Level Descriptions [in round ___] was ...

(5 = *totally adequate* to 1 = *totally inadequate*)

	Round		
	1	2	3
Average Rating	4.63	4.71	4.81

Table 25 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 25: 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		
	1	2	3
Average Rating	4.29	4.55	4.87

At the conclusion of each round, panelists were also asked to respond to statements about performance at the lower borderline. The lower borderline question is worded slightly differently in the first round than it is in the second and third rounds. At the conclusion of the first round, panelists are asked, for each achievement level, to indicate their level of agreement (5 = *totally agree*, 1 = *totally disagree*) with: I was comfortable using the concept of performance at the lower borderline of _______. At the conclusion of the second and third rounds, the question was asked for all three levels combined. Panelists had to respond on a scale from 1 to 5 (1 = *not well formed*, 5 = *very well formed*) to the statement: At the time I provided my round ___ cut score recommendations, my concept of the lower borderline performance of an achievement level was. The mean panelist rating for these questions, by round, is provided in Table 26. The panelist ratings increase by round, as is consistent with patterns of response seen in previous standard setting meetings.

Table 26: Development of Borderline Concept
Panelist Mean Rating on the Evaluation Questions:
I was comfortable using the concept of performance at the lower borderline of ______
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

		Round		
Level	Question Location	1	2	3
Advanced	2-24	4.39	-	-
Proficient	2-25	4.35	-	-
Basic	2-26	4.32	-	-
All Combined		4.35	-	-

At the time I provided my round __ cut score recommendations, my concept of the lower borderline performance was:
(5 = *very well formed*; 3 = *moderately formed*; 1 = *not well formed*)

	Round		
Level	1	2	3
All Combined	-	4.55	4.81

Comfort and Confidence

As shown in Table 27, 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 addition, panelists' confidence in their cut score recommendations (Table 28) started high and increased steadily from round 1 to round 3. Panelist initial level of confidence is higher than usual, but the trend of increasing confidence over rounds is typical of other methods and standard setting meetings ACT has conducted for the Governing Board.

Table 27: Panelist Mean Rating of Comfort Level with Various Features of Mapmark
I think I will be/I was comfortable
(5 = *totally agree*; 3 = *somewhat agree*; 1 = *totally disagree*)

Round	Question Location	Activity	Average Rating
1	1-17	Using a 2/3 or 0.67 probability to interpret the location of an item on my map	4.29
1	2-7	Working through the Ordered Item Book on my own	4.84
1	2-33	Using a 0.67 probability to define mastery in placing my bookmarks	4.06
2	3-31	Choosing scale values instead of placing bookmarks to recommend cut scores	4.35

Table 28: Panelist Mean Rating of Confidence Level in Cut Scores by Round

	Round		
	1	2	3
Average Rating	3.97	4.32	4.77

Usefulness/Helpfulness of Materials and Information

Results in the top panel of Table 29 show that panelists found the whole group and table group KSA activities to be useful. During these activities panelists worked together to identify knowledge, skills, and abilities that a student needs in order to answer each item correctly. The bottom panel of Table 29 shows that the information and materials in the Mapmark process were generally perceived to be helpful. Average ratings for almost all materials and information specific to the Mapmark process were above 4.0 and all were higher than the average rating for the helpfulness of consequences data (the percent of students in achievement levels), at 3.90. This may be regarded as a positive outcome since the consequences data are purely normative information. Besides the consequences data, the least helpful materials were the booklet score plots, although still highly rated at 3.97. This relatively low rating is probably due to the fact that the booklet score plots are only used once to locate booklets. The Booklet Score Charts then provide the same information in more detail and are used repeatedly in round 2. As in previous Mapmark standard settings, the Ordered Item Book was perceived to be most helpful, and, in this case, the mean rating of helpfulness of the OIB was followed closely by that of the Primary Item Map and Achievement Level Descriptions.

Table 29: Usefulness/Helpfulness of Activities/Information

The ________ was

(5 = *very useful*; 3 = *somewhat useful*; 1 = *not at all useful*)

Round	Question Location	Activity	Average Rating
Pre	1-25	Whole group work on the common constructed response items	4.23
1	2-2	Table group review of remaining constructed response items	4.23
1	2-12	Table discussion of the Ordered Item Book was	4.42

During the ALS process, I found the ____________

(5 = *very helpful*; 3 = *somewhat helpful*; 1 = *not at all helpful*)

Round	Question Location	Information/Materials	Average Rating
Post	5-30	Achievement Level Descriptions	4.71
Post	5-31	Ordered Item Booklet	4.84
Post	5-32	Primary Item Map	4.71
Post	5-33	Rater Location Data	4.23
Post	5-34	Consequences Data	3.90
Post	5-35	The Booklet Score Charts	4.26
Post	5-36	The Booklet Score Plots	3.97
Post	5-37	The Cut Score Dispersion Chart	4.39

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 30, 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 4.65 (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 4.45 (5 = *to a great extent*).

Table 30: 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	Average Rating
1	2-32	1.45
2	3-30	1.35
3	4-18	1.65

Amount of Time Allocated for Tasks

The adequacy of time allocated for tasks was an important issue in this ALS because this was the first time that Mapmark with whole booklet feedback had been implemented as an ALS procedure. Details concerning the amount of time allocated for tasks is presented in Table 31. Average ratings in this table indicate that time was sufficient for all tasks as all averages were greater than 3.0. In a few cases, the average was closer to 4.0, in particular in the general orientation to NAEP (3.80), the general introduction to the NAEP Achievement Level Setting process (3.81), and the framework presentation (3.74). These averages are higher than usual for ACT NAEP standard settings. These unusually high ratings may have been due to changes to the briefing book provided in the advance materials. The new briefing book provides much of the information covered in the orientation sessions on the first day, and so some of the information in the presentations may have felt redundant to panelists.

Table 31: Amount of Time Allocated for Activities
(5 = *far too long*; 3 = *about right*; 1 = *far too short*)

Round	Question Location	Activity	Average Rating
Pre	1-3	The General Orientation to the NAEP Program	3.80
	1-8	The General Introduction to the NAEP Achievement Level Setting process	3.81
	1-14	The Mapmark method orientation	3.45
	1-19	The Framework presentation	3.74
	1-22	The whole group KSA review	3.43
1	2-1	The table group KSA review	3.26
	2-4	Independent OIB review	3.39
	2-10	The table discussion of the OIB	3.13
	2-16	The ALD presentation	3.55
	2-28	Placing the bookmarks	3.48
2	3-5	Borderline Proficient exercise	3.33
	3-12	The table group whole booklet review	3.19
	3-26	Round 2 cut score recommendations	3.16
3	4-9	Discussing the consequences data	3.42
Post	5-3	The Consequences Questionnaire	3.19
	5-10	Complete the tasks I was to accomplish during each round	3.39
	5-16	The Exemplar Item Rating Task	3.26

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—the most vocal panelists generally voiced surprise that the percentages at each level above Below Basic were not lower. Despite this surprise, the group cut score did not change from round 2 to round 3 other than a one point increase at the Basic level. This result, along with comments voiced during the whole group discussion, indicates that panelists were not unduly influenced by the introduction of the consequences data to their process, and maintained their commitment to criterion-referenced cut score judgments.

At the conclusion of the ALS, panelists were asked to review the achievement level percentages and to complete a questionnaire indicating the reasonableness of those percentages. They were asked to indicate if they felt that the percentages reflected their expectations about the proportions of students whose NAEP score would be at or above the cut score established for each achievement level and, if not, to indicate if they would raise or lower the cut scores to adjust the percentages. For each achievement level established in the ALS, the majority of panelists (84-90%) indicated that the cut scores yielded reasonable achievement level percentages and so should be left as is (Table 32). Of those who recommended changes, several recommended changes for more than one level. In total, only 8 panelists (26%) suggested changes: 7 suggested raising cuts and 1 suggested lowering cuts.

Table 32: Cut Score Recommendations After Seeing Round 3 Consequences Data

ALS	Should Be Lower N (%)	Leave As Is N (%)	Should Be Higher N (%)
Basic	1 (3)	27 (87)	3 (10)
Proficient	0 (0)	26 (84)	5 (16)
Advanced	0 (0)	28 (90)	3 (10)

Characteristics of the panelists recommending raises to the cut scores were reviewed (Table 33) to determine if all recommenders were sitting at the same table or were in the same rater group. All panelists recommending raises to the cuts had individual cut scores that were higher than the group cut for the level recommended. For example, panelist A1212 recommended raising both the Basic and Proficient cuts. This panelist's final Basic and Proficient cut scores (342 and 379, respectively) were higher than the group cuts (326 and 363, respectively). No other discernible characteristic pattern was identified. The seven panelists recommending raises came from four of the six tables, were in both groups A and B, and represented both teachers and members of the general public. Nonteacher educators were not represented, but there were only four nonteacher educators in the ALS. This additional analysis suggests that those recommending raises to the cut score were thinking independently and were not subject to any undue group influence.

Table 33: Characteristics and Cut Scores of Panelists Recommending Raising Cut Scores

Panelist	Table	Type	Gender	Race	Region	Panelist's Final Round 3 Basic Cut (Cut Given on Consequences Questionnaire)	Panelist's Final Round 3 Proficient Cut (Cut Given on Consequences Questionnaire)	Panelist's Final Round 3 Advanced Cut (Cut Given on Consequences Questionnaire)
		GP	M			325	360	431 **(431)**
		Teacher	F			342 **(328)**	379 **(370)**	420
		Teacher	F			345	381 **(365)**	409
		GP	F			330 **(332)**	376 **(375)**	420 **(420)**
		Teacher	M			323	365 **(366)**	419
		Teacher	F			337 **(336)**	370 **(370)**	407
		GP	F			324	363	419 **(419)**

Bold indicates panelist's recommended new cut.

OUTCOMES OF THE ACHIEVEMENT LEVEL SETTING PROCESS

There are three components of NAEP achievement levels: Achievement Level Descriptions, 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 the Governing Board'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" (National Assessment Governing Board, 2005). The Achievement Level Descriptions were developed by the Governing Board before the ALS meeting (see Appendix C), and were translated into cut scores during the meeting.

On process evaluation questions in both the Pilot Study and in the ALS meeting, panelists reported being satisfied with the ALDs. Table 34 summarizes Pilot Study Mapmark with whole booklet feedback, and ALS panelists' responses to questions concerning their satisfaction with the ALDs. Mean ratings of satisfaction with ALDs is consistently above 4 on a scale of 1-5, for both the Pilot Study and ALS.

Table 34: Pilot Study and ALS Mapmark with Whole Booklet Feedback Panelists' Responses to Questions about ALDs

Question		Mean Rating	
Location	Question	Pilot	ALS
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.	4.13	4.29
2-18	My own level of satisfaction with the Basic achievement level description is:	4.25	4.39
2-19	My own level of satisfaction with the Proficient achievement level description is:	4.38	4.35
2-20	My own level of satisfaction with the Advanced achievement level description is:	4.31	4.35
5-25	I believe that the achievement levels capture meaningful distinctions in economics performance as described in the ALDs.	4.38	4.48

Note: Questionnaire 2 was administered immediately after round 1 bookmark placements. Questionnaire 5 was administered at the conclusion of the meeting.

Panelists' average rating of their understanding of the ALDs is presented by level and round in Table 35. This question is asked for each level only at the conclusion of the first round, and is asked for all levels combined in succeeding rounds. At the conclusion of the first round, panelists were asked to indicate on a scale from 1-5 (1 = *totally inadequate*, 5 = *totally inadequate*): At the time I provided the round 1 bookmark placements, my understanding of the ________ achievement level description was. At the conclusion of the second and third rounds, the same question was asked for all three levels combined. In order to allow for comparison across rounds, the mean panelist rating for all three levels combined was also calculated for round 1. 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 35, panelists reported levels of understanding well above 4.0 early in the process, and understanding continued to increase slightly over rounds as panelists continued to study and apply the ALDs to their tasks.

Table 35: Understanding of ALDs

At the time I provided the/my round __ bookmark placements/cut score recommendations my understanding of the ___ achievement level description was …

(5 = *totally adequate*; 3 = *somewhat adequate*; 1 = *totally inadequate*)

	Round		
Level	1	2	3
Basic	4.61	-	-
Proficient	4.65	-	-
Advanced	4.65	-	-
All Combined	4.63	4.71	4.81

As these ALDs were used to anchor the process for establishing cut scores and, as the responses of panelists in the Pilot Study and ALS meeting to process evaluation questions concerning the ALDs are positive, ACT endorses the ALDs for use in representing the achievement levels set in this project.

Cut Scores

Table 36 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 3 medians, highlighted in yellow in Table 36, as the cut scores for the achievement levels (326 for Basic, 363 for Proficient, and 411 for Advanced). These numbers are on the ACT-NAEP like 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:

1) Distribution of cut scores by round.
2) Reliability of cut scores. This section includes reliability across different types of panelists, rater groups, tables, and meetings.
3) Reasonableness of results when compared to external sources of information. This section includes a comparison to national results on AP Micro and Macro

Economics as well as on the National Council on Economic Education's Test for Economic Literacy.

4) Special Studies Results. This section provides data from two special studies designed to compare classification of items and student booklets based on the ALS results with their classification by an independent panel.

Table 36: NAEP Grade 12 Economics ALS Cut Scores by Panelist and Medians by Group and Table

			Basic Scale Value			Proficient Scale Value			Advanced Scale Value		
Group	Table	ID	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
A	1		336	332	337	358	363	363	408	409	411
			324	328	328	342	359	363	405	405	411
			341	327	327	364	365	365	403	407	411
			324	337	329	358	362	362	402	413	407
			313	327	325	343	360	363	409	404	409
	2		313	336	329	347	368	363	411	411	411
			312	325	325	341	358	358	431	431	425
			320	320	320	350	363	363	408	408	408
			312	322	322	315	359	359	491	408	408
			313	325	325	348	360	360	402	436	431
	3		312	340	326	340	367	367	427	427	427
			339	342	342	375	379	379	420	420	420
			342	345	345	360	381	381	405	409	409
			329	329	329	386	380	380	434	423	419
			332	332	332	360	360	360	405	405	405
B	4		320	320	323	353	365	365	419	419	419
			294	325	325	365	358	358	409	409	409
			294	324	324	360	360	360	427	427	420
			294	294	323	345	368	368	427	427	427
			321	329	330	364	374	376	420	420	420
	5		324	323	323	366	364	364	417	411	411
			321	321	335	366	364	364	409	408	408
			317	337	337	340	386	370	400	409	407
			321	325	338	353	380	380	394	427	427
			321	321	324	365	363	363	427	416	419
	6		320	320	326	337	369	369	417	417	417
			332	323	323	364	359	359	403	403	410
			325	323	323	365	364	364	392	398	407
			325	325	325	350	353	353	403	403	403
			325	325	325	353	359	359	391	403	411
			338	340	340	357	357	357	427	429	416
Medians	**All**		**321**	**325**	**326**	**357**	**363**	**363**	**409**	**411**	**411**
	Group	**A**	**324**	**329**	**328**	**350**	**363**	**363**	**408**	**409**	**411**
		B	**321**	**324**	**325**	**359**	**364**	**364**	**413**	**414**	**414**
	Table	**1**	**324**	**328**	**328**	**358**	**362**	**363**	**405**	**407**	**411**
		2	**313**	**325**	**325**	**347**	**360**	**360**	**411**	**411**	**411**
		3	**332**	**340**	**332**	**360**	**379**	**379**	**420**	**420**	**419**
		4	**294**	**324**	**324**	**360**	**365**	**365**	**420**	**420**	**420**
		5	**321**	**323**	**335**	**365**	**364**	**364**	**409**	**411**	**411**
		6	**325**	**324**	**325**	**355**	**359**	**359**	**403**	**403**	**411**

Distribution of Cut Scores by Round

The variability of cut scores within rounds and levels was assessed. 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. As panelists review results and feedback together, outliers and variability tend to decrease as panelists gain a shared sense of borderline performance and as they become aware of the group cut score. The variability of cut scores across panelists in the Economics ALS decreased by round. Figure 43 is a plot of the Mean Absolute Deviation (MAD) of cut scores of individual panelists from the group cut score by round in the ALS. The ALS MAD was largest for the Advanced level in round 1 and then decreased in subsequent rounds. Differences between panelists' cut score recommendations decrease over rounds with the greatest amount of convergence between rounds 1 and 2. In addition, the lack of large increases in the MAD from round 2 to round 3 indicates that there were no extreme reactions among panelists to student performance data in the ALS. These findings are consistent with results ACT has obtained in previous standard setting work for the Governing Board.

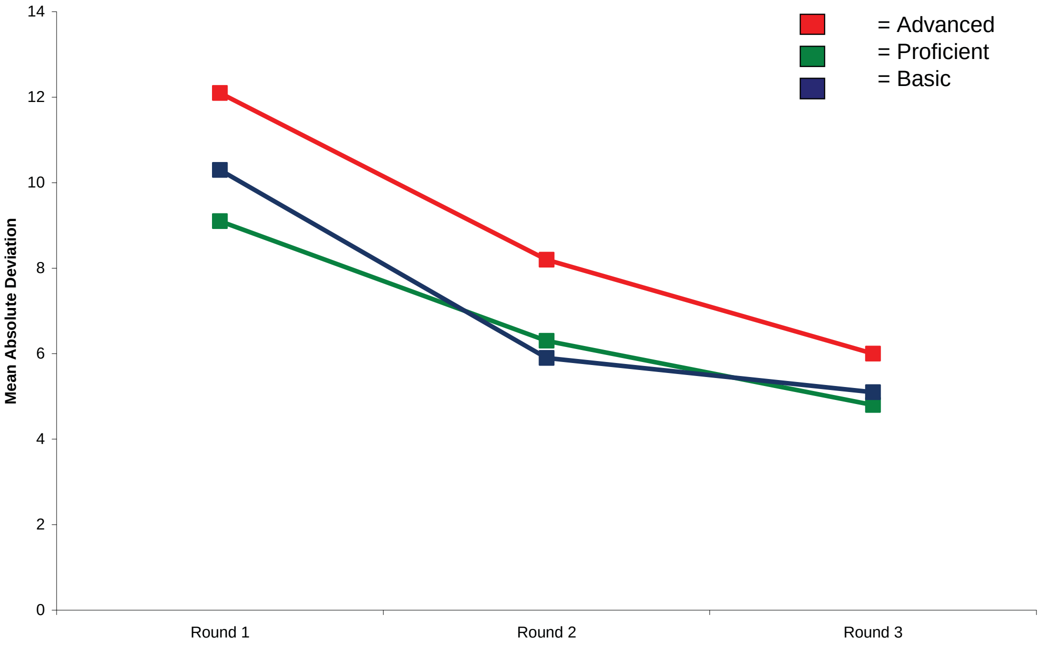


Figure 43. Mean Absolute Deviation (MAD) of cut scores from median by round.

A study of the change in cut scores by level and round provides additional information about how panelists were responding to the feedback provided. Table 37 presents the number and percent of panelists whose cut scores increased from the previous round, decreased, or had no change. The patterns in this table are similar to the patterns seen in previous standard settings for the Governing Board. The largest frequency of change is from round 1 to round 2, indicating the incorporation of information gleaned from the booklets into their judgments. At each level from round 2 to round 3, the majority of panelists did not change their cut scores in response to the student performance data.

In comparison to previous standard setting studies, the proportion of panelists making changes in the ALS at the Basic and Advanced levels from rounds 1 to 2 is somewhat smaller than proportions seen in the past. Process evaluation data indicated that ALS panelists were slightly more confident in their cut scores at each round than panelists have indicated in previous standard settings conducted for the Governing Board, which may explain this difference.

Table 37: Number and Percent of Panelists Who Changed Their Cut Scores Between Rounds in the ALS

	Basic			Proficient			Advanced		
Rounds of Change*	Increase n (%)	No Change n (%)	Decrease n (%)	Increase n (%)	No Change n (%)	Decrease n (%)	Increase n (%)	No Change n (%)	Decrease n (%)
R1 to R2	16 (52)	10 (32)	5 (16)	21 (68)	3 (10)	7 (23)	10 (32)	15 (48)	6 (19)
R2 to R3	8 (26)	19 (61)	4 (13)	3 (10)	26 (84)	2 (6)	8 (26)	16 (52)	7 (23)

As shown in Table 38, differences between the mean and median cut scores were generally small. The largest difference was five points at the Advanced level in round 1. This may have been due to one outlier at the Advanced level in round 1 who set the Advanced cut score at the top of the scale (see panelist A1209 in Table 36). This difference decreases in subsequent rounds as the outlier disappears.

Another observation in Table 38 is that the median tends to be lower than the mean cut score. Seven of the nine signed differences in Table 34 are negative. A predominance of negative values means that panelists cut scores are slightly positively skewed—the highest cut scores recommended by panelists tend to be higher than one would expect in a symmetrical distribution of cut scores.

Table 38: Mean and Median Cut Scores and Difference by Round for ALS

	Median			Mean			Median - Mean		
Level	R1	R2	R3	R1	R2	R3	R1	R2	R3
Basic	321	325	326	321	327	329	0	-2	-3
Proficient	357	363	363	355	365	365	2	-2	-2
Advanced	409	411	411	414	414	414	-5	-3	-3

Reliability of Cut Scores

The reliability of cut scores emerging from a standard setting process is typically thought of in regard to how consistent the cut scores are across tables, rater groups, and panelist type, and how close the final cut scores from the process would be if the process were performed on two occasions with few differences.

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 did not identify any effects that called the results of the ALS meeting into question or raised serious questions about the process.

As there is no satisfactory method of estimating the differences between groups on their median cut scores and as the mean and median cut scores were highly similar, 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.

Item pool (rater group) effects were not statistically significant at any level or round (Figure 44), although group A medians were consistently lower than B's for Advanced and Proficient and higher for Basic.

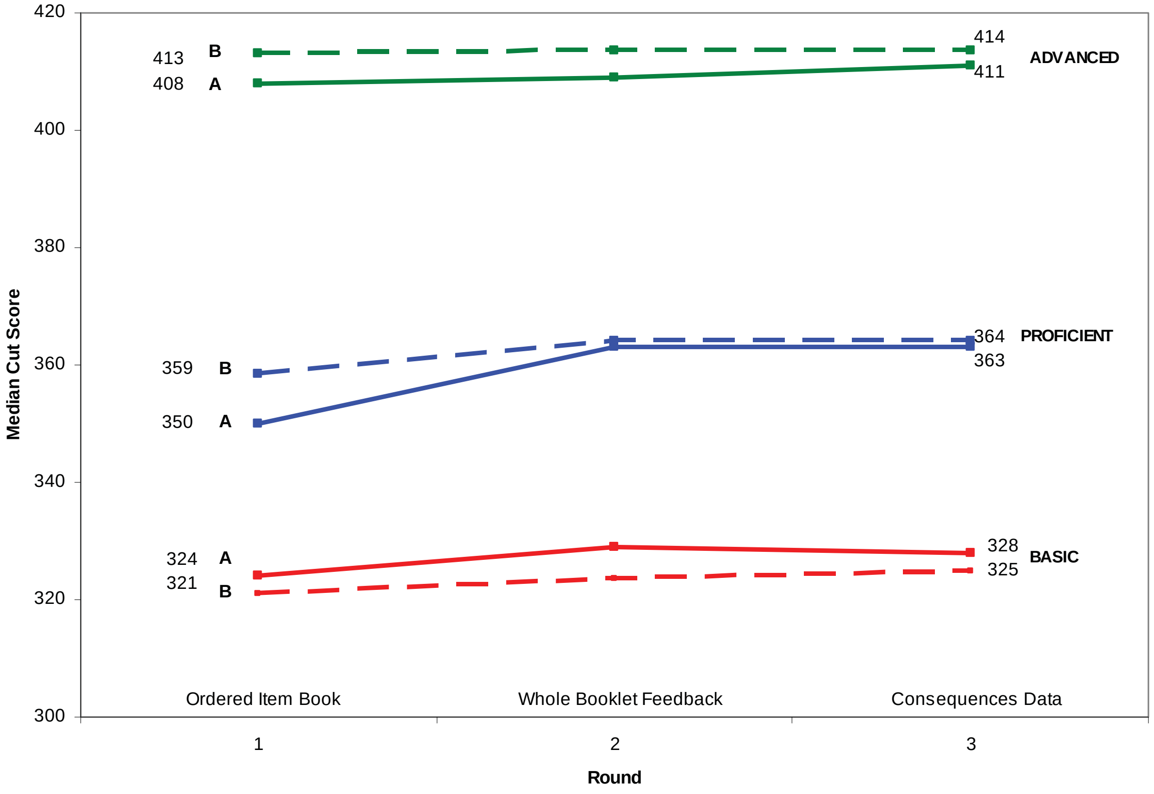


Figure 44. Median cut scores by item pool group.

Figure 45 shows table medians by round and achievement level. Table group effects at the final round were statistically significant only at the Basic level. The graph in Figure 45 of the table group effects based on the table mean cut scores at the Basic level would seem to illustrate that table differences were larger at round 1 than round 3. However, variance in the first round was greater within than between tables, whereas by the final round, the variance had decreased substantially so as to render slight mean differences significant (see Table 36 for cut scores by panelist within groups and tables). Table 39 shows that the largest within-group difference between table median Basic cut scores was 31 points at round 1 and only 11 points at the final round.

Finally, differences in cut scores between different genders, races, geographic regions, and panelist types (teacher, nonteacher, general public) were not statistically significant. Table 39 shows that the largest difference between median cut scores by panelist type was 17.5 points at the Advanced level at round 1, but only 11.5 points at the final round.

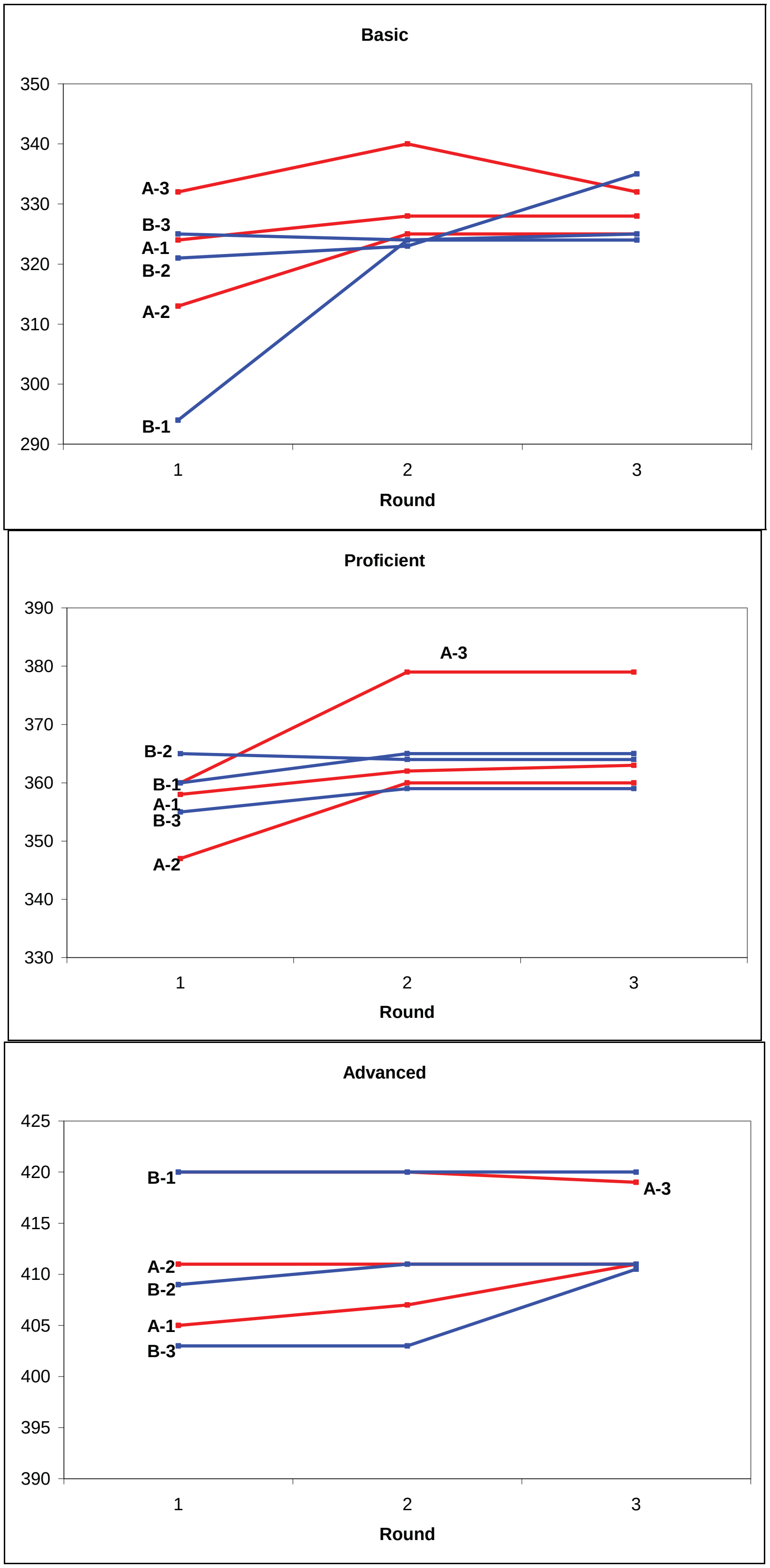


Figure 45. Median cut scores by level and table.

Table 39: Medians and Mean Absolute Difference (MAD) of Cut Scores by Factor Level

	Round 1							Round 3						
		Basic		Proficient		Advanced			Basic		Proficient		Advanced	
Factor	N	Median	MAD	Median	MAD	Median	MAD	N	Median	MAD	Median	MAD	Median	MAD
Type														
Teacher	18	320.5	9.1	359	10.5	409	12.1	18	326	4.8	364	6	411	6
Non Teacher	4	322.5	9.2	351.5	10.8	402.5	14	4	327	4.9	360.5	5.9	407.5	7.3
Gen. Public	9	321	9.1	357	10.3	420	14.2	9	325	4.8	363	5.1	419	7.6
Max. Diff.		2		7.5		17.5			2		3.5		11.5	
Ethnicity														
White	26	321	9.1	353	10.4	409	12.1	26	326	4.8	363.5	5.1	411	6
Non-White	5	325	9.6	360	10.6	408	12.2	5	324	5.3	363	5.1	408	7
Max. Diff.		3.5		7		1			2		0.5		3	
Gender														
Male	15	321	9.1	357	10.3	409	12.1	15	325	4.8	363	5.1	411	6
Female	16	321	9.1	355.5	10.3	409.5	12.1	16	326.5	4.9	363	5.1	411	6
Max. Diff.		0		1.5		0.5			1.5		0		0	
Region														
Midwest	6	321	9.1	353	10.4	405.5	12.7	6	327.5	5	361.5	5.6	418	7.3
Northeast	3	320	9.2	353	10.4	419	13.8	3	326	4.8	365	5.5	419	7.6
South	16	322.5	9.2	362	11.4	410	12.2	16	325.5	4.8	363	6	410.5	6.1
West	6	321	9.1	347.5	12	404	13.3	6	326	2	363	3.5	409	8.5
Max. Diff.		1.5		14.5		15			2.0		3.5		10	
Group														
A	15	324	9.4	350	11.1	408	12.2	15	328	5.1	363	5.1	411	6
B	16	321	9.1	358.5	10.4	413	12.6	16	325	4.8	364	5.2	413.5	6.4
Max. Diff.		3		8.5		5			3		1		2.5	
Table														
A:1	5	324	9.4	358	10.3	405	12.8	5	328	5.1	363	5.1	411	6
A:2	5	313	12	347	12.2	411	12.3	5	325	4.8	360	6.1	411	6
A:3	5	332	13.2	360	10.6	420	14.2	5	332	6.7	379	14.2	419	7.6
Max. Diff.		21		13		15			7		19		8	
B:4	5	294	27.1	360	10.6	420	14.2	5	324	6.6	365	5.5	420	8.1
B:5	5	325	9.6	355	10.4	403	13.7	5	325	4.8	359	6.6	410.5	6.1
B:6	6	321	9.1	365	12.6	409	12.1	6	335	8.3	364	5.2	411	6
Max. Diff.		31		10		17			11		6		9.5	
Overall	31	321	9.1	357	10.3	409	12.1	31	326	4.8	363	5.1	411	6

Cut scores by round for the Pilot Study and ALS using the Mapmark with whole booklet feedback method are presented in Table 40. On a 300-point scale, ranging from 203 to 503, the final Basic cut scores from the two meetings differed by 12 points (326 for the ALS vs. 338 for the Pilot Study), the final Proficient cut scores differed by 5 points (363 for the ALS vs. 368 for the Pilot Study), and the final Advanced cut scores differed by 10 points (411 for the ALS vs. 421 for the Pilot Study).

Table 40: Group Cut Scores by Round and Level for Pilot Study and ALS Using Mapmark with Whole Booklet Feedback

	Basic			Proficient			Advanced		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Pilot Study	332	338	338	365	368	368	423	422	421
ALS	321	325	326	357	363	363	409	411	411
Difference	-11	-13	-12	-8	-5	-5	-14	-11	-10

The standard error of the cut score is an estimate of the uncertainty in the reported cut score (the median cut score across panelists) due to various sources of error. The standard error of the difference of two cut scores combines the estimates of the standard error of each individual cut score. Unfortunately, ACT can recommend no single, sure method for estimating the standard error of the final 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 cut score from the previous round. Panelist cut scores after round 1 are influenced by the group cut score and cut score distribution. Panelists are generally more comfortable being close to the middle so there is a regression to the round 1 group cut score. Estimates of the standard error of the final cut score do not account for a fundamental regression to the median of previous rounds, motivated by panelists' desire for conformity, as well as for the effects of criterion-referenced feedback. For this reason, estimates of the standard error at the final round tend to be smaller and are more likely to underestimate differences between replications of a method using the same item pools but different groups of panelists. In addition, cut scores established in rounds 2 and 3 are based on the baseline established in the first round, and do not tend to vary substantially from the previous round. For this reason, an understanding of the differences between cut scores is most informed by an analysis of results from round 1.

Table 41 presents the standard error estimates for the group cut scores (medians) for round 1 and the final round for each achievement level in the Pilot Study and ALS, with the standard errors calculated using two distinct nonparametric methods (Maritz & Jarrett, 1978; bootstrap, see Efron & Gong, 1983). As expected, the standard errors decreased for both methods from round 1 to the final round. The standard error of the difference between the ALS and Pilot Study cut scores is shown in Table 42 and is compared to the absolute value of the actual difference. The actual difference between the round 1 cut scores was close to one standard error of the difference for Proficient and two standard errors for Basic and Advanced. As estimates of the standard error at the final round are underestimates, the relevant round for interpretation of differences is the first round.

Table 41: Estimates of Standard Error of the Group Cut Scores across Levels and Rounds for the Pilot Study and ALS, Using Two Distinct Nonparametric Methods

	Pilot/ALS	Basic		Proficient		Advanced	
Method		Round 1	Final	Round 1	Final	Round 1	Final
Maritz-Jarrett	ALS	1.8	1.4	3.4	0.8	4.1	2.7
	Pilot	4.4	1.1	4.5	1.8	6.5	3.4
Bootstrap	ALS	1.9	1.4	3.5	0.8	3.8	2.5
	Pilot	3.7	1.0	3.8	1.4	5.5	3.0

Table 42: Estimates of Standard Error of the Difference in the Pilot Study and ALS Group Cut Scores by Levels and Rounds Compared to Absolute Value of Actual Difference

Standard Errors	Basic		Proficient		Advanced	
of the Difference	Round 1	Final	Round 1	Final	Round 1	Final
Maritz-Jarrett	4.8	1.8	5.6	2.0	7.7	4.3
Bootstrap	4.6	1.7	5.2	1.6	6.7	3.9
Observed |D|	11	12	8	5	14	10

Differences in cut scores may be due to factors expected to affect cut scores, which vary across meetings using the same method, but which are not represented in the standard error estimates. Such factors include physical accommodations, presence of observers, interactions among panelists over rounds, random variation, and panelist understanding of the differences in the purpose of the meeting. ACT and our Technical Advisory Committee carefully reviewed procedural validity and internal consistency data from the ALS to determine if differences may have been due to procedural or internal validity factors. In addition, ACT reviewed panelists' qualifications. Results indicated no differences in panelist qualifications between the Pilot Study and ALS, that the ALS procedural results were stronger than or comparable to that of the Pilot Study, and that internal consistency emerged as expected. The conclusion was that there is no reason to doubt the results of the ALS, and that the differences between the results from the two meetings may be due to the following:

- *Panelist understanding of differences in the purpose of the meetings.* Panelists in the Pilot Study clearly understand that the cut scores they establish will not have national implications but, instead, will inform development and refinement of the method. Economics is only the second NAEP subject area for which standards have been set following the January 8, 2002 establishment into law of No Child Left Behind. In this era of great emphasis on accountability, panelists in ALS meetings may be more likely to set lower cut scores because they know that the scores will be used for reporting the national results of student performance in economics. Panelists in the Pilot Study knew the results would not be reported. This difference, along with the changes to the briefing book and the number of panelists, means that the ALS is not an exact replication of the Pilot Study.

- *The interpretability of the Achievement Level Descriptions (ALDs).* In economics, unlike in some subjects, the ALDs do not indicate specific content that students should know or specific skills that students should have at each level, but instead list concepts and indicate that students at the Basic, Proficient, and Advanced levels will have mastery of a “limited,” “broader,” or “extensive” set of these concepts, respectively. This may allow for greater variability in translating these descriptions to scores on the assessment.
- *National variation in economics instruction and standards.* Research conducted by the National Council on Economic Education has indicated that state requirements for economics education vary substantially, and large proportions of students have never had a formal course in economics or personal finance. Economics content is often embedded throughout the curriculum. In addition, there is considerable variation among courses with economic and personal finance content. Given this variation, panelist interpretations of “limited,” “broader,” and “extensive” mastery of economics concepts may vary widely.

Reasonableness of Results when Compared to External Sources of Information

The distribution of student performance relative to the achievement levels, referred to here as the achievement level percentages, provides external information as to the reasonableness of the cut scores. Figure 46 shows the grade 12 economics achievement level percentages resulting from the final cut scores established in the ALS. As indicated earlier, the majority of panelists (between 84% and 90% depending on the achievement level) felt that these percentages were reasonable and that the cut scores should not be changed to alter the percentages.

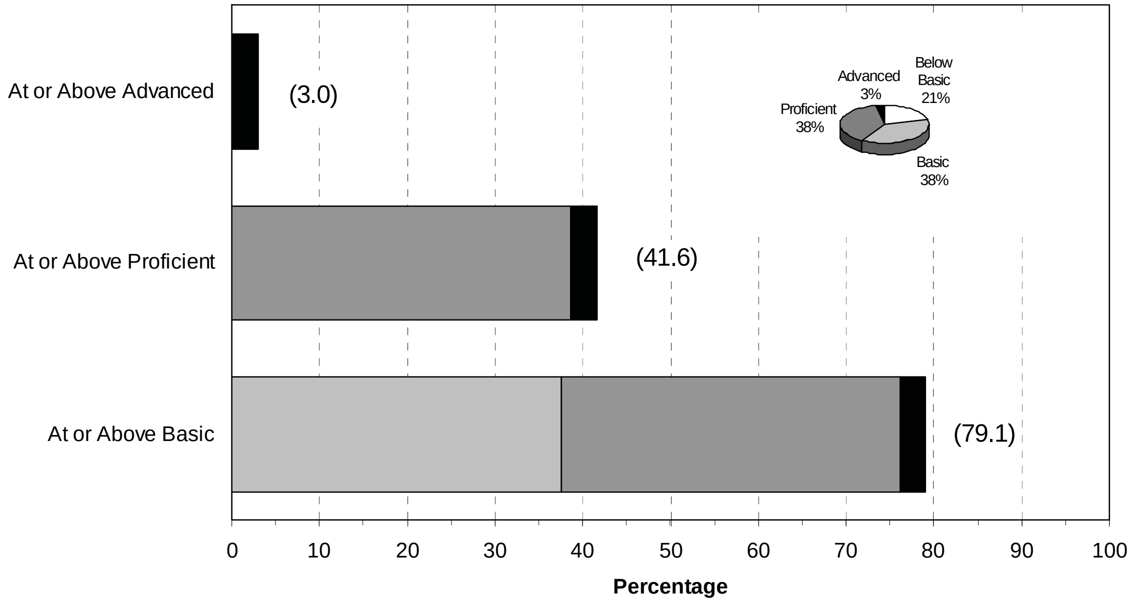


Figure 46. Percent of students at or above each achievement level.

Comparison to Economics Advanced Placement (AP)

Because NAEP results are not reported on an individual level, matching individual scores on the NAEP to scores on other assessments of the same subject area is not possible. However, historically, ACT has requested from the College Board statistics on the proportion of grade 12 students scoring at each level on Advanced Placement (AP) exams in the same subject for comparison to the proportion of students scoring at the Advanced level on the NAEP. To this end, ACT has received and reviewed the percentage of 2006 graduating seniors taking AP MacroEconomics and MicroEconomics exams and the percent of students at each AP Economics score level for each assessment individually and the two assessments combined. Because some of the students in the total assessment count will have taken both exams and been counted twice in the totals, this percent may be slightly inflated. Our TACSS economics content expert has indicated that a score of 3 or higher on either AP exam would be comparable to Advanced performance on the NAEP. As is apparent in Table 43, 1.02% of the 2006 graduating seniors taking AP Economics exams scored a 3 or higher. The percentages provided in Table 43 do not account for students taking AP Economics, honors, and international baccalaureate courses who do not take the AP exam. ACT's TACSS content expert has indicated that some students with this coursework would also be expected to score at an Advanced level, rendering the 3% of students scoring at the Advanced level plausible.

Table 43: Percent of 2006 Graduating Seniors Taking AP Macroeconomics and Microeconomics Exams and Their Corresponding Scores

	Percent of Graduating Seniors		Total Macro	
Score	Macro	Micro	And Micro	
5	0.14	0.10	0.23	1.02%
4	0.27	0.18	0.45	
3	0.21	0.14	0.34	
2	0.26	0.12	0.37	
1	0.33	0.17	0.50	
Total	1.20	0.70	1.90	

Comparison to Test of Economic Literacy (TEL)

Another national assessment of high school student economics knowledge, skills, and abilities is the National Council on Economic Education’s (NCEE) Test of Economic Literacy (TEL). Content in both the TEL and the economics NAEP are based on NCEE’s Voluntary National Content Standards. To assess reasonableness of the NAEP economics achievement level percentages in comparison to TEL results, ACT contracted with three economics experts including our TACSS content expert, who had been actively involved in the development of the NAEP economics assessment and had played key roles in the standard setting process. These experts had all worked with the NCEE, one had played a leadership role in the organization, and all three were familiar with the TEL to varying degrees. Each was provided with the achievement level percentages and the exemplar items illustrating performance at each achievement level. The experts were asked to compare the economics NAEP and the NCEE TEL to determine if the content and difficulty of the two assessments are comparable, and consequently, if the achievement level percentages are comparable.

Two of the three content experts felt that the two assessments should not be compared. Although both assessments were based on the Voluntary National Content Standards in economics developed by NCEE, they serve different purposes. The TEL was developed to serve as a pre-and post-test in an economics principles course. The design of the NAEP exam, by contrast, was to test students on economics content that may be learned across a variety of courses including courses such as consumer economics or personal finance. In part for this reason, the economics NAEP includes many items in Household and Individual contexts, which are items related to personal finance (e.g., earning, spending, saving, borrowing, and investing). These contexts are not explicitly included in the TEL and ACT content experts indicate that fewer than 20% of the total TEL items could be considered as containing content of a Household and Individual nature.

The lack of Household and Individual contexts on the TEL may cause the two assessments to differ in their content. Of the 225 score points on the NAEP assessment, almost half, or 102, are in the Household and Individual contexts. ACT reviewed these score points to determine their distribution across the three achievement levels based on an RP criterion of 0.67, the same criterion used in the ALS and used to select exemplars. The distribution of these items, of all NAEP economics items across the achievement scale, and the median scale value for Household, Individual, and all NAEP economics items are provided in Figure 47. Household and Individual items are slightly less difficult than the entire pool of items at a median scale value of 369 and 366 respectively, compared to the overall median of 379. In addition, very few items in these contexts are at scale values above the Advanced cut score of 411. A count of Household, Individual, and all items in each achievement level is provided in Table 44.

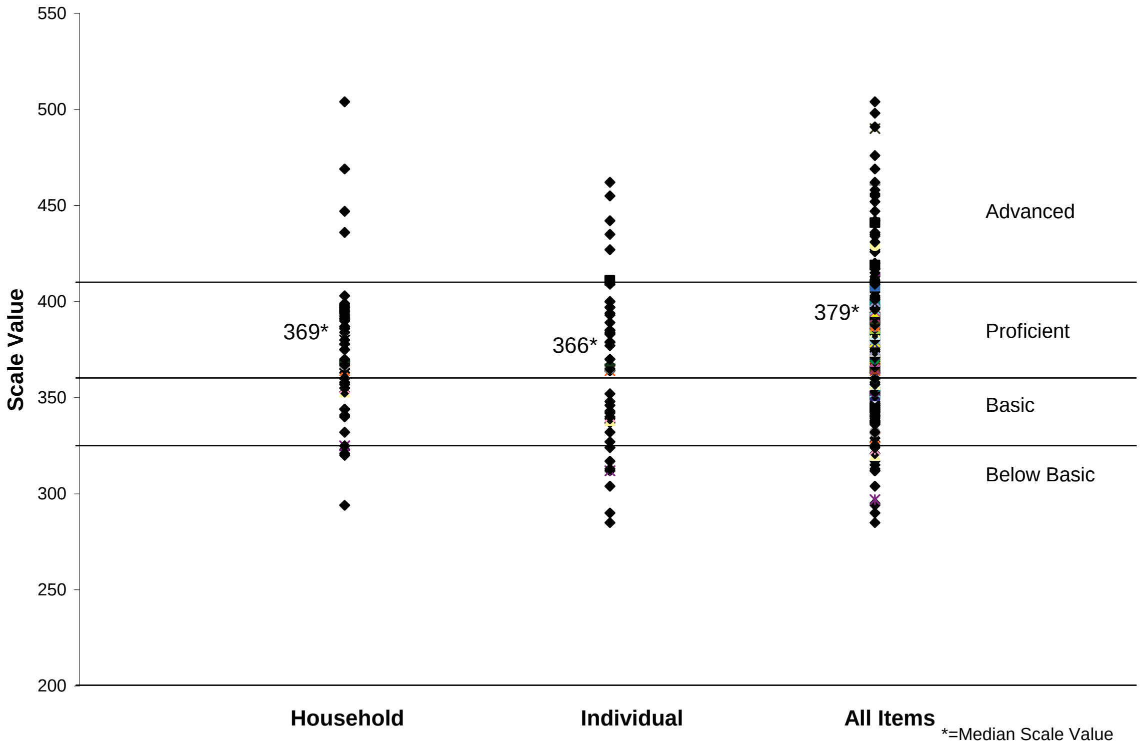


Figure 47. All scale values and median scale value of NAEP items in the Household and Individual contexts and of all NAEP grade 12 economics items using an RP criterion of 0.67.

Table 44: Percent and Number of Individual, Household, and All NAEP Economics Items in Each ALS Achievement Level

			Household +	
	Household	Individual	Individual	All Items
Level	Row % (n)	Row % (n)	Row % (n)	(n)
Below Basic	30%	40%	70%	
	(6)	(8)	(14)	(20)
Basic	26%	22%	48%	
	(14)	(12)	(26)	(54)
Proficient	26%	22%	48%	
	(28)	(24)	(52)	(108)
Advanced	9%	14%	23%	
	(4)	(6)	(10)	(43)

These results illustrate that 70% of the items at the Below Basic level are Household and Individual context items. This may contribute to differences in content and difficulty between the two tests.

The differences in use and content between the NAEP and TEL assessments suggest that a comparison between proportions of students performing at the different achievement levels will not be meaningful. In addition, because achievement levels have not been established on the TEL, such a comparison is not viable. ACT, therefore, does not recommend comparing results from the TEL with results from the NAEP.

Special Studies Results

In addition to analyses of the data collected in the Achievement Level Setting meeting and comparisons of results to results on other assessments, two external Special Studies were conducted. These studies were designed to provide additional information to the Governing Board on the reasonableness of the results of the ALS. They are described in a separate report (ACT, Inc. 2007) and are summarized here.

From January 11-13, 2007, two Special Studies were conducted in St. Louis, Missouri. Both studies included a panel of the same 13 panelists representing teacher and nonteacher educators. The studies began with a Booklet Classification Study which lasted for the first day and a half and ended with an Item Classification Study which concluded the three-day meeting. These studies allow for the comparison of the empirical classifications of the booklets and items from the ALS results to the Special Studies panelist classifications. The empirical classification of a booklet is the achievement level into which the student's score maps based on the ALS cut scores. The empirical classification of an item is the achievement level into which the item maps based on the ALS cut scores (RP = 0.67). If there is a reasonable correspondence between the empirical classifications of student booklets and panelist classifications of booklets completed by a panel comprised of teachers and nonteacher educators who did not participate in the ALS but are familiar with economics, then there is evidence that students performing within the cut score ranges know and can do the types of things that the ALDs specify. By the same logic, if there is a reasonable correspondence between the empirical classifications and panelist classifications of items, then there is evidence that items which students performing within the cut score ranges know and can answer are related to knowledge, skills, and abilities described in that level's ALD. For both studies, the principle is that if a group of panelists separate from those who served in the ALS but with similar characteristics agrees with the empirical classifications of the booklets and items, this supports the translation of the Achievement Level Descriptions to the score scale by the ALS panelists.

Panelist Recruitment

The same pool of nominees was used to recruit panelists for the Pilot Study, ALS, and Special Studies. The panelists were recruited with efforts made to ensure that the panel had proportional representation by gender, race, and geographic region, however, unlike the ALS and Pilot Study, only teachers and nonteacher educators were invited to participate in the Special Studies. No consistent pattern had been discerned in previous studies to indicate that classification would vary significantly by type of panelist (Loomis, 2000). Demographic characteristics of participating panelists are provided in Table 45.

Table 45: Demographics of Panelists Participating in the Special Studies

Type	Males	Females	Caucasian	African American	Hispanic	MW	NE	SO	WE	Total
Teacher	6	4	8	1	1	1	3	4	2	10
Nonteacher	2	1	3			2	1			3
Total	8	5	11	1	1	3	4	4	2	13

Booklet Classification Study

The Booklet Classification Study was conducted similarly to one conducted by ACT for the validation of the 1998 civics standards (Loomis, 2000). In it, panelists were asked to classify examples of student performance on test booklets into the achievement levels using only the Achievement Level Descriptions as the criterion for classification. The booklet classification task was holistic and required panelists to consider the overall performance of the student rather than to estimate the performance of students on each item. The booklet scores were not revealed to panelists, nor were scores on individual items indicated within the booklets. Panelists classified booklets into achievement level categories using the Achievement Level Descriptions to judge the performance represented by the booklet as a whole.

Forty booklets from four forms were selected from the total set of students who participated in the assessment. . Any booklets for which missing data would be considered "not reached," i.e., not administered, were not included. In general, these are booklets where the student failed to answer the last question in one or both of the blocks on the test form, indicating he or she may not have completed the test. The forms used included four blocks, three of which are slated for release, and consisted of about 40% of the items in the assessment. Because the Special Studies were conducted prior to the ALS, the 40 booklets were selected to be in the middle of the achievement level range on the basis of the cut scores set in the Pilot Study. Once the ALS results were calculated, the empirical classifications for these same booklets were calculated, based on the ALS cut scores. Booklets were distributed in each ALS achievement category with 7 in the Below Basic range, 13 in Basic, 13 in Proficient, and 7 in Advanced. Two of these booklets' raw scores were close to the cut score set for Advanced in the ALS (one above and one below the cut) and were used to illustrate borderline Advanced in the ALS. One booklet's score was just below the ALS Proficient cut score, and one was just below the ALS Basic cut score and was at the same scale value as a booklet used to illustrate borderline Basic in the ALS. The mean, minimum, and maximum ACT NAEP-like scale scores for the booklets empirically classified into each achievement level category are provided in Table 46.

Table 46: Mean, Minimum, and Maximum ACT NAEP-Like Scale Values for Performance Scores of Student Booklets at Each ALS Achievement Level

ALS achievement level	N	Mean	Min	Max
Below Basic	7	316.7	310	325
Basic	13	352.3	345	360
Proficient	13	392.3	380	409
Advanced	7	443.6	412	503

Before beginning classification of the 40 booklets, panelists were asked to conduct a practice classification session with 10 booklets. They were given one hour for this practice classification and were told that the rate (10 booklets per hour) was the approximate rate necessary for the actual Booklet Classification Study. Following the practice, panelists were given the opportunity to discuss their classifications. This discussion was a whole group discussion so that all panelists heard all comments. This helped panelists gain a sense of how their classification judgments compared to others in the group.

Panelists were then asked to classify the 40 booklets and were told to use the Achievement Level Descriptions as the criterion for classifying performances represented by these booklets. They were told only that booklets were selected such that the score of at least one booklet fell within the range of each achievement level. Although the scores of the booklets and individual item scores were not revealed, panelists did have scoring rubrics for all items, and they could refer to those rubrics. Panelists were instructed in the method and in marking their classification forms. In particular, they were told that they were to classify the booklets according to the ALDs and to base their classifications on a holistic judgment. The facilitator stressed that scoring booklets was not the task and that booklet scores were not necessary in order to perform the task. Instead, the panelist was to gain a holistic sense of the student's performance and then to place each student's booklet into an achievement level using the Achievement Level Descriptions as reference. Four achievement level categories were available for the panelists to select (Below Basic, Basic, Proficient, and Advanced). Booklets were numbered from 1-40, and the numbering was unrelated to the score of the booklet. The results of this round of classifications were compared to the empirical booklet score classifications that resulted from the achievement levels setting meeting.

Booklet Classification Results

The results of panelist booklet classifications were compared to the empirical booklet score classifications based on cut scores from the achievement level setting meeting. Table 47 shows the correspondence between individual panelists' classifications of booklets and the empirical score classifications that were based on the results of the ALS. The empirical distribution of booklets based on the ALS results was 7 Below Basic, 13 Basic, 13 Proficient, and 7 Advanced.

Table 47: Economics 2007 Booklet Classification Outcomes
Correspondence of Individual Panelist Classifications of Student Booklets into Achievement Level Categories and Empirical Score Classifications of Student Booklets into Achievement Level Categories

Achievement level classification *by empirical scores* of student booklets	Achievement level classification of student booklets by *panelists*			
(ACT NAEP-Like cut scores)	Below Basic	Basic	Proficient	Advanced
Below Basic (<326) (n = 7 booklets, n = 91 classifications)	**59%** **(n = 54)**	41% (n = 37)	-	-
Basic (326-362) (n = 13 booklets, n = 169 classifications)	15% (n = 25)	**79%** **(n = 134)**	6% (n = 10)	-
Proficient (363-410) (n = 13 booklets, n = 169 classifications)	-	25% (n = 42)	**63%** **(n = 106)**	12% (n = 21)
Advanced (> 410) (n = 7 booklets, n = 91 classifications)	-	-	22% (n = 20)	**78%** **(n = 71)**
Total (n = 40 booklets, n = 520 classifications)	15% (n = 79)	41% (n = 213)	26% (n = 136)	18% (n = 92)

Bold entries are for cells that would represent classification agreement.
Simple Kappa = 0.59, Weighted Kappa = 0.72

Overall, there was a 70% agreement between the panelists' individual classifications and the empirical classifications (365 out of 520 total classifications). The lowest level of agreement was 59% for Below Basic and the highest was 79% for Basic. These results were compared to the 1998 civics booklet classification results presented in Table 48 (Loomis, 2000). In the 1998 civics study, the same number of booklets (40) distributed across the 1998 civics achievement scale (7 Below Basic, 13 Basic, 13 Proficient and 7 Advanced) were classified by 11 panelists.

Table 48: Civics 1998 Booklet Classification Outcomes Correspondence of Teachers' Classifications of Student Booklets into Achievement Level Categories and Empirical Score Classifications of Student Booklets into Achievement Level Categories

Achievement level classification *by empirical scores* of student booklets (ACT NAEP-Like cut scores)	Achievement level classification of student booklets *by panelists* Below Basic	Basic	Proficient	Advanced
Below Basic (<149.2) (n = 7 booklets, n = 77 classifications)	**92%** **(n = 71)**	8% (n = 6)	-	-
Basic (149.2-165.39) (n = 13 booklets, n = 143 classifications)	44% (n = 63)	**55%** **(n = 78)**	1% (n = 2)	-
Proficient (165.4-177.89) (n = 13 booklets, n = 143 classifications)	2% (n = 3)	50% (n = 71)	**43%** **(n = 62)**	5% (n = 7)
Advanced (≥ 177.90) (n = 7 booklets, n = 77 classifications)	-	-	52% (n = 40)	**48%** **(n = 37)**
Total (n = 40 booklets, n = 440 classifications)	31% (n = 137)	35% (n = 155)	24% (n = 104)	10% (n = 44)

Bold entries are for cells that would represent classification agreement.
Simple Kappa = 0.41, Weighted Kappa = 0.60

In civics, there was 56% agreement between the panelists' classifications and the empirical classifications of the booklets. The lowest level of agreement was 43% for Proficient and the highest was 92% for Below Basic. Overall, this is almost a 15% lower level of agreement than the economics results. In general, results of the Booklet Classification Study provide support for the ALS cut scores, and indicate a greater degree of agreement than the findings from the 1998 civics Booklet Classification Study.

Item Classification Study

During this study, panelists were presented with all 186 economics items. For constructed response items, panelists were instructed to consider each score point independently for a total of 225 items and score points. First, panelists independently classified the items into the achievement levels based on content and perceived difficulty. For each item, panelists were asked to review the item and then to review each achievement level description. Starting with the Below Basic achievement level, they were instructed to ask themselves, would at least two-thirds of the students at this level be able to answer this item correctly? If the answer to this question was no, they were asked to look at the next higher achievement level description and ask the same question until they were able to identify the achievement level into which to classify the item. Following independent classification, panelists worked together in their table group to come to agreement. Agreement was not forced, but was encouraged. After the discussion, the panelists were asked to finalize their item classifications.

Item Classification Results

Panelists’ classifications of the items were compared to item empirical classifications based on an RP criterion of 0.67, the same criterion that was used in the ALS. If there is a strong relationship between the cut scores and the Achievement Level Descriptions, the expectation was that there would be a high level of agreement between panelist and empirical classifications at the RP criterion used in the ALS. Table 49 shows the correspondence between panelists’ classification of items and the empirical score classifications that were based on the results of the ALS. Final panelists’ classification was the median panelist response for each item after completion of the group task at the end of the Item Classification Study.

Table 49: Economics 2007 Item Classification Outcomes
Panelists’ Judgments vs. Performance Level at RP .67, Economics ALS Cut Scores

		Median Panelist Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification	Below Basic	**45%** **n = 9**	50% n = 10	5% n = 1		n = 20
	Basic	6% n = 3	**76%** **n = 41**	19% n = 10		n = 54
	Proficient	1% n = 1	37% n = 40	**59%** **n = 64**	3% n = 3	n = 108
	Advanced		12% n = 5	67% n = 29	**21%** **n = 9**	n = 43
	Total	n = 13	n = 96	n = 104	n = 12	n = 225

Bold entries are for cells that would represent classification agreement.
Simple Kappa = 0.31, Weighted Kappa = 0.43

Results indicate that the panelists’ classification agreed with the empirical classification for 123 out of the 225 score points (55%). The lowest level of agreement was 21% for Advanced and the highest was 76% for Basic. For Below Basic and Proficient, agreement was at 45% and 59%, respectively. These results were compared to results from a similar study conducted in 1998 to provide support for the civics cut scores (Loomis, 2000). In the civics study, there were 184 items and score points to be classified. Results from civics are provided in Table 50.

Table 50: Civics 1998 Item Classification Outcomes
Panelists' Judgments vs. Performance Level at RP .65, Civics Grade 12 ALS Cut Scores

		Median Panelist Classification				
		Below Basic	Basic	Proficient	Advanced	Total
Empirical Classification	Below Basic	**4%** **n = 1**	83% n = 20	13% n = 3		n = 24
	Basic	2% n = 1	**63%** **n = 41**	35% n = 23		n = 65
	Proficient		26% n = 18	**63%** **n = 44**	11% n = 8	n = 70
	Advanced		12% n = 3	72% n = 18	**16%** **n = 4**	n = 25
	Total	n = 2	n = 82	n = 88	n = 12	n = 184

Bold entries are for cells that would represent classification agreement.
Simple Kappa = 0.22, Weighted Kappa = 0.34

Results for the civics 1998 Item Classification Study indicate that the panelists' classification agreed with the empirical classification for 90 out of the 184 score points (49%). This is six percentage points lower than the results for economics. The lowest level of agreement was 4% for Below Basic and the highest was 63% for Basic and Proficient. For Advanced, agreement was at 16%.

Exemplar Item Ratings

Exemplar item ratings were gathered in the ALS meeting to provide the Governing Board 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 49 potential exemplar items, representing a total of 57 steps, or score points. There were 39 multiple choice items and 10 polytomously scored constructed response items representing 18 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 top of the level could correctly answer the item or attain the score point. For example, at the Proficient level, all items to be released that mapped to a value in between the Proficient and Advanced cut scores were selected as potential exemplars for the Proficient level (see Figure 48). 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.

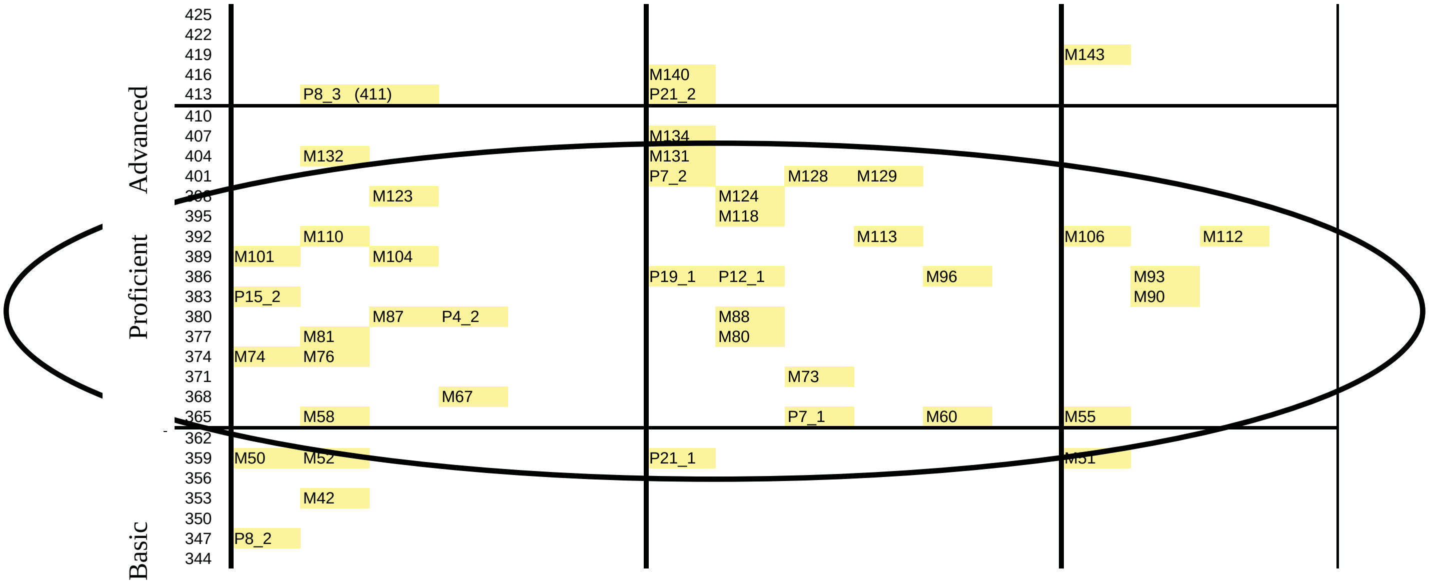


Figure 48. Exemplar items selected to represent the Proficient level.

The number of score points per achievement level overall and by item type is shown in Table 51.

Table 51: Number of Exemplar Score Points Mapped to Level Overall and by Item Type

Level	Multiple Choice	Polytomously Scored	Total
Basic	7	3	10
Proficient	28	6	34
Advanced	4	9	13
Total	39	18	57

For each item, panelists were asked to indicate if they felt the item was very good, OK, or should not be used to illustrate performance at the level with which it was associated. Detailed results of the exemplar item ratings are shown in Appendix P. ACT and our TACSS agree that if one-fifth, or 20%, of panelists checked the Do Not Use category, the item should not be recommended for use as an exemplar. The shaded cells in Appendix P flag items that were eliminated by this criterion. The number of remaining exemplars per achievement level overall and by item type are provided in Table 52.

Table 52: Number of Exemplar Score Points Meeting Rating Criteria and Mapped to Level Overall and by Item Type

Level	Multiple Choice	Polytomously Scored	Total
Basic	6	2	8
Proficient	23	5	28
Advanced	4	9	13
Total	33	16	57

ACT's suggested rating criteria (rated by fewer than 20% as *Do Not Use*) leaves a sufficient number of potential exemplar items of both multiple choice and constructed response type for the Governing Board to choose exemplar items from. Each achievement level was associated with at least two score points on partial credit (polytomously-scored) items that met ACT's suggested ratings criteria.

ALS panelists' responses to process evaluation questions concerning the exemplar items are shown in Table 53. These were questions 19 and 20 on the last process evaluation questionnaire. Mean ratings were positive. Each was above 4.00 on a scale of 1-5 with no individual ratings below partial agreement at 3. This indicates a high level of satisfaction with the items selected as potential exemplars for illustrating performance at each level.

Table 53: Responses of ALS Panelists to Questions about Exemplar Items

Question	5	4	3	2	1	Mean Score	SD
5-19. I believe the exemplar items will be useful for describing the achievement levels.	Totally Agree 20	10	Somewhat Agree 1	0	Totally Disagree 0	4.61	0.56
5-20. The exemplar items I reviewed seemed appropriately matched to their achievement level.	Totally Agree 15	10	Somewhat Agree 6	0	Totally Disagree 0	4.29	0.78

In addition to the ratings information provided by the ALS panelists, each exemplar item was also classified into achievement levels by the Special Studies panelists and was rated by the three content experts who had also compared the economics NAEP results with the TEL. ACT recommends that the Governing Board use the lists of items mapped to the achievement levels in the ALS meeting, the ALS panelist ratings of exemplars, Special Studies classifications, and content experts' ratings, along with other criteria of its choosing, to select exemplar items for the achievement levels. Along with the ALS panelist ratings, the number of content experts indicating that each item should not be used and the number of Special Studies panelists classifying each item into each achievement level are provided in Appendix P for this purpose.

SUMMARY AND CONCLUSIONS

For the purposes of helping the Governing Board set achievement levels for the 2006 NAEP in grade 12 economics, ACT:

- developed content domains for use in the Mapmark with domains standard setting method;
- conducted a Field Trial to develop a new standard setting method, Mapmark with whole booklet feedback, based on the bookmark method;
- conducted a Pilot Study in which cut scores for the 2006 NAEP in grade 12 economics were set using two Mapmark procedures, Mapmark with domains and Mapmark with whole booklet feedback;
- reviewed and compared the results of the two methods used in the Pilot Study with the TACSS;
- implemented Mapmark with whole booklet feedback for the operational ALS meeting on the recommendation of the TACSS and the request of COSDAM; and
- conducted two Special Studies to determine if an independent panel interpreted the Achievement Level Descriptions in a manner consistent with the interpretation of the ALS panelists.

Data from the ALS and Special Studies provide evidence of procedural validity, internal consistency, and reasonableness of the results. The ALS meeting received ratings on the panelist process evaluation questionnaires comparable to or higher than ratings from the previous standard setting meetings across all categories including clarity of instructions, panelist understanding of tasks, and panelist understanding of the meaning of performance at the lower borderline of each achievement level. ALS panelists also indicated that they had more time than they needed to complete their tasks, and did not feel rushed. In addition, panelist ratings of the efficacy of the method in yielding reasonable cut scores were slightly higher than for previous standard setting studies and the vast majority of panelists (30 out of 31) indicated they would sign a statement recommending the use of the resulting cuts. These results indicate that the quality of the ALS procedure used in economics equals or exceeds the quality of processes used to establish achievement levels for other NAEP content areas.

The Achievement Level Descriptions were also well received by the panelists. Panelist ratings of their understanding of the ALDs were high and increased across rounds. By the final round, they felt that their cut scores were highly consistent with the level of performance described in the Achievement Level Descriptions.

Internal consistency of the cut scores was strong. There were no significant differences between mean cut scores by rater groups, panelist type, races, genders, and geographic regions. In addition, there were no significant differences between mean cut scores by table at the Advanced and Proficient levels. The slight table group effect at the Basic level in the final round is due to a reduction in variance within the tables that heightens the amount of variance between.

The achievement level percentages and results from the Special Studies can help to inform the judgment as to the reasonableness of the ALS cut scores. The resulting achievement level percentages were approved by a large majority of the panelists. Those who did not approve the percentages recommended changes to be more consistent with their own individual cut scores, as expected. In addition, both the booklet classification and item classification Special Studies results provide support for the ALS results, which is stronger than the support for the 1998 civics results provided by similar studies in 2000. Finally, comparisons of economics NAEP achievement level percentages to the percentage of grade 12 students taking AP Economics exams and scoring above 3 provides some support for the percentage of students scoring at the Advanced level.

ACT's TACSS reviewed the ALS and Special Studies meetings processes and results and concluded as summarized above that the procedural validity was strong, the ALS cut scores were reliable, and that the panelists' reactions to the consequences data as well as the Special Studies results provide support for the achievement levels. Based on these results, ACT recommends the cut scores from round 3 of the ALS meeting. On the scale transformed by ACT for use in the ALS meetings, these are 326 for Basic, 363 for Proficient, and 411 for Advanced.

ACT also recommends that the Governing Board use the lists of items shown for each achievement level in Appendix P along with panelists' ratings of these items as exemplars, plus other information such as item content and difficulty, in selecting exemplar items for NAEP reports. It is further recommended that the Governing Board consider more strongly those items that received *Do Not Use* ratings from fewer than 20% of the panelists, and that the Governing Board consider classification data from the Special Studies and content experts ratings in their decision-making.

Based on these activities, ACT provided the Governing Board at their May 17-19, 2007, Board meeting with the following input regarding the three recognized outcomes of the Achievement Level Setting process:

- ACT endorses the Achievement Level Descriptions that were used in the operational ALS meeting.

- ACT recommends the cut scores from round 3 of the operational ALS meeting. These cut scores are currently not on the scale that will be used to report the 2006 assessment results.

- ACT recommends that the Governing Board use the lists of potential exemplar items from the ALS meeting in the process of selecting exemplar items. Ratings of these items by ALS panelists, three content experts, and the Special Studies panelists should be taken into consideration in selecting exemplar items.

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 Governing Board's Committee on Standards, Design, and Methodology.

A Board action on May 18, 2007 resulted in the Governing Board's unanimous adoption of the grade 12 economics Achievement Level Descriptions and cut scores. Exemplar items will be selected by COSDAM from the lists of potential exemplar items that emerged from the ALS meeting, as recommended by ACT.

RECOMMENDATIONS FOR FUTURE STANDARD SETTINGS

ACT has several recommendations for future standard setting meetings. They are for changes to the recruitment procedures and changes to the meeting method itself.

Changes to Recruiting Procedures

Recruitment for standard setting meetings has become increasingly difficult. For each NAEP standard setting, ACT has contacted a larger sample, and received smaller response rates. This may be due, in part, to an increase in standard setting across the county in response to No Child Left Behind. Potential panelists who, in the past, may have had few opportunities to participate in such activities may now be receiving invitations from their state and their schools, in addition to NAEP. In addition, members of the general public are typically not rewarded by their employer for participating in standard setting meetings and so are not willing to take their own personal time to volunteer. ACT recommends the following changes to the recruiting process to improve response rates:

- ACT recommends the Governing Board consider allowing contractors to recruit directly from the staff of relevant professional organizations. Staff members of relevant professional organizations (e.g., National Council on Economic Education) are often eager to participate and would have no difficulty in getting a release from their employers for the time necessary to set standards.

- ACT recommends a streamlining of the initial contact materials to potential nominators. In the past, a lengthy introductory letter, accompanied by nomination forms and explanations of requirements has been sent in a 9x12 envelope. For this project, ACT sent these contents to some potential nominators and also sent a short letter in a business envelope to others directing them to a website with more information. Response was much greater to the brief letter.

- ACT recommends a 9-month period prior to the ALS for recruiting purposes.

- ACT recommends payment of a $300-$500 stipend to all participants. This will offset some personal costs associated with taking any unpaid leave, and will make NAEP participation more attractive than other opportunities.

Changes to Meeting Procedures

Evaluations of the Mapmark with whole booklet feedback method were overwhelmingly positive. There were two areas about which some panelists expressed concern. ACT recommends the following changes to those areas:

- ACT recommends the adoption of the new, general overview Briefing Book as was used in the ALS, as opposed to the highly detailed book as used in the Pilot Study. A number of Pilot Study panelists indicated that the highly detailed Briefing Booklet was confusing and was not helpful as an advance material. The streamlined book, which provided a general overview, received no negative comments and seemed to provide panelists with a clearer sense of what to expect.

- ACT recommends a shortening of the orientation information at the beginning of the standard setting meeting. Several panelists commented in their evaluation forms that the amount of time spent on orientation was too much. Observational evidence also suggested that panelists tired from too much time spent listening and not enough spent engaging in activity. The new Briefing Booklet would allow for shortening the presentations to eliminate redundancy.

- ACT would not recommend any additional normative feedback presented to the panelists, such as consequences data for various demographic groups, AP results, or student performance in courses. This information would put too much emphasis on the normative data, and might distract panelists from the criterion-referenced nature of the task.

References

ACT, Inc. (2005). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade 12 mathematics: Process report.* Iowa City, IA: Author.

ACT, Inc. (2007). *Developing achievement levels on the 2006 National Assessment of Educational Progress in grade 12 economics: Special studies report.* Iowa City, IA: Author.

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

Efron, B. and Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *American Statistician, 37,* 36-48.

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. (2000). *Developing achievement levels on the 1998 National Assessment of Educational Progress in civics: Validation research.* Iowa City, IA: ACT

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

Maritz, J.S. and Jarrett, R.G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association, 73*, 194-196.

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 (2005). *NAGB policy.* (Appendix B in Attachment A (*Statement of Work*) to Solicitation No. ED-06-R-0013. *Twelfth Grade Economics Achievement Levels*). Washington, D.C.: Author.

National Assessment Governing Board (2006). *Economics framework for the 2006 national Assessment of Educational Progress* 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., Kolen, M. & Nicewander, W. A. (1999). A rationale for defining achievement levels using IRT-estimated domain scores. *Applied Psychological Measurement*, *23*, 347-362.

Appendix A

Item Map

Name: ______________________

Primary Item Map

Scale	Content Area															
	Market						National						International			
above	P27_2	P29_3					P26_3	P28_4					P25_2			
497							P24_2									
494																
491							P23_2						P22_4			
488																
485																
482																
479																
476							P21_3	P28_3								
473																
470													M154			
467																
464																
461	P20_2						P19_2									
458							P18_2									
455							M153						P17_2			
452							M152									
449																
446							M151									
443	P16_2															
440													P25_1			
437	P14_2	P15_3														
434	P13_2						M150						M149			
431							M148	P26_2								
428							P12_2	P28_2	M146	M147						
425							P11_2									
422																
419	M144						M145						M143			
416	P10_2						M140	M141	M142							
413	P9_2						P21_2									
410	P29_2	P8_3					M135	P18_1	M136	M137	M138	M139				
407							M134									
404	D3	M132					M131	P23_1					M133			
401	M126	M130					P7_2	M127	M128	M129						
398	M121	M122	M123	P6_2	M125		M120	M124								
395	M115	M116	M119				P24_1	M118					P22_3	M117		
392	M107	M110	M114				M105	M108	M111	M113			M106	M109	M112	
389	M101	M102	M104				M100	M103								
386	M95	P27_1	M98				P19_1	P12_1	M92	M94	M96		P17_1	M93	M97	M99
383	P15_2	M91					P5_3	D2					M89	M90		
380	M82	M86	M87	P4_2			M83	M88	P28_1				M84	M85		
377	M79	M81					M78	M80								
374	M74	M76	M77				M75									
371	M68	M69	M70				M71	M72	M73							
368	M64	M65	M66	M67	P3_2		M63	P26_1	P11_1							
365	M56	M58	M61				M57	P2_2	P7_1	M59	M60	M62	M55	P22_2		
362							M54									
359	M50	M52	P14_1				P21_1						M51	M53		
356	M44	M45	M46	M49			M47						M48			
353	M41	M42	M43	P9_1			P5_2									
350	M38	M40					P5_1	M39					M37			
347	P8_2	M35	M36				M34						M32	M33		
344	M27	M28	M29	M30			M26						M31			
341	M21	P6_1	P16_1	P10_1	M24	M25	M22	M23					P22_1			
338	M18	P20_1	P1_2	M19			M20									
335							M17									
332	M16						M15									
329	M13												M14			
326	P15_1	M9	D1	M11	M12		M10									
323	M8	P4_1														
320	M7	P3_1														
317	M6															
314	M5	P29_1														
311	P8_1	P13_1														
308																
305							P2_1									
302																
299																
296	M3						M4									
293	M2															
290	P1_1															
287																
284							M1									
281																

Appendix B

Achievement Level Descriptions

NAEP Achievement Levels Definitions for Grade 12 Economics

Basic

Students performing at the Basic level of achievement should be able to identify and recognize a limited set of economic concepts and relationships that are important for partial understanding of the market economy, national economy, and international economy. A limited set includes some of the following: (a) in the market economy -- scarcity, opportunity cost, incentives, marginal decision-making, markets, prices, demand, supply, competition, economic institutions, income determination, entrepreneurship, investment, and government actions; (b) in the national economy -- economic systems, money, interest rates, economic growth, gross domestic product, unemployment, inflation, fiscal policy, and monetary policy; and (c) in the international economy -- comparative advantage, the benefits and costs of trade, and exchange rates. An example of the level of understanding that students should be able to demonstrate at the Basic level is the ability to recognize the inverse relationship between the market price of a product and the amount buyers are willing and able to purchase.

Students should be able to use a limited set of these economic concepts and relationships in simple applications. For example, when given data or information about an economic event or situation, they should be able to identify a likely economic outcome. Students should be able to interpret data or information presented in simple charts, graphs, or tables, such as those showing changes in economic data over time.

Proficient

Students performing at the Proficient level of achievement should be able to identify and recognize a broader set of economic concepts and relationships that are important for solid understanding of the market economy, national economy, and international economy. A broader set includes many of the following: (a) in the market economy -- scarcity, opportunity cost, incentives, marginal decision-making, markets, prices, demand, supply, competition, economic institutions, income determination, entrepreneurship, investment, and government actions; (b) in the national economy -- economic systems, money, interest rates, economic growth, gross domestic product, unemployment, inflation, fiscal policy, and monetary policy; and (c) in the international economy -- comparative advantage, the benefits and costs of trade, and exchange rates. An example of the level of understanding that students should be able to demonstrate at the Proficient level is the ability to explain the role of shortages in causing market prices to change.

Students should be able to use a broader set of these economic concepts and relationships in more challenging applications that involve analyzing economic problems and decisions, and recommending policies and actions. Students should be able to interpret data or information presented in complex charts, graphs, or tables, such as those relating changes in one or more economic variables to changes in other economic variables, and to analyze economic data and information to describe events and trends.

Advanced

Students performing at the Advanced level of achievement should be able to identify and recognize an extensive set of economic concepts and relationships that are important for thorough understanding of the market economy, national economy, and international economy. An extensive set includes most of the following: (a) in the market economy -- scarcity, opportunity cost, incentives, marginal decision-making, markets, prices, demand, supply, competition, economic institutions, income determination, entrepreneurship, investment, and government actions; (b) in the national economy -- economic systems, money, interest rates, economic growth, gross domestic product, unemployment, inflation, fiscal policy, and monetary policy; and (c) in the international economy -- comparative advantage, the benefits and costs of trade, and exchange rates. An example of the level of understanding that students should be able to demonstrate at the Advanced level is the ability to identify factors that increase or decrease the demand for a product and to explain the effects of these changes on price and quantity.

Students should be able to use these economic concepts and relationships in complex applications that involve analysis and evaluation of economic data and information to explain events and their causes, and policies and their outcomes. Students should be able to use data or information presented in complex charts, graphs, or tables in their analysis and evaluation.

Appendix C

Members of the Technical Advisory Committee on Standard Setting

Technical Advisory Committee For Standard Setting (TACSS)

1. Stephen Buckles
   Professor of Economics
   Vanderbilt University
   Nashville, Tennessee

2. Barbara G. Dodd
   Professor
   Department of Educational Psychology
   University of Texas
   Austin, Texas

3. George Engelhard
   Professor of Educational Measurement and Policy
   Emory University
   Atlanta, Georgia

4. Robert A. Forsyth
   Professor Emeritus
   University of Iowa
   Iowa City, Iowa

5. Mary J. Pitoniak
   Lead Program Administrator
   Educational Testing Service
   Princeton, New Jersey

Appendix D

Domain Development Materials and Definitions

Benchmark Rating Instructions

1. Take some time to review how the benchmarks are defined in the framework. In the framework, there are 3 content areas, 20 standards, and 105 benchmarks. Collectively, the content areas, standards, and benchmarks represent the knowledge, skills, and abilities that represent economics content that students should know and be able to do. Benchmarks are the most specific unit of representation in the framework. Each benchmark in the framework is classified into just one standard and one content area.

2. Ideally, all of the benchmarks would be "mastered" by a student who is exposed to a complete curriculum in economics. By mastery of a benchmark, we mean that a student would be able to correctly answer questions on most of the content and skills represented by the benchmark. By a complete economics curriculum, we mean instruction covering all of the knowledge, skills, and abilities represented in the framework.

3. It is generally assumed that some knowledge and skills can be mastered through courses that expose students to relatively little of a complete economics curriculum and that, in a complete curriculum, some knowledge and skills will be mastered earlier than others. As you read each benchmark in the framework, please think about the content and skills associated with that benchmark and at what point in relation to mastery of other benchmarks in a complete economics curriculum, most of the content and skill associated with that benchmark is mastered. Please use the attached Benchmark Rating Scale and Benchmark Rating Form to record your responses for each benchmark.

4. It might be easier to rate some benchmarks than others because some benchmarks may refer more specifically to content, while others may represent a skill that is acquired and practiced over a more extended period of time and instruction. We also realize that virtually every benchmark represents a variety of content and skills covering a range of time and sequence in an economics curriculum. We ask that you choose a rating based on when you feel the majority of the content and skills represented by the benchmark would be mastered.

5. In some cases you may find that mastery of a benchmark does not appear to fit into a relationship with mastery of other benchmarks. Some benchmarks, for example, may be mastered at no particular point in a complete economics curriculum. If this is the case, please select the "Does Not Apply" option on your rating scale.

6. Before you begin, it might be useful to review all of the benchmarks and their organization in the framework. The benchmark ratings represent where mastery of a benchmark stands in relation to other content and skills in the framework. This relationship crosses boundaries of content areas and standards.

Benchmark Rating Scale	
Rating	Meaning
5	The knowledge, skills, and abilities associated with this benchmark are mastered after the vast majority of other benchmarks have been mastered and late in an instructional sequence in economics the goal of which is mastery of all the benchmarks in the NAEP Framework.
4	Mastery of the knowledge, skills, and abilities associated with this benchmark typically follows mastery of the majority of benchmarks that occur earlier in a sequence.
3	The knowledge, skills, and abilities associated with this benchmark are mastered about midway through an instructional sequence in economics.
2	Mastery of the knowledge, skills, and abilities associated with this benchmark typically follows mastery of some earlier benchmarks.
1	The knowledge, skills, and abilities associated with this benchmark are mastered very early in an instructional sequence in economics.
DOES NOT APPLY	The knowledge, skills, and abilities (KSAs) associated with this benchmark are mastered in an economics curriculum in no particular order in relation to other KSAs.

Item Rating Instructions

Please rate all of the items within a block before proceeding to another block. For each item:

1. Read the item, including all distractors (for multiple choice items) and think of the knowledge, skills and abilities (KSAs) that are required to answer the item correctly or to receive full credit. A scoring key and scoring rubrics are attached at the end of each block of items. You may find these helpful in identifying the KSAs needed to get full credit, especially on constructed response items.

2. If the item requires more than one knowledge, skill or ability for full credit, think about the most difficult or instructionally advanced knowledge, skill or ability needed to obtain full credit on the item.

3. Refer to the Item Rating Scale and select the category that best represents the most difficult or instructionally advanced knowledge, skill or ability required in order to obtain full credit on the item.

4. Record your Item Rating on the form provided.

A category, “does not apply” is available for you to use if you feel that the KSAs required by the item do not fall into any particular position relative to mastery of other KSAs in an economics curriculum whose goal is mastery of all the benchmarks in the framework. Do not treat this as a “do not know” category.

Item Rating Scale

Rating	Meaning
5	The knowledge, skills, and abilities (KSAs) required to get full credit on this item are usually mastered after the vast majority of other KSAs in an economics curriculum have been mastered, the goal of which is mastery of all the benchmarks in the NAEP Framework.
4	Mastery of the KSAs required to get full credit on this item typically follow mastery of the majority of other KSAs in economics.
3	The KSAs required to get full credit on this item are mastered about midway through mastery of all the KSAs in economics.
2	The KSAs required to get full credit on this item typically require mastery of some earlier KSAs in economics.
1	Most of the KSAs needed to get full credit on this item are mastered early in a learning sequence in economics.
DOES NOT APPLY	The KSAs required by this item are mastered in an economics curriculum in no particular timing in relation to other KSAs. They may occur early in some economics curricula and late in others.

Market

M1	**Entrepreneurs** Items in this domain require basic understanding of profit and what it is that entrepreneurs do. Profit, identified as the excess of revenues over costs, is identified as key to business success or failure and a key motivation for entrepreneurs.
M2	**Incentives** Items in this domain address the ways in which people respond to positive and negative incentives. They indicate how incentives cause people to change their behavior in predictable ways. Incentives may include changes in pay, prices, costs, interest rates, and taxes. Responses include changes in purchasing, in job choice or in amount of time spent on a job, and in use of services.
M3	**Markets and Equilibrium** Items in this domain require knowledge of basic concepts such as the nature and purpose of markets, the interaction of buyers and sellers, and equilibrium in a market. Items require students to recognize an equilibrium in a variety of contexts that include text, graphs, and tables.
M4	**Productivity, Income, and Capital** Items in this domain require understanding of how worker incomes and business productivity are affected by experience, education, skills, and job training. Concepts of human resources, human capital, and physical capital are involved. Students may be required to understand the role of these factors in personal and business investment decisions.
M5	**Scarcity and Opportunity Cost** Items in this domain require knowledge and application of the concepts of scarcity and opportunity cost. Items may appear in the context of individuals assessing tradeoffs between short and long term effects of economic choices where the concept of opportunity cost is needed.
M6	**Economic Institutions** Items in this domain address the nature and functions of economic institutions including banks, labor unions, and corporations. Questions may require the student to know the general and specific functions of these institutions such as making loans (banks) and collective bargaining (labor unions) and to understand how business organizations such as sole proprietorships and corporations differ.
M7	**Competition** Items in this domain require understanding of competition – its effects on prices or on a business's ability to control prices, and its effects on product innovation. Items may also address how lack of competition changes the effects on prices or innovation.
M8	**Economic Role of Government** Items in this domain require students to understand the reasons for, and effects of, certain governmental functions such as granting copyrights and providing goods and services that the private sector would not otherwise provide.
M9	**Interaction of Supply, Demand, and Prices** Items in this domain require students to recognize changes in supply or demand or both and may require them to predict the impact of the changes on price and quantity. Students may be asked to explain the law of demand or to predict how a change in demand for a product affects workers who make that product. Some items are presented in the context of government set price floors or price ceilings.
M10	**Additional Costs and Additional Benefits in Decision Making** Items in this domain require the student to understand, identify, or explain business, consumer, and personal choices in terms of the relative magnitude of costs versus benefits. Items may require students to compute and compare additional costs and additional benefits from information in text or tables. Some items require students to identify decisions that would maximize profit for the firm.

National

N1	**Money, Loans, and Interest Rates** Items in this domain require basic understanding of what interest rates and money are, and of key factors affecting loans and interest rates, risk (credit history), and demand for loans.
N2	**Economic Growth** Items in this domain focus on causes and effects of economic growth. The causes include expenditures on physical capital, technology, and training and education. The effects include increases in employment and productivity, standards of living, and decreases in poverty.
N3	**Resource Allocation** Items in this domain require students to know what constitutes resource allocation, how resources are allocated, and what the effects of specific resource allocations might be. Resource allocation includes deciding what goods and services to produce, how to produce them, and who will receive them. Items may require students to understand the role of government, differences among governments and differences among economic systems in resource allocation.
N4	**Government Programs and Taxes** Items in this domain focus on why government programs and services exist and how they are supported financially, through taxes, and politically, through the benefits they provide. Knowledge of different types of taxes, such as income tax, property tax, and progressive tax structures is needed.
N5	**Spending, Income, and Related National Measures** Items in this domain require knowledge of the relationship between total spending, income, standards of living, and of related national measures such as gross domestic product and per capita income.
N6	**Real Interest Rates** Items in this domain require students to know that the real interest rate is the difference between the interest rate and inflation (actual or expected) and they must be able to apply this concept to making real world choices such as whether or when to make a purchase or to invest or borrow money.
N7	**Inflation and Unemployment** Items in this domain require knowledge of inflation and unemployment (and employment), their relationship to each other and to other national measures such as consumer spending, price index, income, and gross domestic product. Students must be able to identify and describe the causes and effects of inflation and unemployment.
N8	**Money Supply** Items in this domain require knowledge of how the money supply is measured and of the relationships between the money supply and other economic factors such as inflation, interest rates, and bank loans.
N9	**Fiscal and Monetary Policy** Items in this domain require students to understand the effects of government fiscal and monetary policies on national debt, budget deficits, money supply, inflation, and unemployment. Fiscal policies include taxes, spending, and borrowing. Monetary policies include actions of the Federal Reserve.

International

I1	**Benefits and Costs of Trade** Items in this domain require understanding the benefits and costs of trade and restrictions in trade (or removing restrictions in trade), other than tariffs. Key concepts include opportunity cost, specialization, and comparative advantage. Items may address the effects of changes in global supply, and change in prices of a good or service on countries that import those goods or services.
I2	**Exchange Rates** The items in this domain focus on the determinants of exchange rates and the effects of changes in exchange rates on the decision-making of people and businesses.
I3	**Tariffs** Items in this domain require understanding of what a tariff is, why tariffs are used, and what effects tariffs have in terms of costs and benefits.

Appendix E

Field Trial Agenda

Agenda
NAEP Economics Field Trial

DAY 1

8:00 AM	Registration and breakfast
8:15 AM	Orientation to NAEP, the achievement level setting (ALS) process, and the mapmark procedure
9:15 AM	Examination
10:15 AM	Break
10:30 AM	**ROUND 1 MAPMARK BEGINS**
	Whole group KSA review (common constructed response items, CROIB)
11:40 AM	Table group KSA review (remaining constructed response items, CROIB)
12:30 PM	Lunch
1:15 PM	Independent KSA review (OIB)
2:45 PM	Break
3:00 PM	Table discussion KSA review (OIB)
4:20 PM	Evaluation #1 and Break
4:30 PM	ALD presentation
5:00 PM	Training in bookmark placement
5:30 PM	Bookmark placements (Proficient, then Basic, then Advanced)
6:00 PM	Evaluation #2 and Adjourn

DAY 2

8:00 AM	Breakfast
8:30 AM	**ROUND 2 MAPMARK BEGINS**
	Feedback from Round 1
8:45 AM	Borderline Proficient booklet review
10:00 AM	Table group student booklet review
12:00 PM	Lunch
1:00 PM	Group-level discussion of bookmark placements and whole booklet feedback
2:00 PM	Round 2 cut score recommendations
2:30 PM	Evaluation #3 & Break
3:00 PM	Feedback from Round 2
	Consequences data
	Evaluation #4
3:45 PM	Debriefing
4:30 PM	Adjournment

Appendix F

List of Panelists from ALS

2006 NAEP Economics Achievement Level Setting
Achievement Level Setting Panelists
March 7-10, 2007

Appendix G

ALS Agenda

AGENDA FOR THE 2006 GRADE 12 NAEP ECONOMICS
ACHIEVEMENT LEVEL SETTING MEETING
March 7-10, 2007
St. Louis, Missouri

Wednesday March 7

Promenade Foyer

8:00 AM Registration and Continental Breakfast

Promenade B

8:30 AM Welcome and Introductions, *Christina Hamme Peterson, ACT*

8:40 AM General Orientation to the NAEP, *Susan Loomis, The Governing Board*

9:10 AM General Introduction to NAEP Achievement Level Setting Process

9:40 AM Break

10:00 AM Panelists' Introductions and NAEP Exam

Promenade C

11:30 AM LUNCH

Promenade B

12:30 PM Orientation to the Method

1:30 PM The NAEP Economics Framework, *R.J. Goodman*

2:15 PM Break

2:30 PM Whole Group KSA Review of Common Extended Constructed Response Items,

4:00 PM Table KSA Review of Remaining Extended Constructed Response Items

5:00 PM Evaluation #1

Cupples Salon C

6:00 PM Get-Acquainted Social Time

6:30 PM DINNER

Thursday, March 8

Promenade Foyer

7:30 AM Continental Breakfast

Promenade B
8:00 AM Continue Table KSA Review of Remaining Extended Constructed Response Items

9:15 AM Independent Review of Ordered Item Book

Promenade C
11:45 AM LUNCH

Promenade B
12:45 PM Table Discussion of Ordered Item Book

2:45 PM Break

3:00 PM ALD Presentation, *R.J. Goodman*

3:45 PM Round 1 Bookmark Training

4:15 PM Round 1 Bookmarks

Evaluation #2

5:30 PM Adjourn

Friday, March 9

8:00 AM Continental Breakfast

Promenade B

8:30 AM Feedback from Round 1

- Cut Scores
- Rater Locations
- Whole Booklets
  - Background
  - Booklet Score Chart
  - Booklet Score Plot

9:45 AM Borderline Proficient booklet review

- Background

Borderline Exercise

- Item Score Table
- Booklet Item Map

Discussion

10:30 AM Break

10:45 AM Independent Student Booklet Review

Promenade C

12:00 PM LUNCH

Promenade B

1:00 PM Whole Group Discussion

1:30 PM Round 2 Cut Score Recommendations
Evaluation #3

2:00 PM Break

3:15 PM Feedback from Round 2

- Cut Scores
- Rater Locations
- Booklet Feedback
- Consequences Data

Whole group discussion

4:00 PM Round 3 Cut Score Recommendations
Evaluation #4

4:30 PM Adjourn

Saturday, March 10

8:00 AM Continental Breakfast

Promenade B

8:30 AM Feedback from Round 3
- Cut Scores
- Rater Locations
- Consequences Data

Consequences Questionnaire

9:15 AM Exemplar Item Ratings

10:30 AM Evaluation #5

10:45 AM Break

11:00 AM Debriefing

12:00 PM Adjourn

Appendix H

ALS Briefing Booklet

Briefing Booklet

2006 Grade 12 Economics NAEP Achievement Level Setting March 7-10, 2007

A Note from the Director

*Congratulations on your selection from a national sample of potential panelists for participation in the grade 12 economics **National Assessment of Educational Progress (NAEP)** Achievement Level Setting! Together with other exceptional teachers, educators, and economics professionals, you will work to determine the student scores on the National Assessment that correspond to various levels of achievement. This work will serve as the basis for national standards which will be used by educators and policymakers at the local, state, and national levels to evaluate student progress and help guide curriculum and instruction.*

*This booklet is designed to provide you with some background on the National Assessment, an overview of the activities and tasks you will engage in during the course of our four day meeting, and an explanation of how the outcomes from that meeting will be used by the **National Assessment Governing Board**. Throughout the booklet you will see terms in bold print. These terms might be unfamiliar or might be used in ways unique to the **Achievement Level Setting (ALS) process**. For this reason, a glossary of terms and list of acronyms are provided in the back of the booklet to help you become familiar with language that will be used throughout the meeting. Some terms in the glossary do not appear in this briefing booklet, but will be discussed in depth at the meeting itself, so I recommend that you bring this booklet along with your other materials as a reference in March.*

It is my hope that you will find this information helpful in preparing you for the achievement level setting meeting in March. I look forward to meeting you, and to working together to determine economics national performance standards.

Yours truly,

Christina Hamme Peterson, Psy.D.

2006 Grade 12 Economics Briefing Booklet

Background on the National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP), also known as "the Nation's Report Card," is the only nationally representative and continuing assessment of what America's students know and can do in various subject areas, including reading, mathematics, science, history, civics, geography, and now, economics. NAEP results are reported not for individual students or schools; but instead for populations of students (e.g., fourth-graders) and groups within those populations (e.g., female students, Hispanic students). For each group, the proportion of students scoring at Below Basic, Basic, Proficient, and Advanced achievement levels are reported, as are the amount and type of instruction these students receive in the subject area, the type of exposure they have had to subject-related content outside of the classroom, and the level of education of their teachers (http://nces.ed.gov/nationsreportcard).

Development of the NAEP is typically a five year process.

- Year 1: Creation of a **framework.** The framework outlines topics and aspects of the subject area that are to be included in the assessment and specifies the relative weight of each topic.
- Year 2: **Item** development. A pool of items is developed. Each item is designed to measure an aspect outlined in the framework and, together, the pool of items measures all aspects specified in the framework.
- Year 3: Item review and field testing. Items are reviewed for quality and are administered in field tests to determine their clarity and performance.
- Year 4: Establishment of **achievement levels**. The NAEP is administered to a nationally representative sample of students at the appropriate grade level. **Achievement Level Descriptions** of what students should know and be able to do in order to be considered performing at the Basic, Proficient, or Advanced levels are developed. A panel of exceptional teachers, educators, and professionals in the subject area identifies **cut scores** on the test that correspond to these achievement levels and recommends **exemplar items** to exemplify performance at each level. You are a member of this panel.
- Year 5: Reporting of results. The proportion of students scoring at each of the three achievement levels (Basic, Proficient, and Advanced) and scoring Below Basic is reported. Also reported is the proportion of students receiving varying levels of instruction in the content area.

The National Assessment Governing Board, created by Congress in 1988 and appointed by the Secretary of Education, sets policy for the National Assessment and oversees all aspects of NAEP, including the development of the test framework and items, administration to a nationally representative sample of students, development of achievement level descriptions and of the cut scores on the test corresponding to those descriptions, and reporting of results. The Governing Board is a 26-member bipartisan group which includes governors, state legislators, local and state school officials, educators, business representatives, and members of the general public.

In economics, the NAEP was administered in the spring of 2006 to a sample of almost 12,000 grade 12 students. Your task in this meeting will be to engage in a series of exercises and activities to help you and the panelists with you identify the scores on the test that correspond to each achievement level and to recommend items that exemplify performance at each of the three levels. The results of this work will allow the Governing Board to determine what proportion of students who took the test in the spring fall into each achievement level.

Activities and Tasks for the Achievement Level Setting Meeting

At the conclusion of this meeting, you will have identified cut scores and exemplar items. Cut scores are scores on the test corresponding to the lower borderline of each achievement level. These establish the minimum standard of performance on the assessment required to be performing at “Basic,” “Proficient,” or “Advanced” levels. Exemplar items are items from the assessment that you feel best illustrate what students should know and be able to do in order to demonstrate Basic, Proficient, or Advanced performance. In order to help you determine the cut scores and identify exemplar items, the achievement level setting staff will walk you through three rounds of cut scores and a series of activities to inform your decision making. A description of these rounds and their associated activities follows.

Round 1

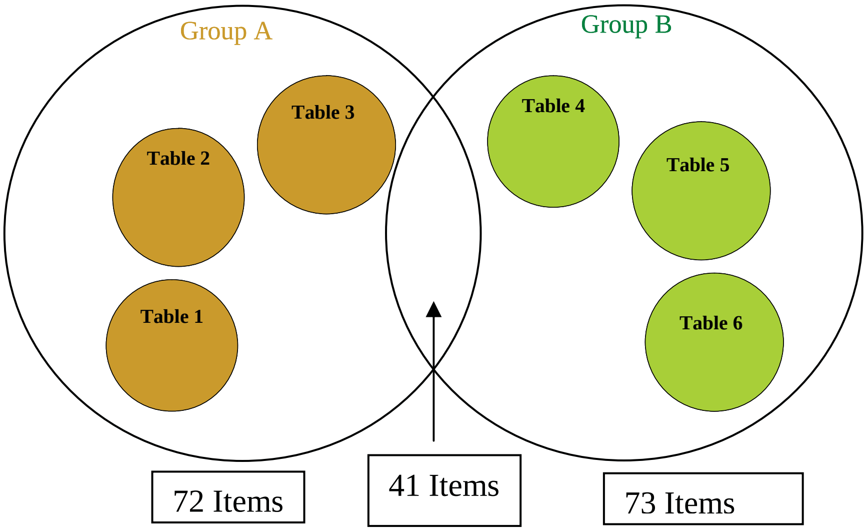
General Orientation and NAEP Exam

The meeting will begin with introductions to the achievement level setting (ALS) staff, an overview of the achievement level setting process (also called “standard setting”), a description of how panelists were selected for participation, and an orientation to the National Assessment of Educational Progress and the history of the National Assessment Governing Board. Following this orientation, the panelists will introduce themselves and the organization of the meeting room will be explained to you.

For the duration of the meeting, you will be assigned a seat at a table along with four or five other panelists. A large amount of discussion and work in this meeting occurs at the table level. The wide-range of experience and unique perspectives represented at each table are invaluable in the standard setting process. Though you should not uncritically accept the opinions and judgments of any other panelist, it is also important to listen to other panelists and take their experiences and perspectives into account in reaching your own judgments and decisions. A strong commitment to mutual respect will help you find the right balance between representing your own background and perspective and learning from other panelists during the standard setting process.

In addition to table groupings, you will belong to one of two **rater groups—**Group A or Group B (see Figure 1). Each rater group looks at a slightly different set of items, called an “**item pool**.” The two item pools have 41 items in common, and 72 or 73 unique items, for a total of about 113 items per rater group. The dividing of panelists into rater groups is done partly because there are a total of 186 items on the test—too many for any one panelist to work with. Panelists at the same table belong to the same rater group and so will be looking at the same items.

Figure 1: Division of Panelists and Items by Table and Rater Group



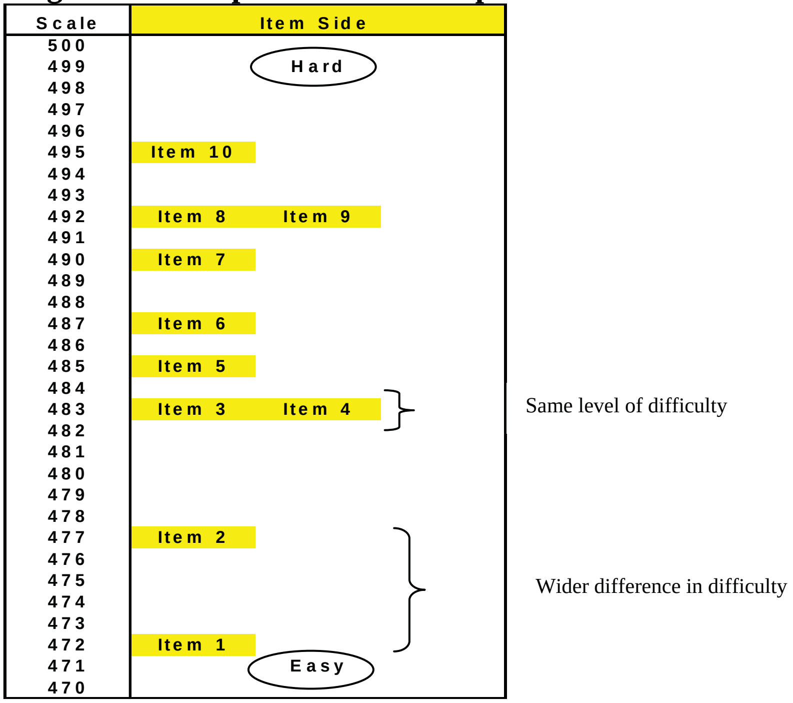
After introductions are made, you will take a form of the 2006 economics NAEP exam for grade 12 under similar conditions as those during the student administration. The purpose of this activity is not to test your economics ability—no score will be reported—but rather to give you an opportunity to experience the exam as the student experiences it. Test conditions such as timing and instructions are important factors to consider in your standard setting work. Also, by taking the exam and scoring your own responses, you will start to become familiar with NAEP test items and their scoring guides.

Orientation to the ALS Method and the Framework
In this session, the standard setting staff will explain general features of the particular method we will be using to establish cut scores on the assessment. This method, called ***Mapmark** with whole booklets,* is a variation of a **Bookmark** procedure which may be familiar to some of you as it is regularly used at the state level. We will describe the judgments you will be making in the process of recommending cut scores, and how we will prepare you to make these judgments. We will also train you in how to understand and use your Mapmark materials.

In this method, you will use an **Ordered Item Book (OIB)** and an **Item Map**. The Ordered Item Book, or OIB, is comprised of all of the items in the panelist's **item pool** ordered from the easiest item to the hardest. Panelists set each of their cut scores by dividing the items in the OIB into two groups—those that they feel are easy enough for a minimally qualified student in the achievement level to have mastered and those too difficult for this expectation.

One of the things that makes Mapmark different from Bookmark is the **Item Map** (Figure 2). The item map also has all of the items on the assessment ordered from easiest to hardest, but on the item map, the difficulty of an item is mapped to an actual scale value. A student scoring at that scale value has a 67% chance of getting that item correct. The item map, therefore, shows "how much" more difficult one item is than another and "how hard" items are for students at different levels of achievement.

Figure 2: Example of an Item Map



Same level of difficulty

Wider difference in difficulty

Following the orientation to the method, the ALS content facilitator will provide you with an orientation to the framework. The content facilitator is a member of the national consensus panel that developed the framework. As a panelist, understanding the test framework is the first step toward reaching a useful understanding of what students should know and be able to do in economics at each achievement level and the ALS staff highly recommends that you read the green framework document provided in your advanced materials prior to the meeting. During the presentation, the topics and aspects of the assessment determined by the framework will be reviewed and discussed.

KSA Review

Following all of the orientation training, you are ready to begin to familiarize yourself with the items in your item pool. This comes in the form of a KSA review which will take place for the remainder of the first day and most of the second day of the meeting. During this review, you will go through your Ordered Item Book (OIB), one item at a time, starting on the first page which is the easiest item in your item pool. For each item, you will think about the Knowledge, Skills, and Abilities (KSAs) needed to correctly answer the item. Because the OIB is in order of increasing difficulty, you will want to consider what a student should know and be able to do *above and beyond* what is necessary to answer an item that appears earlier in the book (an easier item) of similar content. This will allow you to become familiar with the progression of difficulty from one item to the next.

As you work through your OIB, you will locate the items on your Item Map, and check them off. Your checks will give you a visual picture of where items are on the map, how much the items differ in difficulty, and where you are in the range of student achievement represented by all of the items in the assessment.

The KSA review involves multiple stages, including review of constructed response items alone and with multiple choice items, review with the whole group and with just your table group, and independent review. Each stage is designed to help you and the other panelists gain a clearer sense and shared ideas of what the assessment is measuring.

Achievement Level Descriptions (ALDs)
After identifying and discussing the knowledge, skills, and abilities required by test items, the next task is to consider the KSAs students should have in order to be minimally qualified for an achievement level. There are three achievement levels—Basic, Proficient, and Advanced—with a separate description for each level. Each description is meant to communicate to educators and the general public what students at that level should know and be able to do in economics. Your content facilitator will give a presentation about the intent, development, and content of the Achievement Level Descriptions. The presentation will include a discussion of how the ALDs relate to the framework and to the item KSAs that you identified in previous exercises. This session is also intended to help you become comfortable and conversant with the ALDs.

Placing Your Round 1 Cut Scores
After you are familiar with the Achievement Level Descriptions and their relationship to the KSAs you will make your Round 1 cut score recommendations. Your Round 1 judgment is to place, for each achievement level starting with Proficient, then Basic, then Advanced, a bookmark that divides the items in your Ordered Item Book into two groups—items that students at the lower borderline of the achievement level should have mastery of and items that are too difficult for this expectation. You will place your bookmark on the last, or most difficult, item in the first group.

The standard setting task has been defined by the Governing Board as essentially one of translating the ALDs into cut scores. Your concept of knowledge, skills, and abilities that students at the lower borderline of an achievement level should have must be based on the ALDs. The ALD describes what the *typical* student in the level should know and be able to do. Items below your bookmark describe what students performing at the lower *borderline*, or the *minimally qualified* students, know and can do. Students at the borderline of the achievement level are not typical of all students in the level, but they should be qualified to be in the level.

Your Round 1 bookmark placements will focus on one achievement level at a time, beginning with Proficient. You will be instructed in how to place and record your bookmarks. You should place your bookmarks *independently* and should not look to see where other panelists at your table are placing their bookmarks. When placing the bookmark for borderline Proficient, you will work only with the Proficient ALD. Basic and Advanced bookmarks will be placed in the same way in that order. After all three bookmarks are placed, you will have a chance to review them as a group and make final adjustments.

At the end of this process, you will have placed each bookmark—Basic, Proficient, and Advanced—on different pages of your OIB. Each bookmark page identifies the last, or most difficult, item that you feel should be mastered by students performing at the lower borderline of the given achievement level. You will be asked to record the page numbers of your bookmarks on the Cut Score Recommendation Form.

Round 2

Feedback from the Preceding Round

Round 2 begins with **feedback**, a presentation on the results from the previous round. Each panelist's bookmark placement will have been translated to a recommended cut score on the item map scale. For each achievement level, the median cut score over all panelists is determined. The median cut scores (one for each achievement level) are reported as the "Round 1 Cut Scores." The distribution of all panelists' Round 1 cut scores across all three achievement levels will also be presented.

.

You will then be given instructions on how to record the Round 1 cut scores on your item map and in your OIB. The cut score locations give you an overall, visual picture of how the achievement level boundaries are located with respect to each other and to the items on the map and in the OIB. You will be asked to take notice of where your recommended cut scores are located in comparison to the Round 1 cut scores. This will allow you to see whether your recommended cut scores are higher or lower than the Round 1 cut scores. You will not be able to evaluate this information fully without understanding what students at the Round 1 cut scores "can" do and considering whether this is what students "should" be able to do according to the Achievement Level Descriptions. The **booklets** and Item Maps, described below, will help you understand what students at the Round 1 cut scores can do. They will also help you understand what students at your own recommended cut scores can do. Then, if you think your cut scores are too low or too high, you can raise or lower them accordingly in your Round 2 cut score recommendations.

Booklet Review

In addition to the feedback, you will also be provided with 20 examples of student performance. These examples will be in the form of test booklets actually completed by students in spring of 2006. You will have four booklets which scored at the Round 1 cut score and two which scored in the middle of each achievement level and below basic. You will review and discuss the student performance exhibited in these booklets and will consider: 1) if the booklets represent borderline or solid performance given the Achievement Level Descriptions and, 2) where your cut scores fall in relation to the score on each booklet.

Placing your Round 2 Cut Scores

Your review of student booklets will give you a good picture of student performance at the Round 1 cut scores. In establishing your Round 2 cut scores, your task is to determine if you feel performance at the Round 1 cut scores is too high, too low, or just right for the achievement level. Your response to this question will help you to determine if your recommended Round 1 cut scores need to be shifted relative to the median or can remain where they were. Should you opt to shift your cut scores, you will use your Ordered Item Book and Item Map to provide information about the knowledge, skills, and abilities a student at the borderline would have.

Your Round 2 cut score recommendations should be made *independently* and you should not discuss them with any other panelist. You will recommend a cut score for the Proficient level first, then the Basic, and last the Advanced.

Round 3

Feedback from the Preceding Round

Just as at the beginning of Round 2, Round 3 begins with a presentation on the results from the previous round. For each achievement level, the median cut score over all panelists has been determined and these median cut scores (one for each achievement level) are reported as the "Round 2 Cut Scores." The distribution of all panelists' Round 2 cut scores across all three achievement levels will also be presented.

You will be asked to take notice of where your recommended Round 2 cut scores are located in comparison to the Round 2 cut scores. This will allow you to see whether your recommended cut scores are higher or lower than the Round 2 cut scores. You will want to keep the location of your cut score in mind as you review the data provided in Round 3.

Consequences Data and Whole Group Discussion

It is in Round 3 that you will be given information about how student performance on the spring administration of the NAEP is distributed with respect to the Round 2 median cut scores. The percentage of students performing on the Grade 12 economics NAEP at or above the cut scores set for each achievement level will be reported for your consideration and evaluation. This is called **consequences data**.

There will be a whole group discussion centering on whether the consequences data seem reasonable to you in light of the Achievement Levels Descriptions (what students *should know and be able to do*) and in light of what you know about student performance in economics (*what students know and can do*). Having seen these data, do you want to adjust your cut scores? Did students generally perform better or worse than you expected? The consequences data serve largely as a "reality check" for the cut scores. Whatever your reaction to the consequences data, you should keep in mind that it is the Achievement Level Descriptions that take precedence.

Placing Your Final Cut Scores

The process of making Round 3 cut score recommendations is similar to that of Round 2 in that you will choose a scale value as your recommended cut score for each achievement level. Whether your choice is higher or lower than the Round 2 cut score will depend on how your own recommended Round 2 cut score differed from the median (the reported Round 2 cut score) and on your reaction to the consequences data and whole group discussion. We will review the information and materials available for you to use in making your cut score recommendations, the touchstones you will use in considering cut scores, and how you will record your recommended cut scores.

As always, your Round 3 cut score recommendations should be made *independently* and you should not discuss them with any other panelist. You will recommend a cut score for the Proficient level first, then the Basic, and last the Advanced.

Exemplar Item Ratings

It is only after the Mapmark rounds are complete that you will make recommendations for exemplar items. Exemplar items are one of the primary outcomes of the ALS process and are used for reporting student performance on the NAEP relative to the achievement levels. Exemplars are items that illustrate knowledge and skills associated with each achievement level.

Potential exemplar items will be drawn from test **blocks** that the Governing Board plans to release to the public. An item or score level (on a **constructed response item**) from this pool will be identified as a potential exemplar for an achievement level if a student falling at the top of that achievement level has at least a 67% chance of answering that item correctly.

In this task, you will rate the potential exemplar items on whether you feel they should be used to illustrate what students in a particular achievement level should know and be able to do. You will indicate whether you feel the item *should definitely be used, is OK to use,* or *should not be used* as an exemplar for the achievement level. You may discuss potential exemplars with other panelists at your table, but your ratings do not have to be the same.

Process Evaluations

You will be asked to complete evaluation forms after each major activity or phase of the standard setting process. You will also be asked to complete an evaluation form for the process as a whole. The evaluation forms include many statements that you will respond to using a rating scale such as strongly agree to strongly disagree. In addition, you will be asked to provide written responses to more general, open-ended questions and will be given space to comment on any aspect of the ALS process you feel would be helpful to the Governing Board for evaluating the process.

It is very important that you respond to these evaluations carefully and thoughtfully. We will study the evaluations at the end of each day to see if panelists are experiencing any difficulties with performing the tasks. We will also analyze and report the evaluation data in conjunction with the cut score data. The evaluation data are an important source of validity evidence for the cut scores and will also help us improve the process of setting achievement levels.

Glossary of Terms[1]

Achievement Levels

Also known as “standards” or “performance standards.” Three achievement levels will be set for reporting student performance on the NAEP: Basic, Proficient, and Advanced. *Basic* denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade. *Proficient* represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter. *Advanced* signifies superior performance beyond proficient. These achievement levels are represented by Achievement Level Descriptions (see below), cut scores on the test, and exemplar items which exemplify performance at each level.

Achievement Level Descriptions (ALDs)

Statements describing what students should know and be able to do in a specific subject and grade. The achievement level descriptions contain the essential aspects of the assessment framework appropriate to student performance at each level and grade.

Achievement Level Setting (ALS) Process

A judgmental process involving broadly representative panels of educators and noneducators. The process includes developing an understanding of the achievement level descriptions, selecting cut scores to represent what students at the borderline of each level should know and be able to do, and selecting exemplar items to represent what students should know and be able to do at each achievement level.

ACT

ACT, Inc., contractor of the National Assessment Governing Board. Responsible for designing and conducting the ALS process.

Block

A group or set of items forming one section of the NAEP exam. Blocks are timed to allow 25 minutes for students to answer the items. Each block contains approximately 18 items. A NAEP exam is comprised of two blocks.

Booklet

The test form or instrument. A booklet is composed of two blocks and contains approximately 36 test items. The NAEP is administered to a student as one booklet. There are many combinations of blocks to produce 50 different book forms of the economics assessment.

[1] Some terms in this glossary are defined for this specific context. These terms might be used differently outside the context of this achievement level setting process. There is a lot of jargon in this process, and we hope that this glossary will help you become familiar with important terminology.

Booklet Score Chart

A chart that shows the expected total number of points on two test forms as a function of the achievement scale score as well as the location of the 20 booklets panelists will review in relation to the achievement scale. This chart is used to provide cut score and rater location feedback, and to select new cut scores in Round 2.

Booklet Score Plot

A graph showing all possible percent correct scores on a test form. The Round 1 median cut scores are marked on this chart, on all ten student booklets, two at each cut score and one in the middle of each achievement level.

Bookmark

The standard setting method on which mapmark is based. The bookmark method involves a Knowledge, Skills, and Abilities (KSA) review of items in an Ordered Item Book (OIB), and placing a bookmark in the OIB to indicate the cut score for an achievement level. Mapmark extends the bookmark method with the use of spatially-representative item maps, marking the cut score on the map, and using examples of student performance on actual booklets.

Borderline Performance

The level of performance that is minimally acceptable for each achievement level. In other words, the level of performance that just meets the criteria for each level.

Common Items

These are the items that are used by both rater groups. They comprise about one-third of each rater group's item pool.

Cognitive Categories

The type of thinking that an item requires—what it asks the student to do. NAEP economics items are classified into three cognitive categories: knowing, applying, and reasoning. Knowing items rely heavily on recall and recognition. Applying items involve description and explanation of the relationship between info (data, summaries, hardliner, problems, etc.) and economics concepts. Reasoning items require problem solving, evaluation, interpretation, and analysis.

Consequences Data

Information about how student performance is distributed with respect to the cut scores; The percentage of students at or above each cut score, and the percentage of students in each achievement level.

Constructed Response Item

An item that requires examinees to construct or supply a written response. Any item that is not in a multiple-choice format.

Extended Constructed Response Item: Any constructed response item that has more than two score levels. Also called a polytomous item or a polytomously scored item.

Short Constructed Response Item: A constructed response item that has just two score levels: right (=1) or wrong (=0). Also called a dichotomous item.

Constructed Response Ordered Item Book (CR-OIB)

A book of the constructed response items in your item pool including both the common items and the items unique to your group. The CR-OIB includes scoring rubrics and sample student responses for every possible score.

Content Area

Areas of knowledge and skills taught in a high school economics curriculum. The NAEP economics Framework has three content areas: Market Economy, National Economy, and International Economy (see **subscale**). Items are categorized into content areas. Approximately 45% of the assessment covers knowledge and skills in the Market Economy; 40% in the National Economy; and 15% in the International Economy.

Cut Scores

Scale scores that define the borderline of the achievement levels. Cut scores are established through the standard setting process.

Cutpoints (See **cut scores**.)

Educational Testing Service (ETS)

Educational Testing Service (contractor of NCES) responsible for developing the assessment questions according to specifications provided to NCES by The Governing Board, analyzing results, and working with NCES staff to prepare The Nation's Report Card and other reports on student achievement in various subject areas assessed by NAEP.

Exemplar Items

Items that illustrate knowledge and skills associated with each achievement level. Exemplar items are selected by panelists to use in reporting the NAEP results. They are a primary outcome of the ALS process.

Extended Constructed Response Items (ECR)

Constructed response items that have more than two score levels. Also called polytomous items, polytomously scored items, or partial credit items. These items can have up to five score levels (0, 1, …, 4), but most have just three (0, 1, and 2). These items typically require the student to supply a more lengthy or detailed response than do "short" constructed response items.

Feedback

Information provided to panelists between rounds, based on the current set of cut scores. Most of the information is based on the cut scores recommended by panelists during the previous round. Feedback data are presented to the panelists for their consideration and to inform their judgments about cut scores in subsequent rounds.

Framework

The framework defines the aspects of the subject area (i.e., economics) that are to be assessed. It specifies the relative emphasis in measuring each content area and topic within the subject. The framework is the foundation for the assessment and for the ALS process. A national consensus panel developed the framework.

Item

A question or cluster of questions that is a single score unit in the assessment.

Item Handle

A unique identifier for each item or score level (of an extended constructed response item) on the item map. The handle gives information on the item type and scoring (P=polytomously scored, constructed response; D=dichotomously scored, constructed response; M=multiple choice), the difficulty order within type (e.g., M1=easiest multiple choice item; D1=easiest dichotomously scored item), and the score level of polytomously scored items (P1 = a score of 1 for item P1, P1-2 = a score of 2 for item P1, etc.).

Item Label

A box appearing on each page of the ordered item book containing information about the item.

Item Map

A spatially-representative display of items by difficulty and content. The vertical distance between items represents their difference in difficulty. Items are organized into columns representing the three content areas (on the Primary Item Map) and Domains (on the Domain Item Maps). The vertical order of items on the Primary Item Map corresponds to their order in the Ordered Item Book. Items are represented on an item map by a handle (see Item Handle). Item handles are color coded: Group A only items are tan, Group B only items are green, common items (Group A and Group B) are yellow.

Primary Item Map: This map represents all of the items in the 2006 assessment—items in both rater group item pools. The columns on this map correspond to the content areas of the assessment (see **content area**). This map will be used and updated each round with cut score and rater location feedback.

Booklet Item Map: This map represents all of the items in the 2006 assessment. It is identical to the primary item map except that the items specific to a given test booklet are highlighted in blue.

Item Pool

The 186 items in the 2006 NAEP Grade 12 economics assessment have been divided into two pools of items called item pools. Each item pool contains approximately 113 items. The pools have 36 items in common. The item pools are approximately equal with respect to item difficulty, item formats, and other item characteristics. Each panelist works with just one item pool, so there are two "rater groups" of panelists, called Group A and Group B.

Item Score Table

This table represents item scores on one test form for students scoring at different points on the achievement scale. The rows represent all of the items in a test form ordered from easiest to hardest. The columns represent 10 different student booklets, ordered from below basic to above advanced. For each item, a "0" or "1" is indicated, illustrating the number of points a student received on that item.

Mapmark

An extension of the **Bookmark** (see glossary) standard setting method, incorporating spatially-representative item maps.

Map Value

The achievement scale on the item map is divided into intervals 3 points wide. Each interval is identified by its midpoint. An item's map value is the midpoint of the interval that contains the item's scale value (see **scale value**). The item can be found on the item map in the row that shows the item's map value in the achievement scale column.

NAEP (See **National Assessment of Educational Progress**.)

National Assessment of Educational Progress (NAEP)

The National Assessment of Educational Progress or the "Nation's Report Card" is the only national representative and continuing assessment of what America's students know and can do in various subject areas. The first-ever economics NAEP was administered to U.S. students in grade 12 from January to March 2006.

National Assessment Governing Board (NAGB) (See **The Governing Board**.)

Created by Congress to formulate policy guidelines for NAEP. Board membership is broadly representative including K-12 classroom teachers, measurement experts, governors, legislators, and interested citizens.

National Center for Education Statistics (NCES)

An agency of the U.S. Department of Education responsible for reporting education statistics including NAEP results.

Ordered Item Book (OIB)

This booklet contains all of the items in your item pool, ordered by difficulty from the easiest to the hardest. Item difficulty is based on actual student performance data. Scoring rubrics for constructed response items are included. Constructed response items appear multiple times, once for each score level.

Constructed Response Ordered Item Booklet (CR-OIB): See separate definition of this term.

Panelists

Teachers, nonteacher educators, and members of the general public selected to participate in the achievement level setting process.

Partial Credit Item (See **Extended Constructed Response Item** or **Polytomous Item**.)

Pearson Education

Contractor responsible for printing and scoring the NAEP exam.

Polytomous Item

A constructed response item that has more than two score levels (e.g., 0,1, and 2). Also called extended constructed response item or partial credit item.

Primary Item Map (See **item map**.)

Rater Group

Panelists are divided into two rater groups. The groups are approximately equal in terms of panelist type and demographic characteristics. Each group sees different but overlapping sets of items, called item pools.

Response Probability

The probability that determines an item's scale value and map value (see **scale value** and m**ap value**) and defines "mastery" in the mapmark standard setting process. The response probability is 0.67 in this ALS process or a 67% chance of answering an item correctly.

Scale Value

Items are mapped to a certain value on the score scale using a response probability. In this ALS process, an item’s scale value is the score at which a student has a 0.67 probability of correctly answering the item.

Scoring Rubric

A list of correct responses, acceptable variations, and their corresponding scores for a given item. It also includes the rationale for scoring each item and explanations for acceptable answers for each score level. Multiple choice answers are located in the lower right hand corner of the item label.

Sequence

The order in which items appear within a block.

Short Constructed Response Item

A constructed response item that has just two score levels (right=1 or wrong=0). Also called a dichotomous item.

Subscale

A unit of psychometric analysis corresponding to each content area. The economics assessment content areas are Market Economy, National Economy, and International Economy. Student performance on the test items is reported for the test as a whole and for each of the three subscales.

The Governing Board (See **National Assessment Governing Board**.)

Westat

Contractor responsible for the sampling and test administration for the NAEP.

Whole Group

Composed of all the panelists in both rater groups for a given ALS method.

Acronym List

ALD – Achievement Level Description

ALS – Achievement Level Setting

BSC – Booklet Score Chart

CR – Constructed Response

CR-OIB – Constructed Response Ordered Item Booklet

D – Dichotomous Item

ECR – Extended Constructed Response

IRT – Item Response Theory

KSA – Knowledge, Skills, and Abilities

M – Multiple Choice Item

NAEP – National Assessment of Educational Progress

NAGB – National Assessment Governing Board

NCES – National Center for Education Statistics

OIB – Ordered Item Booklet

P – Polytomously Scored Item

RP – Response Probability

NOTES

Appendix I

Consequences Questionnaire

2006 Economics NAEP Achievement Levels Setting
QUESTIONNAIRE ON GROUP CONSEQUENCES DATA

You, together with the other panelists, have set cut scores for each achievement level that represent *what students should know and be able to do* according to your common understanding of the achievement levels descriptions. In particular, you used your understanding of the achievement level descriptions of what students *should know and be able to do* to estimate how students performing at the borderline of each achievement level *would* perform on each item within the ordered item book.

A final piece of information is now being provided to you showing how students *actually* performed on the NAEP relative to the final cut scores computed here.

The percentage reported for students scoring at or above the Basic level includes all students who scored at the Basic, Proficient, and Advanced levels. The percentage reported for students scoring at or above Proficient includes all scores at or above the Proficient and Advanced levels. And the percentage scoring at or above the Advanced level includes only those scores at or above the Advanced level. The cut scores are for the group as a whole, and these are the final cut scores.

In this questionnaire you are asked to evaluate the cut scores set by your group for the achievement levels in light of the information provided here as the consequences of those cut scores, (i.e., information about the percentages of students scoring at or above each level). We are interested in knowing whether or not this information about percentages is compelling enough to you that you would alter the cut scores for your group if you had the opportunity to do so.

For your group, please fill in the following percentages obtained using cut scores for each achievement level, based on the recommendations of panelists in your group.

Your group set the Final Basic cut score at ________. This means that approximately ______ percent of students at this grade would score at or above the Basic level on the Economics NAEP.

Your group set the Final Proficient cut score at ________. This means that approximately ______ percent of students at this grade would score at or above the Proficient level on the Economics NAEP.

Your group set the Final Advanced cut score at ________. This means that approximately ______ percent of students at this grade would score at or above the Advanced level on the Economics NAEP.

Please mark the boxes below that correspond to the statements that best characterize your opinions regarding these percentages and the cut scores your group set.

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 for each of these achievement levels?

   □ Yes (Please skip to Number 4.)
   □ No (Please continue to Number 2.)

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?

   □ Yes (Please continue to Number 3.)
   □ No (Please skip to Number 4.)

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 Economics 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

□ Make no change. I am satisfied with the Basic cutscore.

□ 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 _______.

□ 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 _______.

Proficient

□ Make no change. I am satisfied with the Proficient cutscore.

□ 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 _______.

□ 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 _______.

Advanced

□ Make no change. I am satisfied with the Advanced cutscore.

□ 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 _______.

□ 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 _______.

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

 □ I recommend that the achievement levels be reported as set.

 □ I recommend changes consistent with my answers above. If you wish, comment on the magnitude of change you would recommend.

____

____

____

Appendix J

Process Evaluation Questionnaires

Table 1 - Process Evaluation Questionnaire No. 1
Economics ALS

Question	5	4	3	2	1	Mean Score	SD	N
1. The advance materials I received were adequate to prepare me to fulfill my role in this meeting:	Totally Agree 18	7	Somewhat Agree 4	0	Totally Disagree 0	4.48	0.74	29
2. The organization of the advance materials I received for this meeting was:	Very Good 19	6	Acceptable 4	0	Very Poor 0	4.52	0.74	29
3. The amount of time allocated for the General Orientation to the NAEP Program was:	Far Too Long 5	14	About Right 11	0	Far Too Short 0	3.80	0.71	30
4. The explanation of the NAEP in general was:	Absolutely Clear 18	10	Somewhat Clear 2	0	Not at All Clear 0	4.53	0.63	30
5. The explanation of the development of the Economics NAEP was:	Absolutely Clear 18	9	Somewhat Clear 3	0	Not at All Clear 0	4.50	0.68	30
6. The explanation of the major organizations involved in NAEP and the roles of each was:	Absolutely Clear 18	9	Somewhat Clear 3	0	Not at All Clear 0	4.50	0.68	30
7. I understand the purpose of the NAEP pilot study.	Totally Agree 20	9	Somewhat Agree 1	0	Totally Disagree 0	4.63	0.56	30
8. The amount of time allocated for the General Introduction to the NAEP achievement level setting process was:	Far Too Long 4	17	About Right 10	0	Far Too Short 0	3.81	0.65	31
9. I believe my perspectives and experiences will be important in the NAEP standard setting process.	Totally Agree 16	11	Somewhat Agree 4	0	Totally Disagree 0	4.39	0.72	31
10. I understand the difference between criterion-referenced and norm-referenced standards.	Totally Agree 15	8	Somewhat Agree 7	1	Totally Disagree 0	4.19	0.91	31
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 25	4	Somewhat Agree 1	1	Totally Disagree 0	4.71	0.69	31

Question	5	4	3	2	1	Mean Score	SD	N
12. Taking the Economics NAEP was an informative experience.	Totally Agree		Somewhat Agree		Totally Disagree			
	24	6	1	0	0	4.74	0.51	31
13. Taking the Economics NAEP gave me a good idea of what is expected of students.	Totally Agree		Somewhat Agree		Totally Disagree			
	22	9	0	0	0	4.71	0.46	31
14. The amount of time allocated for the Mapmark method orientation was:	Far Too Long		About Right		Far Too Short			
	2	11	17	1	0	3.45	0.68	31
15. The overview of the method to be followed in this meeting was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	12	12	6	1	0	4.13	0.85	31
16. The explanation of how an item map is constructed was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	14	9	7	0	0	4.23	0.82	30
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		Somewhat Agree		Totally Disagree			
	15	10	6	0	0	4.29	0.78	31
18. The explanation of the information in my Ordered Item Book (OIB) was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	17	10	4	0	0	4.42	0.72	31
19. The amount of time allocated for the Framework presentation was:	Far Too Long		About Right		Far Too Short			
	5	13	13	0	0	3.74	0.73	31
20. The presentation of the Economics Framework was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	12	16	3	0	0	4.29	0.64	31
21. The presentation of the Economics Framework had about the right level of detail.	Totally Agree		Somewhat Agree		Totally Disagree			
	10	12	6	0	2	3.93	1.08	30
22. The amount of time allocated for the whole group KSA review was:	Far Too Long		About Right		Far Too Short			
	4	7	17	2	0	3.43	0.82	30
23. The instructions on what I was to do in the KSA review were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	10	15	5	0	0	4.17	0.70	30

Question	5	4	3	2	1	Mean Score	SD	N
24. My understanding of our tasks in the KSA review was:	Totally Adequate		Somewhat Adequate		Totally Inadequate			
	12	14	4	0	0	4.27	0.69	30
25. The whole group work on the common constructed response items was:	Very Useful		Somewhat Useful		Not at All Useful			
	15	8	6	1	0	4.23	0.90	30

Table 2 - Process Evaluation Questionnaire No. 2
Economics ALS

Question	5	4	3	2	1	Mean Score	SD	N
1. The amount of time allocated for the table group KSA review was:	Far Too Long 1	7	About Right 22	1	Far Too Short 0	3.26	0.58	31
2. The table group review of the remaining constructed response items was:	Very Useful 12	14	Somewhat Useful 5	0	Not at All Useful 0	4.23	0.72	31
3. I understand the score levels of polytomous items.	Totally Agree 22	6	Somewhat Agree 3	0	Totally Disagree 0	4.61	0.67	31
4. The amount of time allocated for the independent OIB review was:	Far Too Long 1	10	About Right 20	0	Far Too Short 0	3.39	0.56	31
5. The instructions on what I was to do for the independent OIB review were:	Absolutely Clear 18	11	Somewhat Clear 1	1	Not at All Clear 0	4.48	0.72	31
6. I understood how to use my item map with the ordered item book.	Totally Agree 25	6	Somewhat Agree 0	0	Totally Disagree 0	4.81	0.40	31
7. I was comfortable working through the ordered item book on my own.	Totally Agree 26	5	Somewhat Agree 0	0	Totally Disagree 0	4.84	0.37	31
8. The ordering of the items in the ordered item book agreed with my perceptions of the relative difficulty of the items.	Totally Agree 8	14	Somewhat Agree 9	0	Totally Disagree 0	3.97	0.75	31
9. The KSA work with the OIB helped me understand what can make one item harder than others.	Totally Agree 8	17	Somewhat Agree 5	0	Totally Disagree 0	4.10	0.66	30
10. The amount of time allocated for the table discussion of the OIB was:	Far Too Long 0	4	About Right 27	0	Far Too Short 0	3.13	0.34	31
11. The instructions on what we were to do in the table discussion of the OIB were:	Absolutely Clear 13	16	Somewhat Clear 1	1	Not at All Clear 0	4.32	0.70	31

Question	5	4	3	2	1	Mean Score	SD	N
12. The table discussion of the ordered item book was:	Very Useful 15	14	Somewhat Useful 2	0	Not at All Useful 0	4.42	0.62	31
13. I feel I made a valuable contribution to my table group's discussion.	Totally Agree 17	13	Somewhat Agree 1	0	Totally Disagree 0	4.52	0.57	31
14. I feel my perspective is being heard by others in my table group.	Totally Agree 21	9	Somewhat Agree 1	0	Totally Disagree 0	4.65	0.55	31
15. I feel that I was being pressured to agree with others in my table group.	Totally Agree 1	1	Somewhat Agree 0	8	Totally Disagree 21	1.48	0.93	31
16. The amount of time allocated for the ALD presentation was:	Far Too Long 3	11	About Right 17	0	Far Too Short 0	3.55	0.68	31
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 12	16	Somewhat Agree 3	0	Totally Disagree 0	4.29	0.64	31
18. My own level of satisfaction with the Basic achievement level description is:	Very Satisfied 13	17	Somewhat Satisfied 1	0	Not at All Satisfied 0	4.39	0.56	31
19. My own level of satisfaction with the Proficient achievement level description is:	Very Satisfied 13	16	Somewhat Satisfied 2	0	Not at All Satisfied 0	4.35	0.61	31
20. My own level of satisfaction with the Advanced achievement level description is:	Very Satisfied 13	16	Somewhat Satisfied 2	0	Not at All Satisfied 0	4.35	0.61	31

Question	5	4	3	2	1	Mean Score	SD	N
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			
	19	12	0	0	0	4.61	0.50	31
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			
	20	11	0	0	0	4.65	0.49	31
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			
	20	11	0	0	0	4.65	0.49	31
24. I was comfortable using the concept of performance at the lower borderline of Basic.	Totally Agree		Somewhat Agree		Totally Disagree			
	15	11	5	0	0	4.32	0.75	31
25. I was comfortable using the concept of performance at the lower borderline of Proficient.	Totally Agree		Somewhat Agree		Totally Disagree			
	15	12	4	0	0	4.35	0.71	31
26. I was comfortable using the concept of performance at the lower borderline of Advanced.	Totally Agree		Somewhat Agree		Totally Disagree			
	16	11	4	0	0	4.39	0.72	31
27. I believe my Round 1 bookmark placements are consistent with the achievement level descriptions.	Totally Agree		Somewhat Agree		Totally Disagree			
	12	16	3	0	0	4.29	0.64	31
28. The amount of time allocated for placing the bookmarks was:	Far Too Long		About Right		Far Too Short			
	4	8	18	1	0	3.48	0.77	31
29. The instructions on how I was to place my bookmarks were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	12	13	3	3	0	4.10	0.94	31
30. My understanding of how to use the ALDs to choose my bookmarks was:	Totally Adequate		Somewhat Adequate		Totally Inadequate			
	16	13	2	0	0	4.45	0.62	31

Question	5	4	3	2	1	Mean Score	SD	N
31. The most accurate description of my level of confidence in my Round 1 bookmark placements is:	Totally Confident		Somewhat Confident		Not at All Confident			
	8	14	9	0	0	3.97	0.75	31
32. I felt pressure to recommend bookmarks that were close to those recommended by other panelists.	Totally Agree		Somewhat Agree		Totally Disagree			
	0	1	1	9	20	1.45	0.72	31
33. I was comfortable using a 0.67 probability to define "mastery" in placing my bookmarks.	Totally Agree		Somewhat Agree		Totally Disagree			
	10	13	8	0	0	4.06	0.77	31
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		Somewhat Agree		Totally Disagree			
	5	23	2	0	0	4.10	0.48	30

Table 3 - Process Evaluation Questionnaire No. 3
Economics ALS

Question	5	4	3	2	1	Mean Score	SD	N
1. I understood the Round 1 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree			
	24	7	0	0	0	4.77	0.43	31
2. I understood what students at the Round 1 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree			
	16	15	0	0	0	4.52	0.51	31
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			
	24	7	0	0	0	4.77	0.43	31
4. I understood the cut score dispersion chart (bar graph).	Totally Agree		Somewhat Agree		Totally Disagree			
	20	11	0	0	0	4.65	0.49	31
5. The amount of time allocated for the borderline proficient booklet exercise was:	Far Too Long		About Right		Far Too Short			
	1	8	21	0	0	3.33	0.55	30
6. The instructions I received for the borderline proficient booklet exercise were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	12	16	3	0	0	4.29	0.64	31
7. The purpose of the borderline proficient booklet exercise was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	14	14	3	0	0	4.35	0.66	31
8. The borderline proficient booklet exercise helped me understand how student booklets illustrate performance at a given cut score.	Totally Agree		Somewhat Agree		Totally Disagree			
	16	13	1	0	0	4.50	0.57	30
9. The borderline proficient booklet exercise helped me understand that student performance on individual items may vary even at the same cut score.	Totally Agree		Somewhat Agree		Totally Disagree			
	18	11	0	0	0	4.62	0.49	29
10. The Form C booklet item map was useful.	Totally Agree		Somewhat Agree		Totally Disagree			
	15	11	5	0	0	4.32	0.75	31

Question	5	4	3	2	1	Mean Score	SD	N
11. We should have also done a borderline basic and borderline advanced booklet exercise.	Totally Agree		Somewhat Agree		Totally Disagree			
	1	2	4	17	7	2.13	0.96	31
12. The amount of time allocated for the table group whole booklet review was:	Far Too Long		About Right		Far Too Short			
	3	2	24	2	0	3.19	0.70	31
13. The instructions I received for the table group whole booklet review were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	13	17	1	0	0	4.39	0.56	31
14. The purpose of the table group whole booklet review was:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	14	15	1	0	0	4.43	0.57	30
15. The item maps showing the items in the booklets were useful.	Totally Agree		Somewhat Agree		Totally Disagree			
	14	12	5	0	0	4.29	0.74	31
16. The item score tables were useful.	Totally Agree		Somewhat Agree		Totally Disagree			
	16	13	2	0	0	4.45	0.62	31
17. The booklet score **chart** was useful.	Totally Agree		Somewhat Agree		Totally Disagree			
	15	13	0	2	1	4.26	1.00	31
18. The booklet score **plots** were useful.	Totally Agree		Somewhat Agree		Totally Disagree			
	9	12	7	2	1	3.84	1.04	31
19. I understood the information presented in the booklet score **chart**.	Totally Agree		Somewhat Agree		Totally Disagree			
	19	9	3	0	0	4.52	0.68	31
20. I understood the information presented in the booklet score **plot**.	Totally Agree		Somewhat Agree		Totally Disagree			
	17	11	3	0	0	4.45	0.68	31

Question	5	4	3	2	1	Mean Score	SD	N
21. I understood the information presented in the item score tables.	Totally Agree 21	8	Somewhat Agree 2	0	Totally Disagree 0	4.61	0.62	31
22. At the time I provided the Round 2 cut score recommendations, my understanding of the achievement level descriptions was:	Totally Adequate 25	4	Somewhat Adequate 1	1	Totally Inadequate 0	4.71	0.69	31
23. At the time I provided my Round 2 cut score recommendations, my concept of the lower borderline performance of an achievement level was:	Very Well Formed 19	10	Moderatly Formed 2	0	Not Well Formed 0	4.55	0.62	31
24. I understand the difference between borderline performance and typical performance within an achievement level.	Totally Agree 22	7	Somewhat Agree 2	0	Totally Disagree 0	4.65	0.61	31
25. I believe my Round 2 cut score recommendations are consistent with the lower borderline of the achievement level descriptions.	Totally Agree 20	8	Somewhat Agree 3	0	Totally Disagree 0	4.55	0.68	31
26. The amount of time allocated for my Round 2 cut score recommendations was:	Far Too Long 0	8	About Right 21	1	Far Too Short 1	3.16	0.64	31
27. The instructions I received for recommending Round 2 cut scores were:	Absolutely Clear 20	10	Somewhat Clear 1	0	Not at All Clear 0	4.61	0.56	31
28. My level of understanding of how I was to choose cut scores for Round 2 was:	Totally Adequate 23	5	Somewhat Adequate 2	0	Totally Inadequate 0	4.70	0.60	30
29. The most accurate description of my level of confidence in my Round 2 cut score recommendations is:	Totally Confident 12	17	Somewhat Confident 2	0	Not at All Confident 0	4.32	0.60	31

Question	5	4	3	2	1	Mean Score	SD	N
30. I felt pressure to recommend cut scores that were close to those recommended by other panelists.	Totally Agree 0	1	Somewhat Agree 0	8	Totally Disagree 22	1.35	0.66	31
31. I was comfortable choosing scale values instead of placing bookmarks to recommend cut scores.	Totally Agree 17	10	Somewhat Agree 3	0	Totally Disagree 1	4.35	0.91	31
32. The work with the whole booklets was helpful for setting my Round 2 cut scores.	Totally Agree 16	10	Somewhat Agree 5	0	Totally Disagree 0	4.35	0.75	31
33. The booklet score chart was helpful to me for selecting a cut score.	Totally Agree 14	12	Somewhat Agree 5	0	Totally Disagree 0	4.29	0.74	31
34. I was comfortable locating my cut score selections in both the OIB and the booklet score chart.	Totally Agree 22	8	Somewhat Agree 1	0	Totally Disagree 0	4.68	0.54	31

Table 4 - Process Evaluation Questionnaire No. 4
Economics ALS

Question	5	4	3	2	1	Mean Score	SD	N
1. I understood the Round 2 median cut scores.	Totally Agree 25	6	Somewhat Agree 0	0	Totally Disagree 0	4.81	0.40	31
2. I understood what students at the Round 2 median cut scores can do.	Totally Agree 23	8	Somewhat Agree 0	0	Totally Disagree 0	4.74	0.44	31
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 25	6	Somewhat Agree 0	0	Totally Disagree 0	4.81	0.40	31
4. I understood the cut score dispersion chart (bar chart).	Totally Agree 27	3	Somewhat Agree 1	0	Totally Disagree 0	4.84	0.45	31
5. I understood the feedback on the booklet score chart.	Totally Agree 28	2	Somewhat Agree 1	0	Totally Disagree 0	4.87	0.43	31
6. I understood the feedback on the booklet score plots.	Totally Agree 26	3	Somewhat Agree 2	0	Totally Disagree 0	4.77	0.56	31
7. I understood the consequences data.	Totally Agree 28	3	Somewhat Agree 0	0	Totally Disagree 0	4.90	0.30	31
8. The instructions I received for using consequences data during Round 3 were:	Absolutely Clear 21	10	Somewhat Clear 0	0	Not at All Clear 0	4.68	0.48	31
9. The amount of time allocated for discussing the consequences data was:	Far Too Long 4	6	About Right 20	1	Far Too Short 0	3.42	0.76	31
10. The most accurate description of my level of confidence in using the consequences data to recommend cut scores in Round 3 is:	Totally Confident 20	9	Somewhat Confident 1	1	Not at All Confident 0	4.55	0.72	31
11. At the time I provided the Round 3 cut score recommendations, my understanding of the achievement level descriptions was:	Totally Adequate 26	4	Somewhat Adequate 1	0	Totally Inadequate 0	4.81	0.48	31

Question	5	4	3	2	1	Mean Score	SD	N
12. At the time I provided the Round 3 cut score recommendations, my concept of the lower borderline performance of an achievement level was.	Very Well Formed		Moderately Formed		Not Well Formed			
	26	4	1	0	0	4.81	0.48	31
13. I understand the difference between borderline performance and typical performance within an achievement level.	Totally Agree		Somewhat Agree		Totally Disagree			
	27	4	0	0	0	4.87	0.34	31
14. I believe my Round 3 cut score recommendations are consistent with the achievement level descriptions.	Totally Agree		Somewhat Agree		Totally Disagree			
	28	2	1	0	0	4.87	0.43	31
15. The instructions I received for recommending Round 3 cut scores were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	25	5	1	0	0	4.77	0.50	31
16. My level of understanding of how I was to choose cut scores for Round 3 was.	Totally Adequate		Somewhat Adequate		Totally Inadequate			
	27	2	2	0	0	4.81	0.54	31
17. 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			
	25	5	1	0	0	4.77	0.50	31
18. I felt pressure to recommend cut scores that were close to those recommended by other panelists.	Totally Agree		Somewhat Agree		Totally Disagree			
	3	0	3	2	23	1.65	1.28	31

Table 5 - Process Evaluation Questionnaire No. 5
Economics ALS

Question	5	4	3	2	1	Mean Score	SD	N
1. I understood the Round 3 median cut scores.	Totally Agree		Somewhat Agree		Totally Disagree			
	26	3	2	0	0	4.77	0.56	31
2. I understood what students at the Round 3 median cut scores can do.	Totally Agree		Somewhat Agree		Totally Disagree			
	27	2	2	0	0	4.81	0.54	31
3. The amount of time allocated for the Consequences Questionnaire was:	Far Too Long		About Right		Far Too Short			
	0	6	25	0	0	3.19	0.40	31
4. I understood the Round 3 consequences data.	Totally Agree		Somewhat Agree		Totally Disagree			
	24	6	0	0	0	4.80	0.41	30
5. The instructions I received for completing the Consequences Questionnaire were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	23	8	0	0	0	4.74	0.44	31
6. I understood how to complete the Consequences Questionnaire.	Totally Agree		Somewhat Agree		Totally Disagree			
	25	6	0	0	0	4.81	0.40	31
7. The instructions on what I was to do during each round were:	Absolutely Clear		Somewhat Clear		Not at All Clear			
	19	11	1	0	0	4.58	0.56	31
8. My understanding of the tasks I was to accomplish during each round was:	Totally Adequate		Somewhat Adequate		Totally Inadequate			
	23	7	1	0	0	4.71	0.53	31
9. 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			
	24	7	0	0	0	4.77	0.43	31
10. 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			
	1	10	20	0	0	3.39	0.56	31

Question	5	4	3	2	1	Mean Score	SD	N
11. I would describe the effectiveness of this achievement level setting method as:	Highly Effective 16	12	Somewhat Effective 3	0	Not at All Effective 0	4.42	0.67	31
12. I felt my input was valued and considered by others in my group.	To a Great Extent 20	8	Somewhat 1	1	Not at All 1	4.45	0.96	31
13. I felt pressured by others in my group to make my cut score recommendations agree with theirs.	To a Great Extent 1	0	Somewhat 1	3	Not at All 26	1.29	0.82	31
14. I felt pressured by staff to make cut score recommendations higher or lower.	To a Great Extent 0	0	Somewhat 8	3	Not at All 20	1.61	0.88	31
15. I felt pressured by staff to keep my cut score recommendations the same.	To a Great Extent 0	0	Somewhat 1	4	Not at All 26	1.19	0.48	31
16. The amount of time allocated for the Exemplar Item Rating Task was:	Far Too Long 0	8	About Right 23	0	Far Too Short 0	3.26	0.44	31
17. The instructions I received for the Exemplar Item Rating Task were:	Absolutely Clear 23	6	Somewhat Clear 2	0	Not at All Clear 0	4.68	0.60	31
18. My understanding of how I was to perform the Exemplar Item Rating Task was:	Totally Adequate 26	3	Somewhat Adequate 2	0	Totally Inadequate 0	4.77	0.56	31
19. I believe the exemplar items will be useful for describing the achievement levels.	Totally Agree 20	10	Somewhat Agree 1	0	Totally Disagree 0	4.61	0.56	31
20. The exemplar items I reviewed seemed appropriately matched to their achievement level.	Totally Agree 15	10	Somewhat Agree 6	0	Totally Disagree 0	4.29	0.78	31
21. I understood the purpose of this meeting.	Totally Agree 28	3	Somewhat Agree 0	0	Totally Disagree 0	4.90	0.30	31

Question	5	4	3	2	1	Mean Score	SD	N
22. I feel that this ALS process provided me an opportunity to use my best judgment to recommend cut scores for the NAEP economics assessment.	To a Great Extent 26	4	Somewhat 1	0	Not at All 0	4.81	0.48	31
23. I feel that this ALS process has produced achievement levels that are defensible.	To a Great Extent 23	7	Somewhat 1	0	Not at All 0	4.71	0.53	31
24. I feel that this ALS process has produced achievement levels that will generally be considered reasonable.	To a Great Extent 23	6	Somewhat 2	0	Not at All 0	4.68	0.60	31
25. I believe that the achievement levels capture meaningful distinctions in economics performance as described in the ALDs.	Totally Agree 18	10	Somewhat Agree 3	0	Totally Disagree 0	4.48	0.68	31
26. 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 16	11	Somewhat 3	1	Not at All 0	4.35	0.80	31
27. I feel that the panelists in this meeting are appropriately qualified for setting NAEP achievement levels.	To a Great Extent 19	10	Somewhat 2	0	Not at All 0	4.55	0.62	31
28. 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 23	6	1	No, definitely not 0	3.73	0.52	30
29. Having observers present influenced my judgments.	To a Great Extent 0	0	Somewhat 2	3	Not at All 26	1.23	0.56	31
30. During the ALS process, I found the Achievement Level Descriptions:	Very Helpful 24	5	Somewhat Helpful 2	0	Not at All Helpful 0	4.71	0.59	31
31. During the ALS process, I found the Ordered Item Booklet:	Very Helpful 27	3	Somewhat Helpful 1	0	Not at All Helpful 0	4.84	0.45	31
32. During the ALS process, I found the Primary Item Map:	Very Helpful 25	3	Somewhat Helpful 3	0	Not at All Helpful 0	4.71	0.64	31

Question	5	4	3	2	1	Mean Score	SD	N
33. During the ALS process, I found the Rater Location Data (the location of my cut scores relative to the median cut scores):	Very Helpful 16	7	Somewhat Helpful 7	1	Not at All Helpful 0	4.23	0.92	31
34. During the ALS process, I found the Consequences Data:	Very Helpful 12	9	Somewhat Helpful 6	3	Not at All Helpful 1	3.90	1.14	31
35. During the ALS process, I found the Booklet Score Charts:	Very Helpful 15	9	Somewhat Helpful 7	0	Not at All Helpful 0	4.26	0.82	31
36. During the ALS process, I found the Booklet Score Plots:	Very Helpful 10	13	Somewhat Helpful 6	1	Not at All Helpful 1	3.97	0.98	31
37. During the ALS Process, I found the Cut Score Dispersion Chart:	Very Helpful 18	7	Somewhat Helpful 6	0	Not at All Helpful 0	4.39	0.80	31
38. I would rate the amount of personal attention and assistance I received from the process facilitator (Christina Peterson) as:	Too Much 0	1	About Right 30	0	Too Little 0	3.03	0.18	31
39. I would rate the amount of personal attention and assistance I received from the content facilitator (RJ Goodman) as:	Too Much 0	0	About Right 31	0	Too Little 0	3.00	0.00	31
40. My employer supported my participation in this meeting:	Totally Agree 27	1	Somewhat Agree 3	0	Totally Disagree 0	4.77	0.00	31
41. I had to take vacation time in order to attend this meeting:	Totally Agree 4	0	Somewhat Agree 0	0	Totally Disagree 27	1.52	0.00	31

Appendix K

Exemplar Ratings

2006 NAEP Economics
Exemplar Item Ratings - Basic

				ALS					Special Studies				Content Experts
			Item	Rating as Exemplar			Percent	Percent	Number Classifying Item into Each Level				
		Content	Page	Very		Do Not	Very	Do Not	Below				Do Not
Item	Scale	Area	Number	Good	OK	Use	Good	Use	Basic	Basic	Proficient	Advanced	Use*
P1_2		Mkt	F-7	23	8	0	74%	0%	5	8	0	0	1
M20		Natl	F-9	21	10	0	68%	0%	4	9	0	0	1
M52		Mkt	F-21	19	12	0	61%	0%	0	12	1	0	
M25		Mkt	F-10	14	17	0	45%	0%	0	10	3	0	
M15		Natl	F-6	10	21	0	32%	0%	1	12	0	0	2
M51		Intl	F-16	12	17	2	39%	6%	0	8	5	0	
P8_2		Mkt	F-11	14	11	6	45%	19%	9	4	0	0	
M42		Mkt	F-14	6	19	6	19%	19%	0	10	2	1	1
P21_1		Natl	F-17	9	10	12	29%	39%	0	8	5	0	2
M50		Mkt	F-15	7	9	15	23%	48%	0	2	11	0	1

*Note: Three content experts were asked to rate the exemplars after the ALS was complete

2006 NAEP Economics
Exemplar Item Ratings - Proficient

				ALS					Special Studies				Content Experts
			Item	Rating as Exemplar			Percent	Percent	Number Classifying Item into Each Level				
Item	Scale	Content Area	Page Number	Very Good	OK	Do Not Use	Very Good	Do Not Use	Below Basic	Basic	Proficient	Advanced	Do Not Use
M113		Natl	F-53	26	5	0	84%	0%	0	0	13	0	
P4_2		Mkt	F-35	25	6	0	81%	0%	0	9	4	0	
M67		Mkt	F-27	24	7	0	77%	0%	0	6	7	0	
M73		Natl	F-28	19	12	0	61%	0%	0	4	9	0	
M87		Mkt	F-33	18	13	0	58%	0%	0	5	8	0	
M90		Intl	F-40	18	13	0	58%	0%	0	0	13	0	
M132		Mkt	F-62	16	15	0	52%	0%	0	5	8	0	
M80		Natl	F-31	25	5	1	81%	3%	0	0	13	0	
M101		Mkt	F-49	22	8	1	71%	3%	0	4	9	0	
M76		Mkt	F-30	21	9	1	68%	3%	0	8	5	0	
M60		Natl	F-26	20	10	1	65%	3%	0	8	5	0	2
M81		Mkt	F-32	20	10	1	65%	3%	0	5	8	0	
M96		Natl	F-48	18	12	1	58%	3%	0	0	13	0	
M55		Intl	F-22	15	15	1	48%	3%	0	5	4	4	3
M106		Intl	F-51	15	15	1	48%	3%	0	0	13	0	
M74		Mkt	F-29	7	23	1	23%	3%	0	4	9	0	
M104		Mkt	F-50	16	13	2	52%	6%	0	0	12	1	
M128		Natl	F-60	15	14	2	48%	6%	0	11	2	0	
M131		Natl	F-63	14	15	2	45%	6%	0	4	9	0	
P7_2		Natl	F-58	10	19	2	32%	6%	0	5	8	0	
M58		Mkt	F-25	15	13	3	48%	10%	0	10	3	0	2
M118		Natl	F-55	10	17	4	32%	13%	0	2	11	0	1
P7_1		Natl	F-23	8	19	4	26%	13%	10	3	0	0	1
M123		Mkt	F-56	21	5	5	68%	16%	0	13	0	0	
P15_2		Mkt	F-37	15	11	5	48%	16%	0	9	4	0	1
P12_1		Natl	F-41	15	11	5	48%	16%	0	2	11	0	
M88		Natl	F-34	7	19	5	23%	16%	0	4	9	0	
M134		Natl	F-64	14	11	6	45%	19%	0	0	9	4	
M129		Natl	F-61	15	9	7	48%	23%	0	13	0	0	1
M124		Natl	F-57	7	15	9	23%	29%	5	8	0	0	
M110		Mkt	F-52	5	17	9	16%	29%	0	9	4	0	1
P19_1		Natl	F-44	8	13	10	26%	32%	0	8	5	0	
M112		Intl	F-54	7	14	10	23%	32%	0	0	4	9	1
M93		Intl	F-47	11	9	11	35%	35%	0	0	13	0	1

2006 NAEP Economics
Exemplar Item Ratings - Advanced

				ALS					Special Studies				Content Experts
			Item	Rating as Exemplar			Percent	Percent	Number Classifying Item into Each Level				
Item	Scale	Content Area	Page Number	Very Good	OK	Do Not Use	Very Good	Do Not Use	Below Basic	Basic	Proficient	Advanced	Do Not Use
P12_2		Natl	F-74	21	9	1	68%	3%	0	0	8	5	
M149		Intl	F-78	21	9	1	68%	3%	0	0	9	4	
M148		Natl	F-77	14	16	1	45%	3%	0	5	8	0	
P8_3		Mkt	F-65	22	7	2	71%	6%	0	9	4	0	
M140		Natl	F-72	20	9	2	65%	6%	0	0	13	0	
P21_3		Natl	F-90	20	9	2	65%	6%	0	0	6	7	
P15_3		Mkt	F-81	16	12	3	52%	10%	0	0	13	0	1
P19_2		Natl	F-87	16	12	3	52%	10%	0	0	13	0	
P21_2		Natl	F-68	18	9	4	58%	13%	0	0	13	0	2
M143		Intl	F-73	16	11	4	52%	13%	0	0	13	0	
P25_2		Intl	F-94	14	13	4	45%	13%	0	0	13	0	1
P13_2		Mkt	F-79	14	12	5	45%	16%	0	9	4	0	
P25_1		Intl	F-84	10	15	5	33%	17%	0	13	0	0	1