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

Technical Report

Presented by ACT, Inc.
July 23, 2007

Redacted by Governing Board

The National Assessment Governing Board

Darvin M. Winick, Chair
President
Winick & Associates
Austin, Texas

Amanda P. Avallone,
Vice Chair
Assistant Principal &
Eighth-Grade Teacher
Summit Middle School
Boulder, Colorado

Francie M. Alexander
Chief Academic Officer,
Scholastic, Inc.
Senior Vice President, Scholastic Education
New York, New York

David J. Alukonis
Chairman
Hudson School Board
Hudson, New Hampshire

Barbara Byrd-Bennett
Executive Superintendent
in Residence
Cleveland State University
Cleveland, Ohio

Gregory Cizek
Professor of Educational Measurement
University of North Carolina
Chapel Hill, North Carolina

Shirley V. Dickson
Educational Consultant
Aliso Viejo, California

David P. Driscoll
Commissioner of Education
Commonwealth of Massachusetts
Malden, Massachusetts

John Q. Easton
Executive Director
Consortium on Chicago School Research
Chicago, Illinois

Alan J. Friedman
Director and CEO
New York Hall of Science
Queens, New York

David W. Gordon
Sacramento County Superintendent of Schools
Sacramento County Office of Education
Sacramento, California

Robin C. Hall
Principal
Beecher Hills Elementary
Atlanta, Georgia

Kathi M. King
Twelfth-Grade Teacher
Messalonskee High School
Oakland, Maine

Honorable Keith King
Member
Colorado House of Representatives
Colorado Springs, Colorado

Kim M. Kozbial-Hess
Fourth-Grade Teacher
Hawkins Elementary School
Toledo, Ohio

James S. Lanich
President
California Business for Educational Excellence
Sacramento, California

Honorable Cynthia Nava
Chair, Education Committee
New Mexico State Senate
Las Cruces, New Mexico

Andrew C. Porter
Director
Learning Sciences Institute
Peabody College
Vanderbilt University
Nashville, Tennessee

Luis A. Ramos
Community Relations Manager
PPL Susquehanna
Berwick, Pennsylvania

Mary Frances Taymans, SND
Executive Director
Secondary Schools Department
National Catholic Education Association
Washington, D.C.

Oscar A. Troncoso
Principal
Socorro High School
Socorro ISD
El Paso, Texas

Michael E. Ward
Former State Superintendent
of Public Instruction
North Carolina Public Schools
Jackson, Mississippi

Eileen L. Weiser
Member
State Board of Education
Michigan Department of Education
Lansing, Michigan

Ex-officio Member
Grover (Russ) Whitehurst
Director
Institute of Education Sciences
U.S. Department of Education
Washington, D.C.

Charles E. Smith
Executive Director

Susan Loomis
Assistant Director

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

Technical Report

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

Technical Report
Table of Contents

List of Figures iii
List of Tables iv
Introduction 1
Psychometric Procedures 2
Description of Item Pool 2
Computation of Item Scale Values 3
Item Handles 6
Item Map Values 7
Whole Booklet Feedback 8
Consequences Feedback 8
Mapping Potential Exemplar Items to Achievement Levels 8
Reliability Estimates 9
Process Evaluations 10
Materials and Procedures 10
Briefing Book 10
Division of Item Pool and Panel into Pools/Groups A and B 10
Test Form Administered to Panelists 11
Ordered Item Book (OIB) 12
Constructed Response Ordered Item Book (CROIB) 12
KSA Notes Template 13
Cut Score Recommendation Form and Data Processing 14
Item Map 15
Scale Value to OIB Page Lookup Table 16
Consequences Feedback and Questionnaire 16
Exemplar Item Rating Form 16
Field Trial, Pilot Study, and Special Studies 17
Computation of Domain Characteristic Curves 18
Domain Item Map Percentages 18
Percent Correct Table 18
Domain Task 1 Form 19
Domain Ordered Item Booklet 20
Domain Task 2 Form 20
Domain Score Chart 22
Domain Score Plots 22
Special Studies 23
Computer Programs 23
Programs for Mapmark 23
References 26

APPENDICES

A. TACSS Meeting Summaries
B. Item Information
C. Frequency Distribution of Student Performance
D. Briefing Booklet
E. Primary Item Map
F. Domain Titles and Definitions

List of Figures

Figure 1. Contents of Constructed Response Ordered Item Book by group....13
Figure 2. Sample KSA Notes template....14
Figure 3. Panelist Cut Score Recommendation Form....15
Figure 4. Consequences data....16
Figure 5. Exemplar Item Rating Form for the Basic achievement level....17
Figure 6. Portion of the Percent Correct Table for Market content areas with mastered domains highlighted....19
Figure 7. Domain Task 1 form for the International content area, Group A....20
Figure 8. Domain Task 2 form for the Proficient level....21
Figure 9. Domain Score Plot for a set of domains....22

List of Tables

Table 1. Summary of Item Pool by Block 3
Table 2. Transformation Constants and Weights to Form Composite 4
Table 3. Content Area Correlations 5
Table 4. Marginal Content Area Theta Means and Standard Deviations 5
Table 5. Item Handles, Scale Values, and Map Values for Hardest and Easiest Items within Item Type 7
Table 6. Estimates of Standard Error of Cut Scores 10
Table 7. Summary of Item Pools A and B 11
Table 8. Summary Information about Test Form Taken by Panelists 12
Table 9. Special Studies 17

INTRODUCTION

In September 2006, the National Assessment Governing Board (NAGB) contracted with ACT to conduct research and other activities for setting achievement levels on the 2006 National Assessment of Educational Progress (NAEP) in Grade 12 economics. The contract called for a series of reports, including a Technical Report documenting the technical aspects of ACT's contract activities.

This Technical Report is provides information on the materials and process of the Achievement Level Setting (ALS) meeting that was held in March 2007. The data used in the meeting consisted of items, item statistics, and estimates of student achievement from the test of the 2006 NAEP in Grade 12 economics. The methodology used to set the achievement levels was Mapmark with whole booklets, a bookmark-based procedure that includes item maps and whole booklet feedback.

This report also provides information for the technical aspects of several Special Studies ACT conducted in the course of the project. These studies included a Field Trial, a comparison of feedback methods done during a Pilot Study, and two studies intended to assess the reasonableness of the results of the standard setting process.

In addition to this Technical Report, the following reports contain information about ACT's activities in this project:

1. The Process Report (ACT, 2007a) provides an overview of the Field Trial and Pilot Study and a detailed description of the process and results used in the ALS meeting.

2. The Special Studies Report (ACT, 2007b) provides a description of the purpose, methods, materials, results, and conclusions of two Special Studies conducted in this project. The Special Studies were designed to evaluate the reasonableness of the outcomes of the ALS.

This document is divided into four primary sections: psychometric procedures, materials and procedures, Special Studies, and computer programs.

The psychometric procedures section deals with the statistical characteristics and calculations used during the achievement level setting process. This includes descriptive information for the items used, the description of the statistics used in the meetings and the subsequent analysis of the results. If necessary, the method for calculating the statistic is also given.

The materials and procedures section shows the materials that were given to the panelists during the ALS meeting. This includes preparatory materials the panelists received prior to the meeting as well as materials for each round. A description is given, along with an example of the material.

The Special Studies section describes the other studies conducted as part of the achievement level setting process. These include a Field Trial to test the whole booklet

feedback portion of the Mapmark process; the Pilot Study, which compared the whole booklet feedback method with a domain based feedback method; and two special studies to investigate the consistency of the cut scores from the ALS. The section details the studies including any technical information and materials used.

The final section lists the computer programs used in the study. The name of the program, along with a brief description of what the program does and the inputs and outputs are given.

The Technical Report accounts for technical advice ACT received throughout this project. ACT relied on the advice of a Technical Advisory Committee on Standard Setting (TACSS), the Contract Officer's Representative (COR), and the Committee on Standards, Design, and Methodology (COSDAM). The TACSS is a five-member group that collectively represents expertise in standard setting, economics education, and experience with the NAEP. The COR was Dr. Susan Loomis, the Assistant Director of Psychometrics for NAGB. The COSDAM is a committee of the NAGB Board. The TACSS met four times over the course of the project and provided technical advice concerning all aspects of the project. This input is presented in the form of meeting summaries in Appendix A of this report and is also described in this and other reports described above. Meetings were held with the COSDAM at critical decision points during the contract. Input from these meetings was used to guide the processes that were finally used in the ALS meeting. Additionally, internal to ACT, there was a Technical Advisory Team (TAT) that provided guidance for technical issues arising during preparations for the meetings.

PSYCHOMETRIC PROCEDURES

Description of Item Pool

The Achievement Level Setting (ALS) meeting used items, item statistics, and student performance data from the 2006 NAEP in Grade 12 economics.

Table 1 presents a summary of the scored items used in the ALS meeting. The items were organized into ten blocks, labeled 1 through 10. There were 17 to 20 items in each block of the 186 scored items, 154 were multiple choice, 3 were dichotomously scored constructed response, and 29 were polytomously scored constructed response. The polytomously scored items represented a total of 68 score points, or 30% of the points in the item pool. Dichotomously scored items represented 1% of the points, and multiple choice items represented 68% of the points. The total number of points was 225. There were 187 numbered items in the test booklets, but only 186 items with item statistics. The difference in counts is attributed to the fact that Block 5, Item 7 was dropped by the assessment development contractor for lack of fit to the psychometric model. Table 1 shows how the items were distributed by content area and item type.

Table 1. Summary of Item Pool by Block

Block	All Items	Number of Items with Item-Statistics: Content Area[a] Mkt	Natl	Intl	Item Type[b] MC	DI	Poly	P[c] Points
1	17	6	8	3	13	0	4	9
2	18	9	6	3	15	0	3	7
3	19	9	8	1	15	1	3	6
4	18	8	7	3	15	0	3	7
5	20	10	8	2	17	0	3	7
6	18	10	5	3	15	0	3	8
7	18	7	8	3	15	0	3	7
8	20	10	8	2	17	1	2	5
9	20	10	6	4	17	1	2	4
10	18	8	7	3	15	0	3	8
Total	**186**	**87**	**72**	**27**	**154**	**3**	**29**	**68**
		47%	39%	15%	68%	1%		30%

[a] Mkt = Market Economy, Natl = National Economy, Intl = International Economy
[b] MC = Multiple choice; DI = Dichotomously scored constructed response; Poly = Polytomously scored constructed response
[c] P Points = the number of score points represented by polytomously-scored items

Computation of Item Scale Values

Each item in the assessment is calibrated separately to one of the three content areas shown in Table 2: Market Economy, National Economy, or International Economy. The slope and intercept determine the characteristic of each of the three content areas and were provided by the Governing Board's subcontractor. The weights are determined as part of the framework development.

The computation of item scale values in the Mapmark procedure begins with the computation of score probabilities conditional on the content areas. Let U_{ij} represent the random score on item i associated with subscale j and let θ_j represent student achievement on subscale j. For multiple choice and dichotomously scored items, the following item response theory model was used:

$$P\left(U_{ij}=1 \mid \theta_j\right) = p_{ij} = c_{ij} + \frac{1-c_{ij}}{1+\exp\left[-Da_{ij}\left(\theta_j - b_{ij}\right)\right]}, \quad (1)$$

where D is 1.7, a_{ij} is the item discrimination parameter, b_{ij} is the item difficulty parameter, c_{ij} is the pseudo-guessing parameter for multiple choice items or $c_{ij} = 0$ for dichotomously scored constructed response items. For polytomously scored items, the following item response theory model was used:

$$P(U_{ij} = h \mid \theta_j) = p_{ijh} = \frac{\exp\left[\sum_{r=0}^{h} Da_{ij}(\theta_j - b_{ij} + d_{ijr})\right]}{\sum_{k=0}^{m_{ij}} \exp\left[\sum_{r=0}^{k} Da_{ij}(\theta_j - b_{ij} + d_{ijr})\right]}, \quad (2)$$

where m_{ij} is the maximum score on the item, and d_{ijr} is the threshold parameter for score r, $r=0,1,\ldots,m_{ij}$, and $d_{ij0} = 0$.

The composite scale score, η, is related to subscale thetas, $\theta = \{\theta_1, \theta_2, \theta_3\}$, through the transformations:

$$y = A\theta + b \quad (3)$$

and

$$\eta = w^t y, \quad (4)$$

where A is a diagonal matrix of constants, b is a column vector of constants, and w is a column vector of weights summing to 1. Table 2 shows the transformation constants used to create the composite score scale used in the ALS meeting.

Table 2. Transformation Constants and Weights to Form Composite

Content Area Notation (j)	Content Area	Slope (diag A)	Intercept (b)	Weight (w)
1	Market Economy	37.781	148.982	0.45
2	National Economy	41.083	148.292	0.40
3	International Economy	31.996	151.470	0.15

To obtain the probability of scoring at or above h, conditional on η, a regression procedure based on Donoghue (1997) was used. The following integral was approximated numerical integration

$$P(U_{ij} \geq h \mid \eta) = \int_{-\infty}^{\infty} P(U_{ij} \geq h \mid \theta_j) f(\theta_j \mid \eta) \partial\theta_j \quad , \quad (5)$$

where

$$P(U_{ij} \geq h \mid \theta_j) = \sum_{k=h}^{m_{ij}} P(U_{ij} = k \mid \theta_j), \text{ for } h = 1 \text{ or } h = 1, 2, \ldots, m_{ij}, \quad (6)$$

$$f(\theta \mid \eta) \sim N\left(\mu_j + \frac{\sigma_j \rho_{j\eta}(\eta - \mu_j)}{\sigma_j}, \sigma_j^2(1 - \rho_{j\eta}^2)\right), \quad (7)$$

where μ_j and σ_j are the mean and standard deviation of θ_j, μ_η and σ_η are the mean and standard deviation of the composite scale value η, and $\rho_{j\eta}$ is the correlation between θ_j and η, calculated as:

$$\rho_{j\eta} = \frac{Cov(\theta_j, \eta)}{\sigma_j \sigma_\eta}, \tag{8}$$

Where $Cov(\theta_j, \eta)$ is the covariance between θ_j and η ,

$$Cov(\theta_j, \eta) = \sum_{k=1}^{3} w_k A_{kk} Cov(\theta_j, \theta_k) = \sum_{k=1}^{3} w_k A_{kk} \rho_{jk} \sigma_j \sigma_k . \tag{9}$$

The correlations between content area thetas (ρ_{jk}) based on the item statistics used in the ALS meeting are shown in Table 3. The marginal means (μ_j) and standard deviations of the subscale thetas (σ_j) are shown in Table 4. Elements of the weight vector (w_k) and the diagonal elements of the slope matrix A (A_{kk}) are shown in Table 2. The mean and standard deviation of student achievement on the composite score scale (μ_η and σ_η) were, respectively, 150.00 and 34.33.

Table 3. Content Area Correlations

Content Area Notation (j)	Content Area	1	2	3
1	Market Economy	1.0000		
2	National Economy	0.9576	1.0000	
3	International Economy	0.9190	0.9079	1.0000

Table 4. Marginal Content Area Theta Means and Standard Deviations

Content Area Notation (j)	Content Area	Theta Mean (μ_j)	Theta SD (σ_j)
1	Market Economy	0.0269	0.9264
2	National Economy	0.0416	0.8519
3	International Economy	-0.0459	1.0939

An *item scale value* was obtained for every score point greater than 0 on an item. Let η_{ijh} represent the composite scale value of item score h ($h > 0$) on item i associated with subscale j. The value of η_{ijh} was the lowest integer value of η that satisfied the following condition:

$$P(U_{ij} \geq h \mid \eta) \geq RP, \tag{10}$$

where *RP* stands for the response probability criterion (RP). For the ALS meeting, an RP of 0.67 was used. If the left side of Equation 10 was less than RP when $\eta = 300$, then η_{ijh} was set to 301.

In the economics ALS process, 203 was added to the item scale value obtained as described with reference to Equation 10. This was done in order to disguise the true scale values from panelists, who may have been familiar with the cut scores from other NAEP assessments. This addition produced item scale values ranging from 204 to 504. There was no item for which the conditional probability was 0.67 or higher at scale values less than 284. Item scale values on the Mapmark scale (285 to 504) are shown in the "Scale Value to OIB Page" section of Appendix B.

Item Handles

An item handle is a short character string that represents the item on the item map. Polytomously scored items had more than one item handle—one for each score point above zero.

The first character in the item handle is "M" if the item is multiple choice, "D" if the item is dichotomously scored constructed response, and "P" if the item is polytomously scored.

For multiple choice (M) and dichotomously scored (D) items, the remaining characters in the item handle indicate the rank of the item by its scale value, from easy to hard, with the easiest item having a rank of 1. Items were ranked separately by item type. For example, the multiple choice item handles were numbered M1 to M154. The 3 dichotomously scored item handles were numbered D1 to D3.

Table 5 shows the handles, scale values, and map values for the easiest and most difficult items within each type. Some of these items have scale values outside the range of score intervals on the item map—280 to 498—and are, therefore, located in the rows or categories on the item map labeled "above."

The item handle for a score on a polytomously scored item shows the score that is being represented specifically, and also shows the difficulty order of the highest possible score on the item. For example, the handle P1_2 represents a score of "2" on item P1. More precisely, item P1 is the easiest polytomously scored item in terms of the level of achievement (scale value) that corresponds to a 0.67 probability of earning full credit on the item (a score of 2). As shown in Table 5, each score level of Item P1, as well as each score level of every other polytomously scored item, is indicated by a distinct item handle. Each of these score levels is represented separately and in different locations on the item

map and in the Ordered Item Book corresponding to their respective scale values or map values.

Table 5. Item Handles, Scale Values, and Map Values for Hardest and Easiest Items within Item Type

Item Type	Item Handle	Scale Value	Map Value
Multiple Choice	M154	469	470
	M153	456	455
	M152	452	452
	M151	447	446
	.		
	.		
	.		
	M4	297	296
	M3	295	296
	M2	294	293
	M1	285	284
Dichotomously Scored	D3	403	404
	D2	384	383
	D1	325	326
Polytomously Scored	P28 4	off scale	above
	P28 3	476	476
	P28_2	427	428
	P28_1	381	380
	.	.	.
	.	.	.
	.	.	.
	P2 2	364	365
	P2_1	304	305
	P1_2	338	338
	P1 1	290	290

Item Map Values

An item's map value was the midpoint of the score interval in which the item was located on the item map. The map was divided into 51 score intervals, plus an extreme catch-all category labeled "above." The score intervals were three units wide and represented scale scores ranging from 280 to 498. (The interval midpoints ranged from 281 to 497 in steps of 3.) Items with scale values outside this range were represented in the "above" category. Of the 225 item scale values, five were represented as "above" 498.

Whole Booklet Feedback

Feedback was given to the panelists in the form of student performance on NAEP test booklets. Booklets were selected to represent specific ranges of performance and given to panelists to review. The booklets were selected close to the cut scores of the various levels, as well as at the midpoints of the levels. The level of a booklet was determined using an expected number correct END score. The ENC score for a given scale value is given as:

$$ENC = \sum_{ijk} P\left(I_{ijk} = 1 \mid \eta\right), \tag{11}$$

where η is the composite scale value and I_{ijk} is an indicator function for a score of at least k on item i in content area j. The index k will equal 1 for all multiple choice items and range from 1 to m_{ij} for constructed response items, where m_{ij} is the total number of score points possible on the item.

The ENC is calculated for each possible scale value. Booklets were then stratified by the number of total points received, and booklets closest to the ENC at the cut scores were considered. If the scale score associated with the given number correct score was within one scale point of the cut score, then two booklets were selected at that score level. If there was no booklet within one scale point, then the closest booklet above and below the ENC at the cut score were used. For booklets within an achievement level, a similar method was followed, using the scale score that was at the midpoint of the achievement level. For the Advanced level, the scale score was the midpoint between the cut score for that level, and the scale score associated with the most difficult item. For the Below Basic level, the scale score was the midpoint between the cut score for the Basic level, and the scale score associated with the easiest item.

Consequences Feedback

Consequences feedback was the percentage of students expected to perform at or above cut scores at each achievement level. The empirical distribution of student achievement based on the 2006 assessment was provided to ACT by the test development contractor in the form of the relative frequency distribution shown in the Frequency Distribution of Student Performance table in Appendix C.

Mapping Potential Exemplar Items to Achievement Levels

Potential exemplar (or sample) items in the ALS meeting were drawn from three blocks (Blocks 1, 2, and 4) that had been selected for eventual release to the public. Each score level above zero on a polytomously scored item was treated as a separate item in mapping potential exemplars to achievement levels. Each item was mapped to the lowest achievement level for which the following condition was satisfied:

$$P\left(U_{ij} > h \mid \eta_h\right) > RP, \tag{12}$$

where η_h represents the highest score value for the Basic and Proficient levels on the η scale. For the Advanced level, the score associated with the most difficult item was used.

Reliability Estimates

The term "reliability" is used here to represent the notion that cut scores from two different Achievement Level Setting meetings or from review of different but equal groups of items in the same meeting should not differ if the same method, assessment and Achievement Level Descriptions are used. Cut score reliability was evaluated by examining the standard error of the cut score. More reliable cut scores have smaller standard errors.

The median was used as the cut score in this process, and, as such, the usual standard deviation measures do not give an exact measure of the variability of the process. In general, the standard error of the median is a function of the underlying shape of the distribution of the cut scores. Since this is an unknown, estimates based on approximations are considered.

The first approximation is based on the Maritz-Jarrett procedure (Maritz & Jarrett, 1978). This procedure provides an estimated standard deviation for any percentile. If n is the number of observations and is even, then the k^{th} moment of the median is given by:

$$E[median]^k = \int x^k \binom{n}{n/2-1} \binom{n/2+1}{1} (F(x))^{n/2-1} (1-F(x))^{n/2} f(x) dx \qquad (13)$$

where $f(x)$ is the probability density function of the median, and $F(x)$ is the cumulative distribution function. A similar expression holds when n is odd. This integral can be transformed to an integral of the beta probability density function using the transformation $y = F(x)$. At the i^{th} ordered cut score, the value of y is i/n. So, the integral can be approximated as:

$$\sum_{i=1}^{n} \left(i/n\right)^k \left\{F_\beta\left(i/n, n/2, n/2+1\right) - F_\beta\left(i-1/n, n/2, n/2+1\right)\right\} \qquad (14)$$

where $F_\beta(x, \alpha_1, \alpha_2)$ is the cumulative distribution function at the point x for a beta distribution with parameters α_1 and α_2.

The second estimator of the standard deviation of the median is based on the bootstrap technique (Efron & Gong, 1983). In this procedure, repeated samples with replacement are taken from the original distribution of cut scores, and the median is calculated for each resample. The standard deviation of these medians is then calculated and used as the estimate. In this case, 1,000 samples were created.

The standard errors for these two procedures are given below. Theoretically, the estimates are only valid for the first round of cut scores, since cut scores for subsequent rounds are influenced by the location of the cut scores for the other panelists, and so are not truly independent values. Table 6 below shows the standard errors for both estimators for round 1 and round 3.

Table 6. Estimates of Standard Error of Cut Scores

Method	Basic		Proficient		Advanced	
	Round 1	Round 3	Round 1	Round 3	Round 1	Round 3
Maritz-Jarrett	1.8	1.4	3.4	0.8	4.1	2.7
Bootstrap	1.9	1.4	3.5	0.8	3.8	2.5

Additional analysis was done on the stability of cut scores across groups and panelist types. An ANOVA was done for each characteristic of interest, including gender, ethnicity, panelist type, table, and group. None of these showed any significant differences.

Process Evaluations

At the conclusion of each round and each day, a process evaluation form was provided to panelists. Panelists were asked to indicate their degree of understanding of process tasks, materials, and instructions. Results from the process evaluations were used both to clarify areas of confusion during the course of the meeting and to provide evidence of procedural validity. The responses in the process evaluations were on a 5-point Likert scale. For each item, the mean value for the responses and the standard deviation were calculated.

MATERIALS AND PROCEDURES

Information on materials and procedures used in the ALS is provided in this section. For each, a brief description of the material is given, along with an illustrative example. Additional information and descriptions of other materials can be found in the Process Report (ACT, 2007a), including:

- Agenda
- General Contents of Ordered Item Book (OIB)
- General Contents of Constructed Response Ordered Item Book (CROIB)
- Consequences Questionnaire

Briefing Book

Panelists were given a Briefing Book, which provided a broad overview of the standard setting procedures. The Briefing Book sent to panelists in advance of the ALS meeting is shown in Appendix D. Originally, the Briefing Book gave a more detailed description of the process, but was changed after the Pilot Study based on suggestions from the TACSS and the panelists.

Division of Item Pool and Panel into Pools/Groups A and B

The item pool and panelists were divided into two corresponding sets, A and B, in order to minimize the fatigue effect and the amount of time necessary if each panelist was required to review every item (186 in this case). The division also creates a design that allows the reliability of the process to be evaluated (see *Reliability* section).

There were 31 panelists in the ALS meeting. Fifteen panelists were assigned to group A, 16 to 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 and tables as equal as possible with respect to panelist type, gender, region, and race/ethnicity.

The item pool was divided into equivalent, but overlapping, pools. Each pool contained about 60% of the items in the assessment. Items included in both pools are referred to as *common items*. Equivalence was monitored with regard to: (a) item difficulty, (b) subscale representation, (c) item type representation, and (d) number of items per domain. Domains are subcategories of content created for the Mapmark with domains method used in the Pilot Study.

The equivalent pools were created in two steps: (1) assigning six blocks of items to each pool with two blocks in common, and (2) adjusting for number of items per domain. In Step 1, the common blocks are ones that have been selected for release to the public. These were blocks 2 and 4. The remaining blocks are assigned to groups to achieve the desired equivalence between pools. This is not too difficult because item blocks are generally constructed to be similar in terms of subscale representation and difficulty (see Table 1).

After the initial assignment by blocks, a few items were transferred from one group to another so that each pool would contain at least two items and at least three score points within each domain. This reassignment had very little effect on the equivalence of the pools through simple block assignment.

Table 7 presents a summary of the item pools by group and then overall. It can be seen that the item pools are equivalent as intended.

Table 7. Summary of Item Pools A and B

Group	Items	Points by Subscale[a]			Points by Item Type[b]			Item Difficulty (Scale values at RP[c] of 0.67)				
		Mkt	Natl	Intl	MC	DI	Poly	Points	Mean	SD	Min	Max
A	113	62	58	17	92	2	43	137	329	43	235	454
B	114	64	52	21	96	2	39	137	331	43	235	454
Total	186	104	89	32	154	3	68	225	331	43	235	454

[a] Mkt = Market Economy, Natl = National Economy, Intl = 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)

Test Form Administered to Panelists

Near the beginning of the ALS meeting, panelists took a form of the assessment. Table 8 presents summary information about the test form that was administered to panelists. The

form taken by panelists was composed of the two blocks, [redacted] and [redacted], that were common to item pools for both groups A and B.

Table 8. Summary Information about Test Form Taken by Panelists

Block	All Items	Number of Items: Content Area[a] Mkt	Natl	Intl	Item Type[b] MC	D1	Poly	P[c] Points
[redacted]	18	9	6	3	15	0	3	7
	18	8	7	3	15	0	3	7
Total	36	17	13	6	30	0	6	14
		18%	76%	6%	68%	0%	32%	

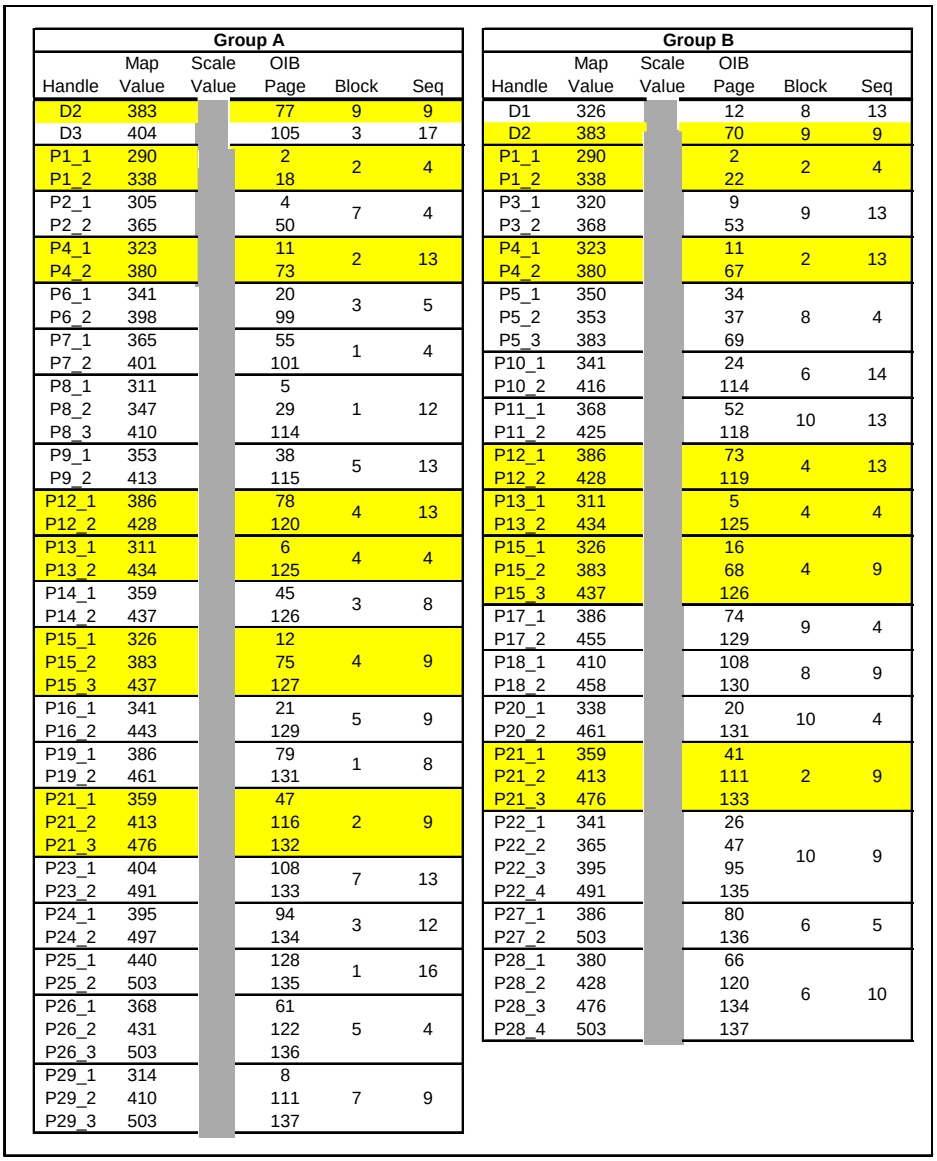
[a] Mkt = Market Economy, Natl = National Economy, Intl = International Economy
[b] MC = Multiple choice; DI = Dichotomously scored constructed response; Poly = Polytomously scored constructed response
[c] P Points = the number of score points represented by polytomously scored items

Ordered Item Book (OIB)

The Ordered Item Book (OIB) contains the items in order of their scale values, from easiest to hardest. Groups A and B have different OIBs since they have different sets of items. The actual order of items in the OIBs and the difficulty of each item on the scale is shown in Appendix B. Items are identified in this appendix by handle, map value, scale value, block, and sequence.

Constructed Response Ordered Item Book (CROIB)

The contents of the group A and B CROIBs are identified by item handles in Figure 1. Items appeared in the CROIB in the order they are listed in Figure 1. For each dichotomously scored and polytomously scored item, the CROIB contained one or more pages showing the text of the item, the scoring rubric, and one example of a student response at each score level, including 0. Items were separated by tabbed dividers with all score levels of a polytomously scored item contained within the same tab.



Group A

Handle	Map Value	Scale Value	OIB Page	Block	Seq
D2	383		77	9	9
D3	404		105	3	17
P1_1	290		2	2	4
P1_2	338		18		
P2_1	305		4	7	4
P2_2	365		50		
P4_1	323		11	2	13
P4_2	380		73		
P6_1	341		20	3	5
P6_2	398		99		
P7_1	365		55	1	4
P7_2	401		101		
P8_1	311		5		
P8_2	347		29	1	12
P8_3	410		114		
P9_1	353		38	5	13
P9_2	413		115		
P12_1	386		78	4	13
P12_2	428		120		
P13_1	311		6	4	4
P13_2	434		125		
P14_1	359		45	3	8
P14_2	437		126		
P15_1	326		12		
P15_2	383		75	4	9
P15_3	437		127		
P16_1	341		21	5	9
P16_2	443		129		
P19_1	386		79	1	8
P19_2	461		131		
P21_1	359		47		
P21_2	413		116	2	9
P21_3	476		132		
P23_1	404		108	7	13
P23_2	491		133		
P24_1	395		94	3	12
P24_2	497		134		
P25_1	440		128	1	16
P25_2	503		135		
P26_1	368		61		
P26_2	431		122	5	4
P26_3	503		136		
P29_1	314		8		
P29_2	410		111	7	9
P29_3	503		137		

Group B

Handle	Map Value	Scale Value	OIB Page	Block	Seq
D1	326		12	8	13
D2	383		70	9	9
P1_1	290		2	2	4
P1_2	338		22		
P3_1	320		9	9	13
P3_2	368		53		
P4_1	323		11	2	13
P4_2	380		67		
P5_1	350		34		
P5_2	353		37	8	4
P5_3	383		69		
P10_1	341		24	6	14
P10_2	416		114		
P11_1	368		52	10	13
P11_2	425		118		
P12_1	386		73	4	13
P12_2	428		119		
P13_1	311		5	4	4
P13_2	434		125		
P15_1	326		16		
P15_2	383		68	4	9
P15_3	437		126		
P17_1	386		74	9	4
P17_2	455		129		
P18_1	410		108	8	9
P18_2	458		130		
P20_1	338		20	10	4
P20_2	461		131		
P21_1	359		41		
P21_2	413		111	2	9
P21_3	476		133		
P22_1	341		26		
P22_2	365		47	10	9
P22_3	395		95		
P22_4	491		135		
P27_1	386		80	6	5
P27_2	503		136		
P28_1	380		66		
P28_2	428		120	6	10
P28_3	476		134		
P28_4	503		137		

Figure 1. Contents of Constructed Response Ordered Item Book by group.

The items highlighted in yellow in Figure 1 were *common items*. These items were reviewed by the whole group (groups A and B combined) in KSA Activity 1 (see Process Report, ACT, 2007a), which was led by the Mapmark content and process facilitators. In KSA Activity 2, the panelists reviewed the remaining items in their CROIB at the table group level.

KSA Notes Template

For each item score level in the CROIB, panelists recorded their notes (KSA notes) on a yellow post-it note. When they were finished with an entire item (e.g., had recorded notes on three post-it notes for a 3-point polytomously scored item), they placed their post-it notes on the KSA Notes template.

The KSA Notes template was a stapled set of 11” x 17” pages with locations designated for ten post-it notes per page. The template differed for each group according to the different items they reviewed. Figure 2 illustrates an example of a page of the KSA Notes template.

The *OIB Page* number shown in Figure 1 was printed in the CROIB on each item so that panelists could locate where on the KSA Notes template to place their yellow post-it notes. Panelists used the OIB page number for the item score level to find the appropriate location in the template for the corresponding post-it note. When panelists were finished with the CROIB, the post-it notes were attached to the template in order of the page numbers in the OIB. The post-it notes were subsequently transferred into the OIB.

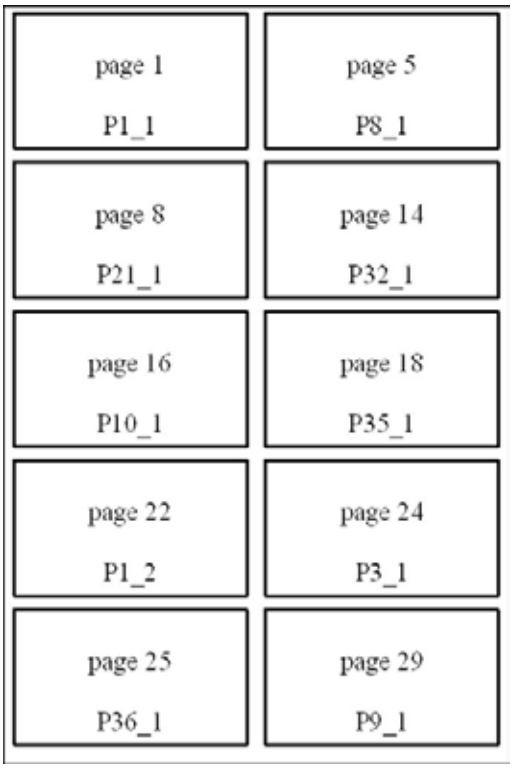


Figure 2. Sample KSA Notes template.

Cut Score Recommendation Form and Data Processing

Figure 3 shows the form that was used by panelists to record their bookmarks. In addition to the information shown in this figure, panelists’ names and IDs were printed on the form. Panelists recorded their bookmark placements and scale value selections for cut scores on this form.

In round 1, the page numbers that panelists had recorded on their Cut Score Recommendation Form for each achievement level were converted to scale values using the Scale Value to OIB Page Lookup Table shown in Appendix B. The scale values corresponding to the bookmarked page numbers were handwritten on the panelist’s Cut Score Recommendation Form, just beneath the boxes where the page numbers were recorded. (Panelists recorded these scale values on their materials in round 2.) The scale values were also entered into an Excel® spreadsheet on the same row as the panelists’ ID

number, which had been pre-entered. Once all the data were entered, the median cut scores across all panelists were computed and were reported as the cut scores for that round.

In round 2 and subsequent rounds, panelists entered scale values for their cut score recommendations on their Cut Score Recommendation Form. This form was collected and returned to panelists after each round. The scale values were entered into an Excel spreadsheet, and the median across all panelists was computed, as in round 1.

Round 1

Basic Bookmark on Page #	Proficient Bookmark on Page #	Advanced Bookmark on Page #

Round 2

Basic Cut Score at Scale Value	Proficient Cut Score at Scale Value	Advanced Cut Score at Scale Value

Round 3

Basic Cut Score at Scale Value	Proficient Cut Score at Scale Value	Advanced Cut Score at Scale Value

Figure 3. Panelist Cut Score Recommendation Form.

Item Map

In the Primary Item Map, items were organized into columns corresponding to content areas of the assessment. The map is shown in Appendix E. In the ALS meeting, the maps were printed on 8 ½" x 14" paper.

Item handles in the item maps were color coded to indicate whether they were exclusively in the group A item pool (tan), group B item pool (green), or were in both item pools (yellow).

The item handles, color code characters, and position information for the item handles in the item maps were created by a SAS® program, primary_map.sas. In the process of importing the output of the program into an Excel spreadsheet, the item handles were put into the correct cells in the map. Cells with a given color code (e.g., "G" for green) were highlighted and colored the appropriate color and the color code was removed.

Scale Value to OIB Page Lookup Table

In round 2 , panelists referred to both the Booklet Score Chart and their OIB to select a scale value for their cut score recommendation. The Booklet Score Chart shows the expected total number of points on the two test forms reviewed by each group 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. To help panelists identify what OIB page numbers corresponded to each scale value, panelists were given a Scale Value to OIB Page Lookup Table shown in Appendix B.

Consequences Feedback and Questionnaire

Consequences feedback was presented to panelists in the form of Figure 4. This display existed as an Excel® pie chart laid on top of an Excel® bar chart. The input data for the display was obtained from the Frequency Distribution of Student Performance table in Appendix C.

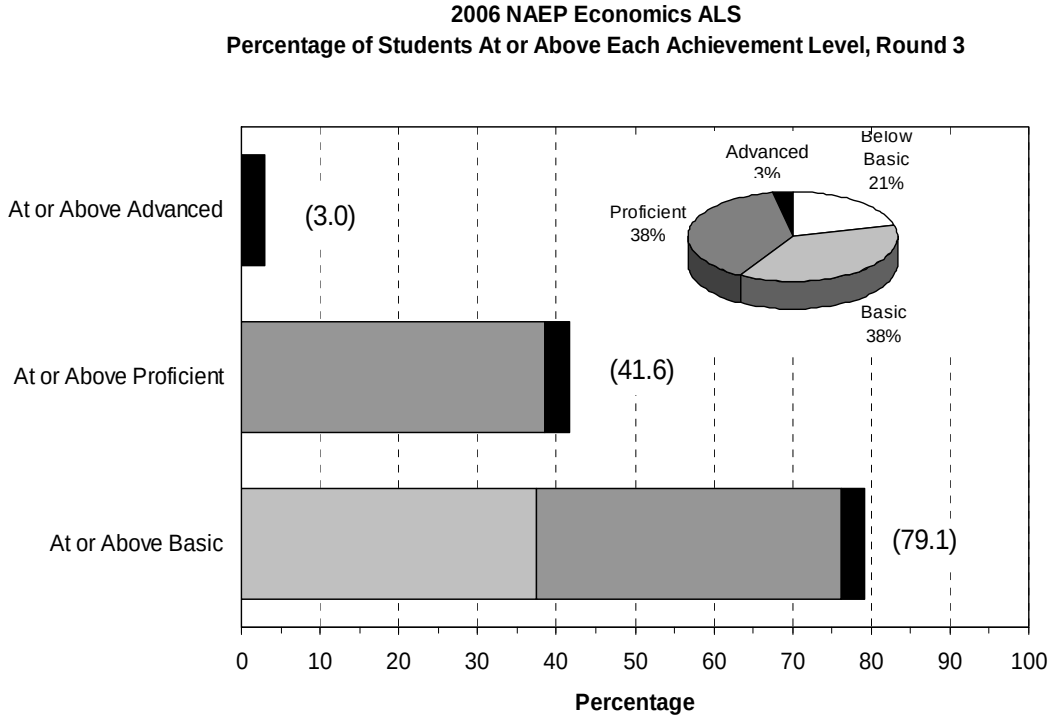


Figure 4. Consequences data.

After reviewing the final consequences data, panelists were asked to complete a consequences questionnaire indicating if they felt the proportion of students scoring at or above each level should be higher, lower, or was about right. The questionnaire is shown in the Process Report (ACT, 2007a).

Exemplar Item Rating Form

An Exemplar Item Rating Form for each achievement level was produced. First, items were classified by achievement level as explained in the Psychometric Procedures section of this report. For each group, the item handle and page number for that item in the OIB

were identified and output in a file. This file was then pasted into an Excel spreadsheet that contained formatting like that shown in Figure 5 for the Basic achievement level. The program that identified the achievement levels associated with each item used the round 3 median cut scores as input.

Item	OIB Page # Group A	OIB Page # Group B	Rating as Exemplar: Very Good	OK	Do Not 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 5. Exemplar Item Rating Form for the Basic achievement level.

FIELD TRIAL, PILOT STUDY, AND SPECIAL STUDIES

Table 9 lists the Field Trial, Pilot Study, and Special Studies that were conducted in this project. Both the Field Trial and Pilot Study involved a Mapmark method similar to that used in the ALS meeting.

Table 9. Special Studies

Study	Date	Purpose
Field Trial	October 2006	Whole Booklet Feedback development
Pilot Study	December 2006	Whole Booklet Feedback relative to Domains
Special Studies	January 2007	Consistency of results to ALS

The item statistics, materials, transformation constants, and all technical procedures used for the Mapmark with whole booklet feedback in the Pilot Study are exactly as described previously in this technical report.

In the Pilot Study, one of the methods used domain level feedback. The technical details and the materials used in domain development and the feedback rounds are described below.

Computation of Domain Characteristic Curves

For the Mapmark with domains method, subareas of content, or domains, were created. Items were classified into these domains based on content. The expected percent correct in each domain was determined at each score point on the achievement scale and was plotted on a domain characteristic curve. Let $E(U_{ij} \mid \eta)$ represent the expected score on item *I* in content area j, conditional on the composite scale score, η. Let D_k be the set of items in domain k. The expected percent correct score on a given domain, *k*, conditional on the composite score, η, was computed as:

$$\mathrm{EPC}(D_k \mid \eta) = 100\left(\frac{\sum_{j \in D_k} \sum_{i \in D_k} \mathrm{E}\left(U_{ij} \mid \eta\right)}{\sum_{j \in D_k} \sum_{i \in D_k} m_{ij}} \right), \qquad (15)$$

where $E(U_{ij} \mid \eta)$ is calculated as:

$$E\left(U_{ij} \mid \eta\right) = \int_{-\infty}^{+\infty} E\left(U_{ij} \mid \theta_j\right) f\left(\theta_j \mid \eta\right) d\theta_j \qquad (16)$$

Equation 16 was approximated by Gauss-Hermite quadrature over 40 equally-spaced points ranging from -4 to +4. Equation 15 is equivalent to taking a weighted average proportion correct score (converted to a percentage), where weights are determined by m_{ij}, the total possible score on the item.

Domain Item Map Percentages

As part of the domain development, and in the Pilot Study, a domain item map was used. This is identical to the usual item map, with the exception that the items are grouped in columns by domains. At the bottom of the map was an expected percent correct score for that domain. For the domain development, this percentage was calculated conditional on the content area scale score needed to yield 67% of expected items correct within that content area. That is, let the content area be denoted by j, let the set of items in content area j be denoted as C_j, and let η_0 be the minimum value of η such that

$$\mathrm{EPC}(C_j \mid \eta_0) = 100\left(\frac{\sum_{i \in C_j} E\left(U_{ij} \mid \eta_0\right)}{\sum_{i \in C_j} m_{ij}} \right) \geq 67 . \qquad (17)$$

The expected percent correct for domain k used in the domain item map is then

$\mathrm{EPC}(D_k \mid \eta_0)$, as defined in equation 16. For the Pilot Study domain item maps, the expected percent correct was calculated using equation 16, with the value of η being equal to the cut score for the Basic, Proficient and Advanced levels, respectively.

Percent Correct Table

After each round, the panelists were provided with a table showing the percent correct for each of the domains at that round's cut scores. An example is shown in Figure 6. The

percent correct is calculated using equation 16. In each table, the highest percentage and the lowest percentage were circled, corresponding to the easiest and hardest domains. Any percentage close to 67 was also circled, to highlight an area that was close to "mastery."

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%

Figure 6. Portion of the Percent Correct Table for Market content area, with mastered domains highlighted.

Domain Task 1 Form

Domain task 1 was designed to elicit panelist feedback on the coherence of the domains, and to begin to get them to think in terms of groups of items which cover similar topics. There were three sheets for Domain Task 1, one for each of the three content areas. Every domain within that content area is listed by title. Within each domain, all items within that domain are listed with the item handle, with the easiest items coming first. The Domain Task 1 chart for the International content area is shown in Figure 7.

Domain Task 1 – Group A

International Economy Domain	Item Handle	I see how this item is like other items in its domain. (Check ✓)		
		Yes	Not Sure	No
I1) Benefits and Costs of Trade	M14			
	M33			
	M55			
	M89			
	M90			
	M106			
	M133			
I2) Exchange Rates	M48			
	M112			
	M117			
I3) Tariffs	M97			
	M149			
	P25_2			

Figure 7. Domain Task 1 form for the International content area, Group A.

Domain Ordered Item Booklet

To assist with responding to Domain Task 1, the panelists were given a Domain Ordered Item Booklet (DOIB). The DOIB is similar to the OIB, with items listed in order, one to page. The difficulty and subsequent ordering of the items was determined as described in the Ordered Item Book section of this report. The items are separated by domain, with each domain separated by a tab. Within a domain, the order of the items was the same as on the Domain Task 1 form. Note that, unlike the OIB, the polytomous items were listed only once, using the scale value associated with achieving the highest possible score on that item. Scoring rubrics and examples of student responses were not included. A complete list of the domain titles and definitions used to construct the DOIB is included in Appendix F.

Domain Task 2 Form

The Domain Task 2 form lists the domains within a content area. The expected percent correct, as calculated from equation 17 is given for each content area, with the given scale value equal to the cut score for that achievement level. There are three sheets to the form, one for each of the three achievement levels. Figure 8 shows the Domain Task 2 form for the Proficient level.

Economics Pilot Study
Domain Task 2
Borderline PROFICIENT

Content Area	Domain	Expected Percent Correct	I think the percentage correct score at the **lower borderline of PROFICIENT** should be... (check the appropriate cell)		
			Lower	**OK**	**Higher**
Market	M1. Entrepreneurs	58%			
	M2. Incentives	53%			
	M3. Markets and Equilibruim	57%			
	M4. Productivity, Income, and Capital	49%			
	M5. Scarcity and Opportunity Cost	49%			
	M6. Competition	40%			
	M7. Economic Institutions	39%			
	M8. Internation of Supply, Demand, and Prices	38%			
	M9. Economic Role of Government	33%			
	M10. Additional Costs and Benefits in Decision Making	31%			
National	N1. Money, Loans, and Interest Rates	56%			
	N2. Spending, Income, and Related National Measures	41%			
	N3. Resource Allocation	42%			
	N4. Economics Growth and Productivity	38%			
	N5. Government Programs and Taxes	38%			
	N6. Real Interest Rates	24%			
	N7. Inflation and Unemployment	27%			
	N8. Money Supply	29%			
	N9. Fiscal and Monetary Policy	21%			
International	I1. Benefits and Costs of Trade	39%			
	I2. Exchange Rates	35%			
	I3. Tariffs	24%			

Figure 8. Domain Task 2 form for the Proficient level.

Domain Score Chart

The Domain Score Chart was a 3 page form, one for each of the three achievement levels. For each level, the scale scores were listed down the left-hand side, and the EPC, as given in Equation 17, are listed for each domain, conditioned on the level cut score. The score scale ranges from 10 points below the lowest recommended cut score to 10 points above the highest recommended cut score. The median value is highlighted. Within each domain, the circles were manually added to the places where the EPC was 67. The panelists were then instructed to circle their cut score on the chart in the left hand column. Figure 9 shows the Domain Score Chart for Proficient at round 2.

Domain Score Plots

Domain Score Plots were used to help the panelists visualize the differences between performances on domains. These plots are just smoothed versions of the EPC values shown across the score scale. These plots were not given to the panelists, but shown to them during the presentations. Various versions of these plots were used to call attention to different topics of discussion. Figure 9 shows one of the plots used in the presentation. Note that vertical lines are drawn at the cut scores, and a horizontal line at the RP value is drawn to draw attention to the level needed for Mastery.

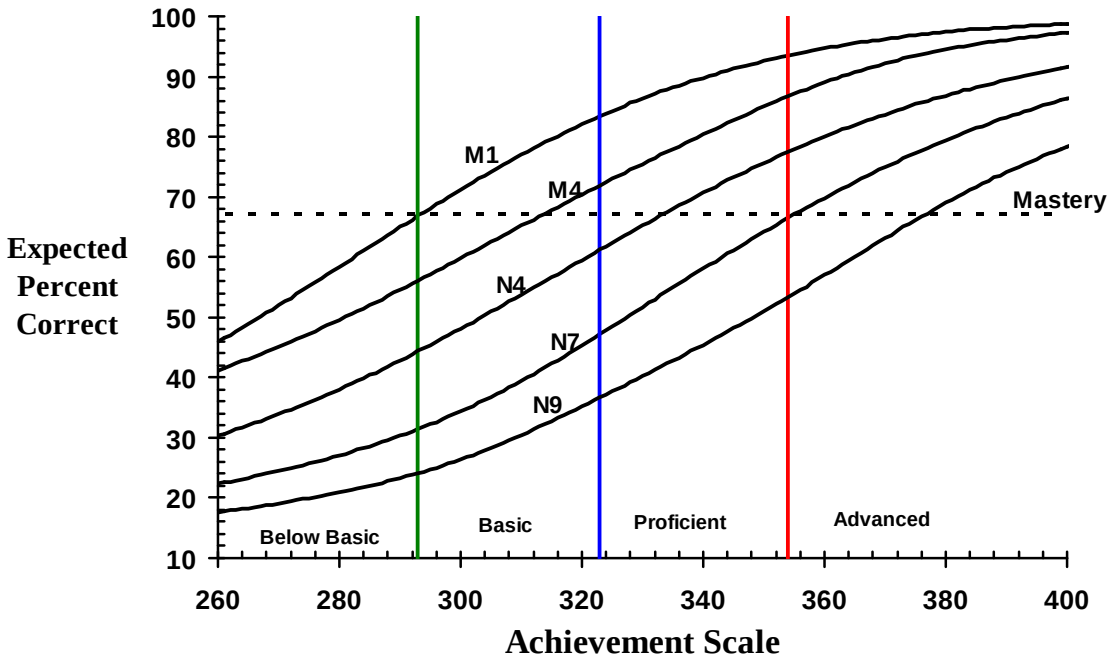


Figure 9. Domain Score Plot for a set of domains.

Special Studies

The Special Studies consisted of a Booklet Classification Study and an Item Classification Study. A description of the studies and results for the studies are given in the Special Studies Report (ACT, 2007b). In the Booklet Classification Study, panelists were asked to classify booklets into an achievement level, based on a holistic judgment of the work on the assessment. The empirical level of the booklet is defined by the scale value that would give the Expected Number Correct equal to the observed number of score points correct, as given by equation 11. In the item classification study, panelists were asked to classify items into an achievement level, using the RP criterion. Using equation 12, the empirical level of an item was established using the first level where the RP criterion was exceeded at one of the scale scores within that level.

For each study, the percentage of agreement between panelist classification and empirical classification was calculated. This was done for both the individual panelists' classifications, and for the median of the panelists' classifications.

Process Evaluations were done after each study, to ascertain the understanding of the tasks and materials used in the studies. A five point Likert scale was used for each question, and the results were summarized by calculating the mean and standard deviation for each question, using 1 to 5 as the response values.

COMPUTER PROGRAMS

A large number of computer programs were developed over the course of the project. The following is a summary of programs that contained essential psychometric algorithms and/or produced key results for materials and data displays. Programs containing FORTRAN source code are named using the extension *.for,* but the executable versions have the extension *.exe*.

Programs for Mapmark

A FORTRAN program **naepg12.for** computes item score probabilities conditionally on subscale thetas and regresses these onto the composite score scale. Two input files are needed:

- *naep12.cc* contains the mean and standard deviation of student achievement on the η scale, transformation constants (Table 2**Error! Reference source not found.**), and subscale correlations (Table 3).
- *g12_irt_info* contains NAEP ID, block, sequence, subscale, item type, and item parameter estimates. (See Item Statistics in Appendix B.)

Four output files are created:

- *naepg12out1.out* contains, for each score point in the assessment (N = 225), the cumulative probability given in Equation 6 conditional on the corresponding subscale theta, θ_j, for values of θ_j that are obtained by applying the inverse of

Equation 3 to y_j = 0, 1, …, 300. Only the y_j values used for the conditioning are reported in the file.

- *naepg12out2.out* contains, for each score point in the assessment, the cumulative probability given in Equation 5, conditionally on values of η = 0, 1, …, 300.
- *naepg12out3.out* contains, for each item in the assessment (N=186), the expected score (item true score), as defined in Equation 15 conditionally on the corresponding subscale theta, θ_j, for values of θ_j that are obtained by applying the inverse of Equation 3 to y_j = 0, 1, …, 300. Only the y_j values used for the conditioning are reported in the file.

$$E\left(U_{ij} \mid \theta_j\right) = \sum_{h=0}^{m_{ij}} hP\left(U_{ij} = h \mid \theta_j\right) \tag{16}$$

- *naepg12out4.out* contains, for each item in the assessment, the expected item true score as defined in Equation 16 conditionally on values of η = 0, 1, …, 300.

The program **Mapmark1.sas** collates item information from various sources and creates a SAS data set, Set9, which is used as input to other SAS programs. Input to the program includes the following files:

- *g12_economics_irt* is a file of item statistics received from the test development contractor, essentially like the Item Statistics table in Appendix B.
- *content.prn* is a file that contains NAEP item identification and classifications of items into the assessment framework.
- *naepg12out2.out* is one of the output files from naepg12.for with the first two lines removed (see above).

Besides the SAS data set, Set9, one output file is produced:

- *labels* is a list containing information that will be printed on the label for each item in the OIB, and CROIB. The information includes the item's handle, content area, map value, scale value, complexity classification, block, and sequence number.

The program **Mapmark2.sas** uses Set9, the SAS file created by **Mapmark1.sas**. It produces output for assembling most of the materials for Mapmark including the Ordered Item Book and the Constructed Response Ordered Item Book:

- *groupa_all.txt* is a list for assembling the group A OIB containing page number, item handle, item map value, item scale value, block, and sequence within block.
- *groupa_cr1.txt* is a list for assembling the group A CROIB.
- *groupa_cr2.txt* is a list for creating the KSA Note Template for group A.

Additional output files include files for group B materials corresponding to those described for group A. These files have *groupb* in their name.

The program **primary_map.sas** also uses Set9 from **Mapmark1.sas**. It produces the output file *primary.map*. The file is incorporated into Excel® spreadsheets to create the item maps.

The program **mapmark-exemplars.sas** uses Set9 from **Mapmark1.sas**, plus the input file *naepg12out2.out* to create a file, *mapmark-exemplars1.out*, that is used as input to the program **exemplar-mapmark.for** (see below).

The program **exemplar-mapmark.for** is used to map potential exemplar items to achievement levels using the method described earlier in this report with reference to Equation 12, and to produce output for creating the Exemplar Item Rating Form. Three input files are needed:

- *mapmark-exemplars1.out* contains item handle, block, sequence, page number for groups A and B, and the conditional probability. This file was generated by the SAS program **mapmark_exemplars.sas** (see above).
- *pctatabove.txt* contains percent of students at or above each scale score.
- *cutscores.txt* contains final cut scores for each achievement level. The file name needs to be provided when running the executable file.

REFERENCES

ACT (2005a). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade twelve mathematics: Domain development report.* Iowa City, IA: Author.

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

ACT (2005c). *Developing achievement levels on the 2005 National Assessment of Educational Progress in grade twelve mathematics: Special studies report.* Iowa City, IA: Author.

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

ACT (2007a). *Developing achievement levels on the 2006 National Assessment of Educational Progress in grade twelve economics: Process report.* Iowa City, IA: Author.

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

Donoghue, J. R. (March, 1997). *Item mapping to a weighted composite scale.* Paper presented at the annual meeting of the American Educational Research Association, Chicago.

Efron, B. & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, Vol. 37, No. 1, pp. 36-48.

Maritz, J. S. & Jarrett, R. G. (1978). A note on estimating the variance of the sample median. *Journal of the American Statistical Association*, Vol. 73, No. 361, pp. 194-196.