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

Technical Report

Presented by ACT, Inc.
May 11, 2005

Redacted by Governing Board

The National Assessment Governing Board

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

Technical Report

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

Technical Report
Table of Contents

APPENDICES

List of Figures

List of Tables

INTRODUCTION

In September, 2003, the National Assessment Governing Board (NAGB) contracted with ACT to conduct research and other activities for setting achievement levels on the 2005 National Assessment of Educational Progress in Grade 12 Mathematics. 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 primarily concerned with the materials and process of the achievement level setting (ALS) meeting that was held in November 2004. The data used in the meeting consisted of items, item statistics, and estimates of student achievement from the 2004 field test of the 2005 NAEP in Grade 12 Mathematics. The methodology used to set the achievement levels was Mapmark, a bookmark-based procedure that includes item maps and domain score feedback.

This report also addresses issues that are informed by results from special studies ACT conducted in the course of the project. Mapmark was developed and evaluated through special studies conducted prior to the ALS meeting. Technical issues concerning outcomes of the ALS meeting, such as the reliability of cut scores obtained in the ALS meeting, and the statistical criteria for pairing exemplar items with achievement levels, are informed by the results from these meetings.

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

1) The *Process Report* provides a detailed description of the process and outcomes of the ALS meeting. It also contains an overview of the entire project, including key results and conclusions from the special studies and a description of the domains that were developed in this project for use in the ALS meeting.
2) The *Special Studies Report* provides a description of the purpose, methods, materials, results, and conclusions of the special studies conducted in this project. The special studies allowed the Mapmark method to be developed and evaluated. They included a Pilot Study in which Mapmark was compared to an Angoff-based methodology using the same data that were used in the ALS meeting.
3) The *Domain Development Report* describes how the domains that were used in the ALS meeting were developed.

The Technical Report also accounts for technical advice ACT received throughout this project. ACT relied on the advice of a Technical Advisory Committee on Standard Setting (TACSS). The TACSS is a five-member group that collectively represents expertise in standard setting, mathematics education, and experience with the NAEP. The TACSS met six times over the course of the project and provided technical advice concerning all aspects of the project. TACSS input is presented in the form of meeting minutes in Appendix A of this report. TACSS input on issues is also described in this and other reports described above.

PSYCHOMETRIC PROCEDURES

Description of Item Pool

The Achievement Level Setting (ALS) meeting used items, item statistics, and student performance data from the 2004 field test of the 2005 NAEP in Grade 12 Mathematics.

Table 1 presents a summary of items used in the ALS meeting by block. The items were organized into ten blocks, labeled 3 through 12. The number of items per block ranged from 17 to 21. There were a total of 180 items of which 119 were multiple choice (M), 24 were dichotomously-scored constructed response (D), and 37 were polytomously-scored constructed response (P). The polytomously-scored items represented a total of 95 score points, or 40% of the points in the item pool. Dichotomously-scored items represented 10% of the points, and multiple choice items represented 50% of the points. The total number of points was 237.

Table 1 shows how the items were distributed by content area. Items are tabulated separately for Measurement (Meas) and Geometry (Geo), but these two content areas were combined into a single subscale (Measurement/Geometry) in the achievement level descriptions, the construction of the student achievement scale, and in most of the activities of the ALS meeting.

Table 1. Summary of Item Pool by Block

Block	Number of Items with Item-Statistics										Student Performance (Percent Correct on Items)			
	All	Content Area					Item Type			P				
		Num	Meas	Geo	Data	Alg	M	D	P	Points	Mean	Std Dev	Min	Max
3	17	3	4	2	4	4	12	1	4	8	48	20	24	82
4	17	3	2	3	5	4	11	2	4	12	46	20	10	84
5	18	4	2	5	3	4	11	3	4	12	47	21	6	81
6	17	2	3	3	4	5	14	1	2	4	41	22	12	82
7	18	2	3	3	3	7	13	0	5	11	47	19	10	81
8	21	5	2	5	4	5	10	4	7	18	49	30	2	95
9	18	6	3	3	2	4	7	5	6	13	41	22	5	84
10	18	3	2	2	4	7	11	5	2	6	46	22	5	87
11	18	2	3	3	4	6	15	2	1	5	47	19	16	75
12	18	2	3	3	4	6	15	1	2	6	45	21	11	84
	180	**32**	**27**	**32**	**37**	**52**	**119**	**24**	**37**	**95**	**45.8**	**22**	**2**	**95**
		18%	15%	18%	21%	29%	50%	10%		40%				

Note: Num = Number Properties and Operations; Meas = Measurement; Geo = Geometry; Data = Data Analysis and Probability; Alg=Algebra and Functions; M= Multiple Choice; D=Dichotomously-scored constructed response; P=Polytomously scored constructed response; P Points = the number of score points represented by polytomously-scored items.

Student performance data are indicated in the last panel of Table 1. The percent correct for multiple choice and dichotomously-scored items is the percent of students who answered the item correctly. The percent correct on polytomously-scored items is the mean item score (over all students) divided by the total possible score, times 100. It can be seen that the items were difficult, on average, with a mean percent correct score of less than 50% over all items.

There were 185 numbered items in the test booklets, but only 180 items with item statistics. The difference in counts is attributed to the following scoring changes:

- Three multiple choice items within Block 9 (Items 4, 5, and 6) were treated as a single, three-point polytomously-scored item.
- Items 6 and 7 within Block 4 were treated as a single, four-point polytomously-scored item.
- Item 6 within Block 6 was not scored.
- Item 17 (a multiple choice item) within Block 5 was combined with Item 18 to make a three-point, polytomously-scored item.

Field Trial Statistics

Item statistics and student achievement distributions used in the ALS meeting were based on the 2004 field test. Materials were created for the Pilot Study in July (see Special Studies Report) using a preliminary set of statistics that was delivered to ACT by the testing contractor in June. A second set of statistics, which involved non-response adjustments, was delivered after the Pilot Study.

Materials were not recreated for the ALS meeting using the second set of statistics because the second set of statistics was very similar to the first set and would have had no effect on the ALS meeting outcomes.

Ultimately, results from the ALS meeting will have to be translated to the 2005 NAEP scale by making adjustments for differences between statistics from the 2004 field test and statistics from the operational 2005 assessment.

Computation of Item Scale Values

Each item in the assessment is calibrated exclusively to one of the four subscales shown in Table 2. The computation of item scale values in the Mapmark procedure begins with the computation of score probabilities conditional on the subscales. 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:

$$P(U_{ij} \quad 1|\theta_j) \quad p_{ij} \quad c_{ij}+\frac{1-c_{ij}}{1+\exp[-Da_{ij}(\theta_j-b_{ij})]}, \tag{1}$$

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:

$$P(U_{ij} \quad h|\theta_j) \quad p_{ijh} \quad \frac{\exp\left[\sum_{r\ 1}^{k} Da_{ij}(\theta_j-b_{ij}+d_{ijr})]\right]}{\sum_{h\ 1}^{m_{ij}}\exp\left[\sum_{r\ 1}^{h} Da_{ij}(\theta_j-b_{ij}+d_{ijr})\right]}, \tag{2}$$

where m_{ij} is the maximum score on the item, and d_{ijr} is the threshold parameter for score *r, r=1,2,…,m*$_{ij}$.

The composite scale score, η, is related to subscale thetas, $\boldsymbol{\theta} \quad \{\theta_1,...,\theta_4\}$, through the transformations:

$$\boldsymbol{y} \quad \boldsymbol{A\theta} + \boldsymbol{b} \tag{3}$$

and

$$\eta \quad \boldsymbol{w}^t \boldsymbol{y}\,, \tag{4}$$

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 preliminary transformation constants used to create the composite score scale used in the ALS meeting.

Table 2. Transformation Constants and Weights to Form Composite

Subscale Notation (*j*)	Subscale	Slope (diag ***A***)	Intercept (***b***)	Weight (***w***)
1	Number Properties	37.943	148.845	0.10
2	Measurement & Geometry	35.470	150.075	0.30
3	Data Analysis	34.765	150.371	0.25
4	Algebra	33.211	151.514	0.35

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 by summing over η = 0, 1,…, 300.

$$P(U_{ij} \geq h \mid \eta) \quad \int_{-\infty}^{+\infty} P(U_{ij} \geq h \mid \theta_j) f(\theta_j \mid \eta) d\theta_j\,. \tag{5}$$

where

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

$$f(\theta_j \mid \eta) \quad N(\mu_j + \frac{\sigma_j \rho_{j\eta}(\eta - \mu_\eta)}{\sigma_\eta}, \sigma_j^2(1-\rho_{j\eta}^2))\,, \tag{7}$$

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

and

$$Cov(\theta_j, \eta) \quad \sum_{k \quad 1}^{4} w_k Cov(\theta_j, \theta_k) \quad \sum_{k \quad 1}^{4} A_{kk} \rho_{jk} \sigma_j \sigma_k\,. \tag{9}$$

The correlations between subscale thetas (ρ_{jk}) based on the preliminary field trial 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 and σ_k) 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, 150.0001 and 34.1547.

Table 3. Marginal Subscale Theta Distributions

	Theta	
Subscale (j)	Mean (μ_j)	S.D. (σ_j)
1	0.0304	0.9224
2	-0.0021	0.9868
3	-0.0107	1.0068
4	-0.0456	1.0539

Table 4. Subscale Correlations

Subscale	1	2	3	4
1	1.0000			
2	0.9402	1.0000		
3	0.9369	0.9233	1.0000	
4	0.9263	0.9482	0.9236	1.0000

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:

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

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 η=300, then η_{ijh} was set to 301.

For the achievement level setting meeting, 100 was added to the item scale value obtained as described with reference to Equation 10. This addition produced item scale values ranging from 141 to 401. There was no item for which the conditional probability was 0.67 or higher at scale values less than 141. Item scale values on the Mapmark scale (141 to 401) are shown in the “Item Handles and Teacher Domain Assignments” section of Appendix C.

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 M119. The 24 dichotomously-scored item handles were numbered D1 to D24.

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—203 to 355—and are, therefore, located in the rows or categories on the item map labeled “above” or “below.”

The item handle for a score on a polytomously-scored item shows the score that is being represented specifically, and also shows the easiness-rank of the highest possible score on the item. For example, the handle P2_4 represents a score of “4” on item P2. The “2” in this handle means that this polytomously-scored item is the second-easiest among all polytomously-scored items, in terms of earning full credit on the item. More precisely, item “P2” is the second-easiest polytomously-scored item in terms of the level of achievement (scale value of 264) that corresponds to a 0.67 probability of earning full credit on the item (a score of 4). As shown in Table 5, each score level of Item P2, 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.

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 fifty-one score intervals, plus two extreme catch-all categories labeled “above” and “below.” The score intervals were three units wide and represented scale scores ranging from 203 to 355. [The interval midpoints ranged from 204 to 354 in steps of 3.] Items with scale values outside this range were represented in the “above” or “below” categories. Of the 237 item scale values, six were represented as “below” 203 and six were represented as “above” 355.

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

Item Type	Handle	Scale Value	Map Value
Multiple Choice	M119	380	above
	M118	375	above
	M117	357	357
	M116	336	336
	.	.	.
	.	.	.
	M4	205	204
	M3	201	below
	M2	160	below
	M1	159	below
Dichotomously Scored	D24	356	above
	D23	333	333
	.	.	.
	.	.	.
	D1	223	222
Polytomously Scored	P37_4	off scale	above
	P37_3	341	342
	P37_2	313	312
	P37_1	233	234
	.	.	.
	.	.	.
	P2_4	264	264
	P2_3	222	222
	P2_2	193	below
	P2_1	157	below
	P1_2	245	246
	P1_1	141	below

Computation of Domain Characteristic Curves

Let $\mathrm{E}(U_{ij} \mid \eta)$ represent the expected score on item i within strand j, conditional on the composite scale score, η. The expected percent correct score on a given domain, k, conditional on the composite score, η, was computed as:

$$\mathrm{EPC}(D_k \mid \eta) \quad 100\left(\frac{\sum_{j \in D_k} \sum_{i \in D_k} \mathrm{E}\left(U_{ij} \mid \eta\right)}{\sum_{j \in D_k} \sum_{i \in D_k} m_{ij}} \right). \qquad (11)$$

where

$$E(U_{ij} \mid \eta) \quad \int_{-\infty}^{+\infty} E(U_{ij} \mid \theta_j) f(\theta_j \mid \eta) d\theta_j \,, \tag{12}$$

$$E(U_{ij} \mid \theta_j) \quad \sum_{h \quad 0}^{m_{ij}} hP(U_{ij} \quad h \mid \theta_j) \,, \tag{13}$$

and Equation 12 was approximated by Gauss-hermite quadriture over 40 equally-spaced points ranging from -4 to +4. Equation 11 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.

The association of items with teacher domains is shown in Appendix C. EPC scores were computed for “score domains,” which consisted of one or more teacher domains, as shown in Table 6.

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 2004 field test was provided to ACT by the test development contractor in the form of the relative frequency distribution shown in the first three columns of the Frequency Distribution of Student Performance table in Appendix C. The additional columns in Appendix C were created to facilitate the production of consequences data displays during the ALS meeting.

Mapping Potential Exemplar Items to Achievement Levels

Potential exemplar items in the ALS meeting were drawn from three blocks (Blocks 3, 4, and 12) 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:

$$P(U_{ij} \geq h \mid \eta_{mp}) \geq RP \,, \tag{14}$$

where η_{mp} stands for the midpoint (for Basic and Proficient) or median (for Advanced) value of the achievement level on the η scale. For Basic and Proficient levels, η_{mp} is defined as:

$$\eta_{mp} \quad \eta_L + \frac{\eta_U - \eta_L}{2} \,. \tag{15}$$

Table 6. Titles of Teacher Domains and the Correspondence between Teacher and Score Domains by Subscale of the 2005 Assessment

Number Properties and Operations

Teacher Domain	Title	Score Domain
N1	Perform Basic Operations	N--1
N2	Determine Correct Operations	N--2
N3	Place Value and Notation	N--3
N4	Multistep Problems	N--4

Measurement/Geometry

M1	Basic Measurement	M--1
M2	Symmetry, Motion, and Proportionality	M--2
M3	Identifying Geometric Objects	
M4	Angles	M--3
M5	Perimeter, Area, and Volume	
M6	Coordinates and Their Applications	M--4
M7	Triangle Properties and Measurements	
M8	Geometric Relationships	M--5

Data Analysis

D1	Common Data Displays	D--1
D2	Elementary Probability and Sampling	D--2
D3	Central Tendency	D--3
D4	Advanced Data Displays	
D5	Abstract Reasoning	D--4

Algebra

A1	Reading Tables and Graphs	A--1
A2	Algebraic Expressions, Equations, and Inequalities	
A3	Systems of Equations	A--2
A4	Slope and Rates	
A5	Creating and Recognizing Expressions	A--3
A6	Advanced Functions and Concepts	

Where η_L is the cut score for the achievement level and η_U is the cut score for the next higher achievement level. When η_{mp} had a decimal value at 0.5, it was rounded up to the next integer.

For the Advanced level, η_{mp} was the median η value among students in the Advanced achievement level:

$$P[(\eta_{Advanced} \leq x \leq \eta_{mp}) \mid x \geq \eta_{Advanced}] \approx 0.5, \qquad (16)$$

where x stands for a composite scale score and $\eta_{Advanced}$ is the Advanced cut score. Because the η scale is discrete, η_{mp} was obtained by choosing the value that made the left hand side of Equation 16 closest to 0.5.

MATERIALS AND PROCEDURES

The Process Report provides illustrations and descriptions of most of the materials and forms used in the ALS meeting, including:

- Agenda
- General Contents of Ordered Item Book (OIB)
- General Contents of Constructed-Response Ordered Item Book (CROIB)
- Domain Task 1 Form
- Domain Task 2 Form
- Consequences Questionnaire

In addition, the Process Report contains a general description of the division of the item pool into two pools, A and B, to which panelists' Groups A and B were assigned. Tables in the Process Report document the equivalence of item pools with regard to number of each item type, total number of score points, and difficulty and location of items on item maps in terms of item scale values.

More specific information on materials is provided in this section.

Briefing Book

The Briefing Book sent to panelists in advance of the ALS meeting is shown in Appendix B.

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

The item pool and panelists were divided into two corresponding sets, A and B, because there are too many items in the assessment (180 in this case) for any one panelist to review, and the division creates a design that allows the reliability of the process to be evaluated. (See Reliability section below.)

There were thirty-one panelists in the ALS meeting. Fifteen panelists were assigned to Group A, sixteen 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 in both pools are referred to as "common items." Equivalence was monitored with regard to: 1) item difficulty, 2) subscale representation, 3) item type representation, and 4) number of items per Teacher Domain.

The equivalent pools were created in two steps: 1) assigning six blocks of items to each pool with two blocks in common, and 2) adjustments for number of items per Teacher Domain. In Step 1, the common blocks are generally the ones that have been selected for eventual release to the public. These were blocks 3 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 term 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 teacher 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. It can be seen that the item pools are equivalent as intended. More detailed comparisons of the item pools with regard to difficulty and representation of subscales and item types is presented in Appendix E of the Process Report. This information is presented by item block within pool, and then aggregated over each pool. The representation of Teacher Domains by each item pool can be seen from the color coding of the item handles in the Domain Item Maps which are presented in Appendix D of this Technical Report.

Table 7. Summary of Item Pools A and B

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

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

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

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

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

Test Forms Administered to Panelists

Near the beginning of the ALS meeting, panelists take a form of the assessment. Each group of panelists takes a different test form. The selection of the forms to administer to panelists is guided by two general considerations:

1. The forms should be as equivalent as possible in terms of difficulty and proportional representation of item types.
2. Panelists should take a form that does not contain items in their item pool. This is to prevent some items (those taken by the panelist) from having more influence in the process than other items in the panelist's pool.

Table 8. Summary Information about Test Forms Taken by Panelists

Group A:

Block	Number of Items										Student Performance (Percent Correct on Items)			
	N	Content Area					Item Type			P				
	Items	Num	Meas	Geo	Data	Alg	MC	Dich	Poly	Points	Mean	Std Dev	Min	Max
	21	5	2	5	4	5	10	4	7	18	49	30	2	95
	18	3	2	2	4	7	11	5	2	6	46	22	5	87
	39	**8**	**4**	**7**	**8**	**12**	**21**	**9**	**9**	**24**	**47.6**	**27**	**2**	**95**
		21%	10%	18%	21%	31%	39%	17%		44%				

Group B: Form M

Block	Number of Items										Student Performance (Percent Correct on Items)			
	N	Content Area					Item Type			P				
	Items	Num	Meas	Geo	Data	Alg	MC	Dich	Poly	Points	Mean	Std Dev	Min	Max
	18	4	2	5	3	4	11	3	4	12	47	21	6	81
	18	6	3	3	2	4	7	5	6	13	41	22	5	84
	36	**10**	**5**	**8**	**5**	**8**	**18**	**8**	**10**	**25**	**44.0**	**22**	**5**	**84**
		28%	14%	22%	14%	22%	35%	16%		49%				

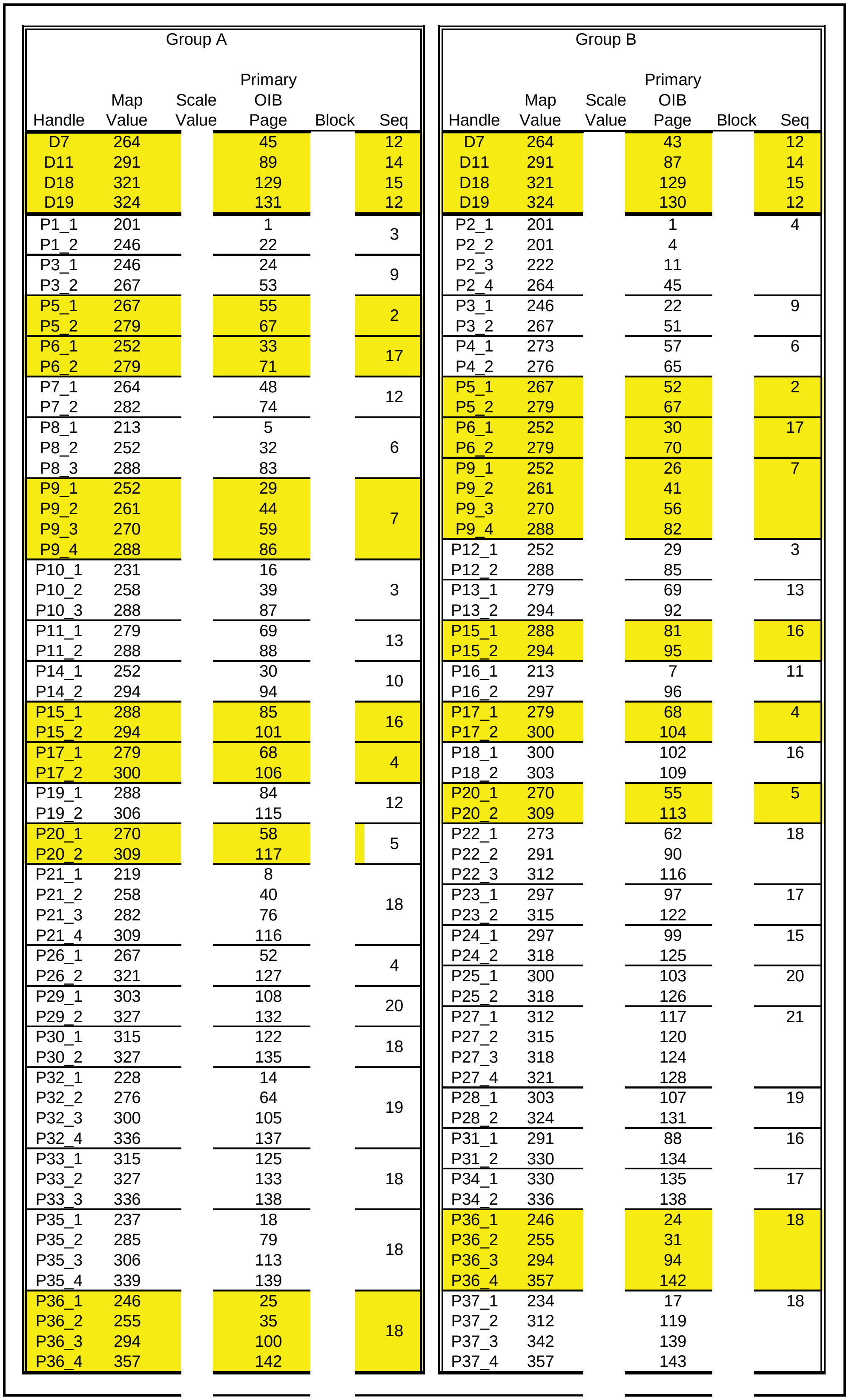
Table 8 presents summary information about the test forms that were administered to panelists. Group A was administered Form M , Group B Form M . The at the end of the Form IDs indicates that both forms allowed students to use a . As shown in the table, the forms were very similar in difficulty and in the proportional representation of item types. Also, the forms taken by panelists included blocks that were not in their item pool.

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 is shown in Appendix C. 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 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 tabs with all score levels of a polytomously-scored item contained within the same tab.



Group A

Handle	Map Value	Scale Value	Primary OIB Page	Block	Seq
D7	264		45		12
D11	291		89		14
D18	321		129		15
D19	324		131		12
P1_1	201		1		3
P1_2	246		22		
P3_1	246		24		9
P3_2	267		53		
P5_1	267		55		2
P5_2	279		67		
P6_1	252		33		17
P6_2	279		71		
P7_1	264		48		12
P7_2	282		74		
P8_1	213		5		6
P8_2	252		32		
P8_3	288		83		
P9_1	252		29		7
P9_2	261		44		
P9_3	270		59		
P9_4	288		86		
P10_1	231		16		3
P10_2	258		39		
P10_3	288		87		
P11_1	279		69		13
P11_2	288		88		
P14_1	252		30		10
P14_2	294		94		
P15_1	288		85		16
P15_2	294		101		
P17_1	279		68		4
P17_2	300		106		
P19_1	288		84		12
P19_2	306		115		
P20_1	270		58		5
P20_2	309		117		
P21_1	219		8		18
P21_2	258		40		
P21_3	282		76		
P21_4	309		116		
P26_1	267		52		4
P26_2	321		127		
P29_1	303		108		20
P29_2	327		132		
P30_1	315		122		18
P30_2	327		135		
P32_1	228		14		19
P32_2	276		64		
P32_3	300		105		
P32_4	336		137		
P33_1	315		125		18
P33_2	327		133		
P33_3	336		138		
P35_1	237		18		18
P35_2	285		79		
P35_3	306		113		
P35_4	339		139		
P36_1	246		25		18
P36_2	255		35		
P36_3	294		100		
P36_4	357		142		

Group B

Handle	Map Value	Scale Value	Primary OIB Page	Block	Seq
D7	264		43		12
D11	291		87		14
D18	321		129		15
D19	324		130		12
P2_1	201		1		4
P2_2	201		4		
P2_3	222		11		
P2_4	264		45		
P3_1	246		22		9
P3_2	267		51		
P4_1	273		57		6
P4_2	276		65		
P5_1	267		52		2
P5_2	279		67		
P6_1	252		30		17
P6_2	279		70		
P9_1	252		26		7
P9_2	261		41		
P9_3	270		56		
P9_4	288		82		
P12_1	252		29		3
P12_2	288		85		
P13_1	279		69		13
P13_2	294		92		
P15_1	288		81		16
P15_2	294		95		
P16_1	213		7		11
P16_2	297		96		
P17_1	279		68		4
P17_2	300		104		
P18_1	300		102		16
P18_2	303		109		
P20_1	270		55		5
P20_2	309		113		
P22_1	273		62		18
P22_2	291		90		
P22_3	312		116		
P23_1	297		97		17
P23_2	315		122		
P24_1	297		99		15
P24_2	318		125		
P25_1	300		103		20
P25_2	318		126		
P27_1	312		117		21
P27_2	315		120		
P27_3	318		124		
P27_4	321		128		
P28_1	303		107		19
P28_2	324		131		
P31_1	291		88		16
P31_2	330		134		
P34_1	330		135		17
P34_2	336		138		
P36_1	246		24		18
P36_2	255		31		
P36_3	294		94		
P36_4	357		142		
P37_1	234		17		18
P37_2	312		119		
P37_3	342		139		
P37_4	357		143		

Figure 1. Con ts of Con cted Response Orde Item Bo y group.

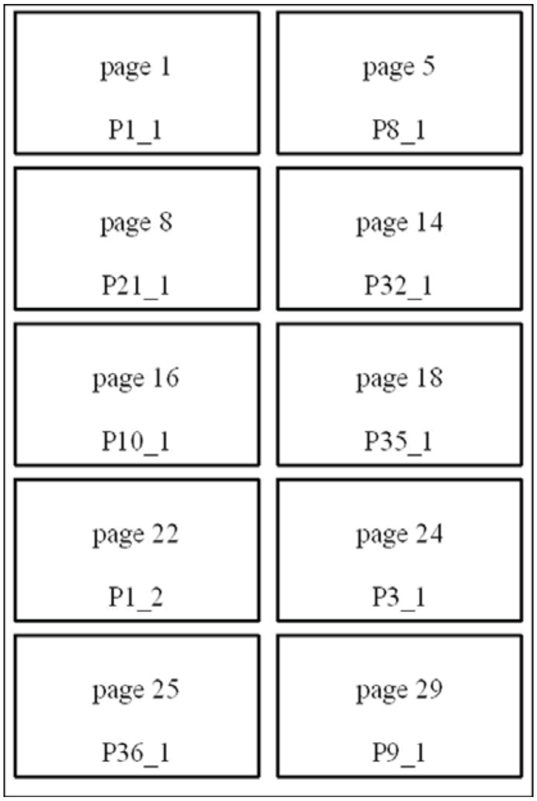
The items highlighted in yellow in Figure 1 were “common items.” These items were reviewed by the whole group (Groups A and B combined) in KSA Activity 1 (see Process Report) which was led by the Mapmark content facilitator. In KSA Activity 2, the panelists reviewed the remaining items in their CROIB at the table group level.

KSA Note Template

For each item score level in the CROIB, panelists recorded their notes (KSA notes) on a yellow stickie. When they were finished with an entire item (e.g., had recorded notes on 3 stickies for a 3-point polytomously-scored item) they placed their stickies into the KSA Note Template.

The KSA Note Template was a stapled set of tabloid-size pages with locations designated for 10 stickies per page. The template differed for each group according to the different items they reviewed. Figure 2 illustrates the first page of the Group A KSA Note Template.

The “Primary OIB Page” number shown in Figure 1 was printed in the CROIB on each item so that panelists could locate where on the KSA Note Template to place their yellow stickies. Panelists used the Primary OIB page number for the item score level to find the appropriate location in the template for the corresponding stickie. When panelists were finished with the CRIOB, the stickies were attached to the template in order of the page numbers in the OIB. The stickies were transferred into the OIB with one pass through the template and OIB.



page 1 P1_1	page 5 P8_1
page 8 P21_1	page 14 P32_1
page 16 P10_1	page 18 P35_1
page 22 P1_2	page 24 P3_1
page 25 P36_1	page 29 P9_1

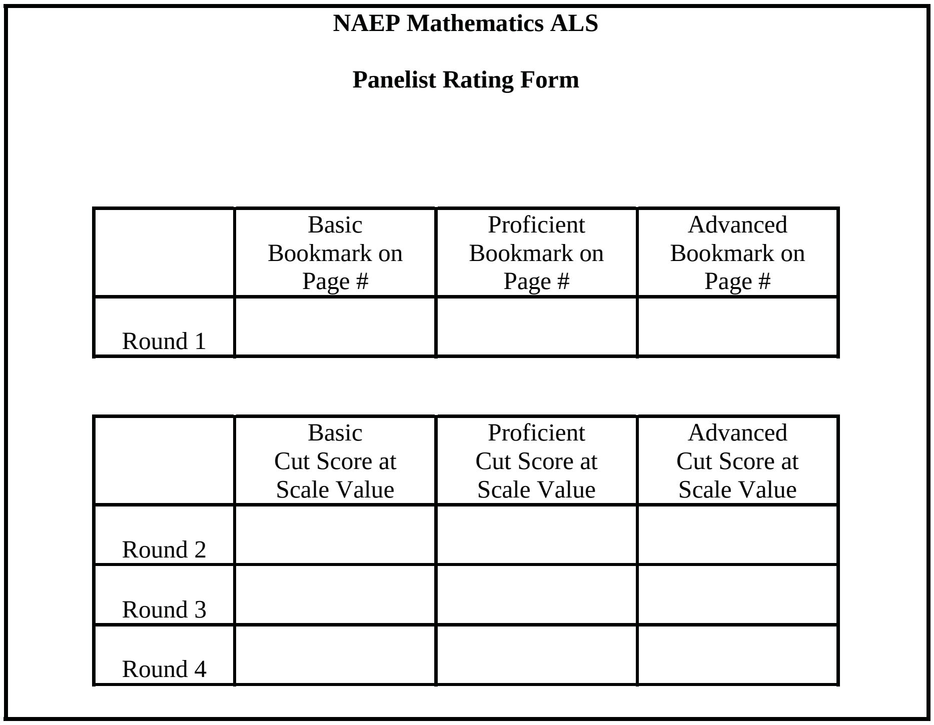
Figure 2. Page 1 of Group A KSA Note Template.

Panelist Rating Form and Data Processing

Figure 3 shows the form that was used by panelists to record their bookmark. 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 Panelist Rating Form for each achievement level were converted to scale values using the Lookup Table shown in Appendix B. The scale values corresponding to the bookmarked page numbers were hand-written on the panelist's Panelist Rating 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. Panelists' names and ID numbers were printed on their Panelist Rating Form. Once all the data were entered, Excel macros were used to compute the median cut score across all panelists, which was reported as the cut score for that round.

In Round 2 and subsequent rounds, panelists entered scale values for their cut score recommendations on their Panelist Rating 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.



NAEP Mathematics ALS

Panelist Rating Form

	Basic Bookmark on Page #	Proficient Bookmark on Page #	Advanced Bookmark on Page #
Round 1			

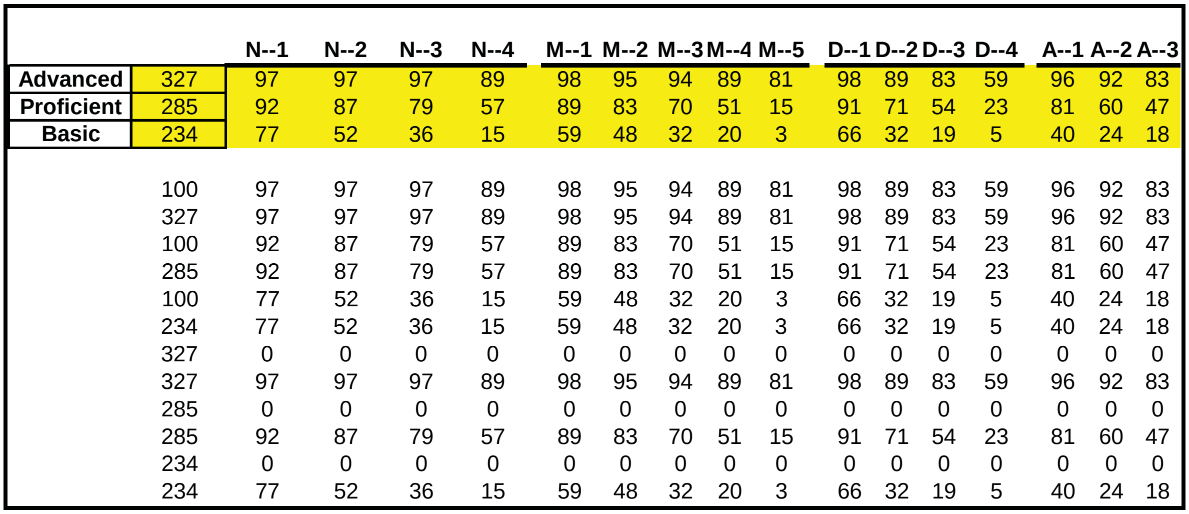
	Basic Cut Score at Scale Value	Proficient Cut Score at Scale Value	Advanced Cut Score at Scale Value
Round 2			
Round 3			
Round 4			

Figure 3. Panelist Rating Form.

Source Spreadsheet

Domain score feedback was shown on the Domain Item Maps, in the Percent Correct Table, in domain score plots, and in the Domain Task 2 Form. These materials existed as spreadsheets in a single Excel book for each round. The spreadsheets contained references to the cells of a “source spreadsheet” within the book. Rows corresponding to the median cut scores from the previous round were copied from a Domain Score Table (see Appendix C) within the book and pasted into the source spreadsheet. Thus, three copy-and-paste steps (for Basic, Proficient, and Advanced separately), produced the domain score feedback information for all materials in a given round.

Figure 4 shows a section of the source spreadsheet. The cells highlighted in yellow were copied in three steps (one step per row) from the Domain Score Table. The remaining cells are partial coordinates for vertical lines illustrating the achievement level boundaries on domain score plots. Many of the coordinates were linked to cells in the yellow-highlighted area.



		N--1	N--2	N--3	N--4	M--1	M--2	M--3	M--4	M--5	D--1	D--2	D--3	D--4	A--1	A--2	A--3
Advanced	327	97	97	97	89	98	95	94	89	81	98	89	83	59	96	92	83
Proficient	285	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
Basic	234	77	52	36	15	59	48	32	20	3	66	32	19	5	40	24	18
	100	97	97	97	89	98	95	94	89	81	98	89	83	59	96	92	83
	327	97	97	97	89	98	95	94	89	81	98	89	83	59	96	92	83
	100	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	285	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	100	77	52	36	15	59	48	32	20	3	66	32	19	5	40	24	18
	234	77	52	36	15	59	48	32	20	3	66	32	19	5	40	24	18
	327	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	327	97	97	97	89	98	95	94	89	81	98	89	83	59	96	92	83
	285	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	285	92	87	79	57	89	83	70	51	15	91	71	54	23	81	60	47
	234	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	234	77	52	36	15	59	48	32	20	3	66	32	19	5	40	24	18

Figure 4. Section of Source Spreadsheet for domain score feedback.

Links to PowerPoint Presentations

There was extensive linking between feedback and data collection forms that existed within Word and Excel files, and PowerPoint presentations. In many cases, such as with the Percent Correct Table (see below), panelists were given only one, generic version of the form, but multiple versions of the form were used in the PowerPoint presentation to focus the discussion on different aspects of the form or task. Each version existed as a separate Excel spreadsheet in specific sections and data were highlighted in advance or by staff in response to the results of the previous round.

Item Maps

There were two kinds of item maps—the Primary Item Map and Domain Item Maps. In the Primary Item Map, items were organized into columns corresponding to subscales of the assessment. There was one Domain Item Map for each subscale. In the Domain Item Maps, items were organized into columns corresponding to teacher domains. Solid vertical lines in Domain Item Maps separated items into Score Domains. When a Score Domain included more than one Teacher Domain, the Teacher Domains were separated by a dashed vertical line. Both types of maps are shown in Appendix D. In the ALS meeting, the maps were printed on 11 by 17 inch 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).

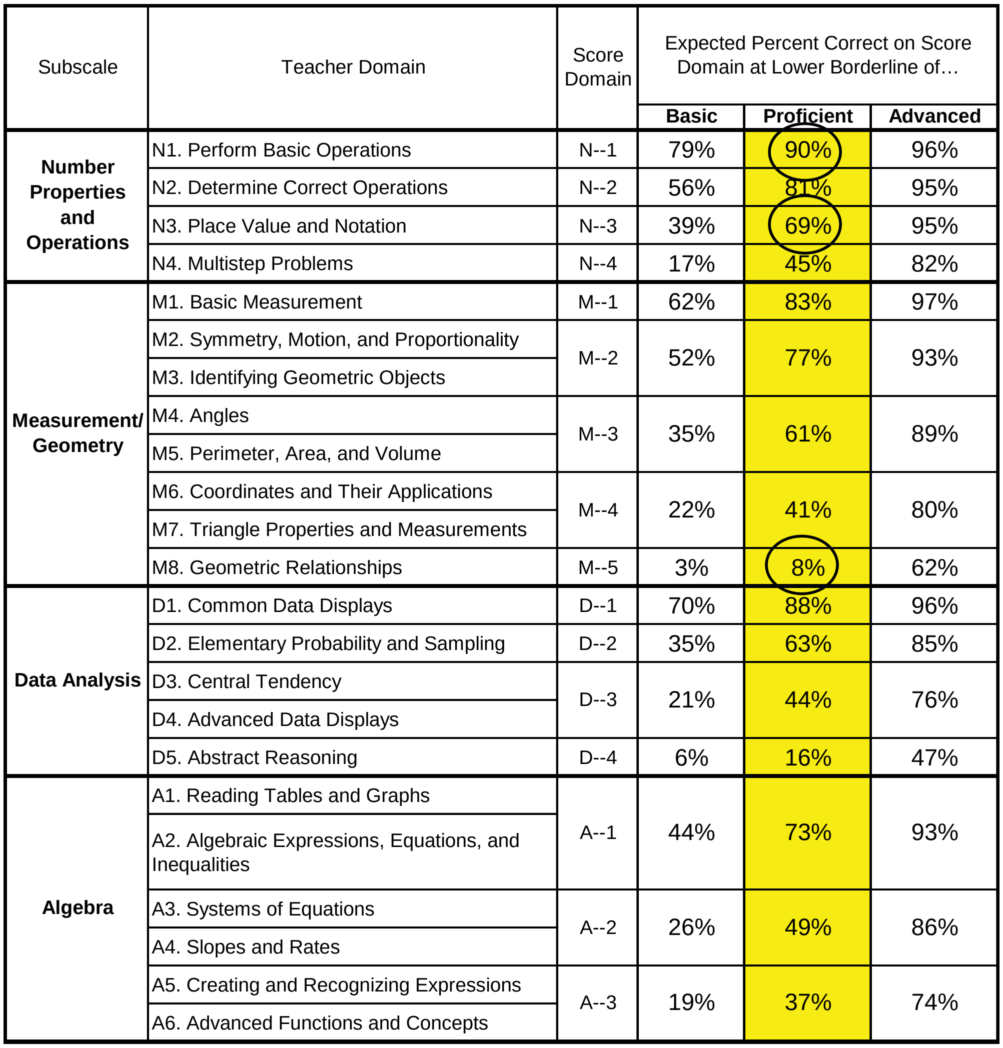
The item handles, color code-characters, and position information for the item handles in the item maps were created by a SAS program. In the process of importing the output of the SAS program into an Excel spreadsheet, the item handles were put into the correct cells in the map. Cells with a given color-code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

For feedback to panelists, horizontal lines were manually added to the Domain Item Maps to indicate the position of the cut scores from the previous round. The lines bordered the bottom of the row/interval containing the cut score. Expected domain scores conditional on the cut scores were incorporated into the maps through links between the cells on the map containing the percent correct scores and the Source Spreadsheet.

Percent Correct Table (PCT)

Figure 5 shows the Percent Correct Table. The cells containing percents in the PCT were linked to cells in the Source Spreadsheet. The PCT was updated after each round of Mapmark.

When the PCT was first introduced to panelists in Round 2, three different versions—one for each achievement level—were presented in PowerPoint displays. These versions existed as separate spreadsheets in the Excel book for Round 1 feedback, differing primarily in the column that was highlighted in yellow. Figure 5 shows the version that was presented for the Proficient achievement level. The circles around the highest and lowest domain scores were present in the PCT for Proficient before the ALS meeting since the scores for the easiest and hardest overall domains would be the highest and lowest regardless of the cut score. A third circle was manually added (only after Round 1) before the table was printed to indicate a domain score that was close to 0.67.



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

Figure 5. Percent Correct Table with Borderline Proficient scores highlighted.

Domain Task 1 Form

The Domain Task 1 Form, a section of which is shown in Figure 6, did not contain any feedback and was, therefore, prepared before the ALS meeting. This form existed as five pages, one page per subscale, in a Word file. Items were listed in this form by their handles in order of increasing scale value within teacher domain within subscale. Teacher domains were arranged on the form in the order they appeared from left to right in the Domain Item Maps.

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

Figure 6. Section of Domain Task 1 Form for Group A.

Domain Ordered Item Book (DOIB)

Panelists used a Domain Ordered Item Book to respond to the statement on the Domain Task 1 Form, "I see how this item is like other items in its domain" (Yes, No, Not Sure). In the DOIB,

- Items were presented in the same order as their handles appeared in the Domain Task 1 Form (by increasing scale value within Teacher Domain within subscale).
- Subscales were identified and separated by tabs.
- A Teacher Domain Definition was placed at the beginning of each Teacher Domain. Teacher Domain Definitions for all Teacher Domains within the Number Properties and Operations subscale are shown in Appendix E.
- Scoring rubrics and examples of student responses to constructed response items were not included.
- Score points of polytomously-scored items were not represented separately in different locations. Rather, polytomously-scored items were located in the DOIB by the scale value of their highest point. They were located in the same way in the Domain Task 1 Form, where they were identified only by the handle for their highest score point (e.g., see Figure 6).

Domain Task 2 Form

The Domain Task 2 Form, which was used in Round 2 only, consisted of one page per achievement level, where each page existed as a separate spreadsheet in the aforementioned Excel book. Cells containing domain scores in this form were linked to the Source Spreadsheet. Figure 7 shows the Domain Task 2 Form for the Proficient achievement level.

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

Figure 7. Domain Task 2 Form for the Proficient achievement level.

Domain Score Chart

The Domain Score Chart was a three-page form—one page each for Basic, Proficient, and Advanced. Figure 8 shows the page for Proficient. This form was created by copying the necessary rows from a spreadsheet that was essentially a Domain Score Table (see Appendix C) with borders between columns into another spreadsheet that already had a title and column headings. The necessary rows contained scale values ranging from 10 minus the lowest recommended cut score to 10 plus the highest recommended cut score from the previous round. The additional markings in Figure 8 (highest/lowest lines, highlighted median, and circles around 67% correct scores) were added by staff during the meeting. The circle indicating a panelists' location (Panelist "X") was added by the panelist.

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

Figure 8. ***in Score Chart showing Round 1 results and location of Panelist "X" for the Proficient achievement level.***

Domain Score Plots

Domain score plots were used extensively in Round 2, and to a lesser extent in Round 3. They were used only in PowerPoint presentations, where they existed as links to charts in the Excel book containing the Source Spreadsheet. Various forms of the plots existed to call attention to different topics of discussion. Figure 9 shows the basic plot for score domains representing items calibrated to the Data Analysis and Probability subscale. There was one such plot for each subscale. The coordinates of the vertical bars in the plots, which indicated the location of the cut scores, were linked to the Source Spreadsheet and were updated accordingly. The dashed horizontal line at 67% correct was present on some versions of the plot, but not on others, depending on the topic of discussion.

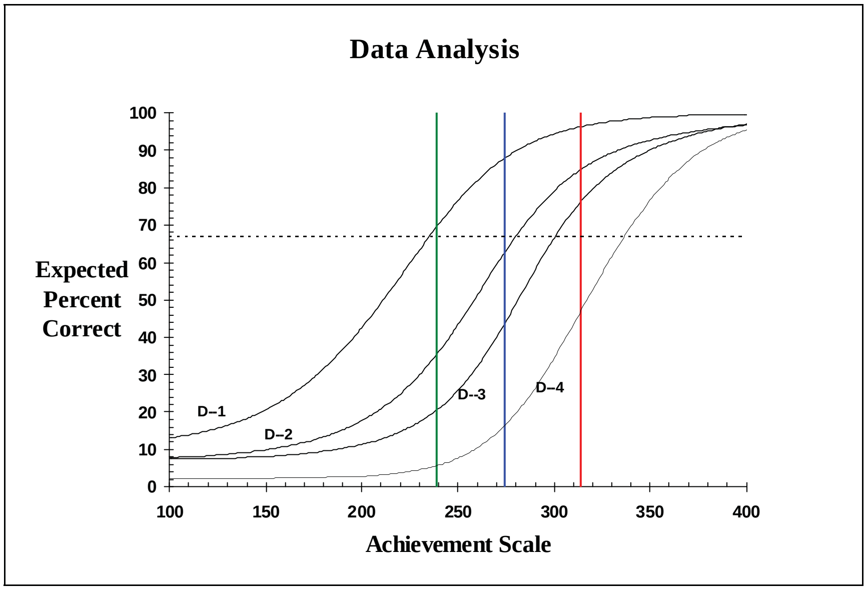


Figure 9. Percent correct score plot for Data Analysis and Probability subscale.

Scale Value to OIB Page Lookup Table

In Round 3 of Mapmark, panelists referred to both the Domain Score Chart (DSC) and their OIB to select a scale value for their cut score recommendation. To facilitate their task, they were given a Scale Value to OIB Page Lookup Table shown in Appendix C. This table was new to the ALS meeting, but panelists' remarked so favorably on it that it should be regarded as important material for Round 3 of the Mapmark process.

Consequences Feedback and Questionnaire

Consequences feedback was presented to panelists in the form of Figure 10. 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 Achievement table in Appendix C. The questionnaire is shown in the Process Report. It did not present any consequences data itself, so it was printed prior to the ALS meeting.

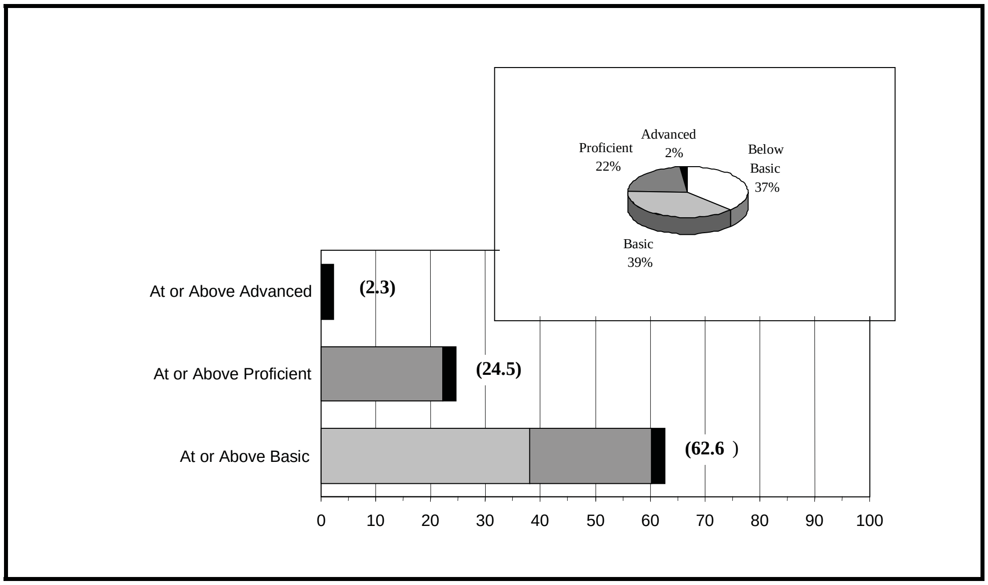


Figure 10. Consequences data presented to panelists in Round 4.

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

Figure 11. Exemplar Rating Form for the Basic achievement level.

Exemplar Item Rating Form

An Exemplar Item Rating Form for each achievement level was produced by running a computer program and copying the output of the program into an Excel spreadsheet that contained formatting like that shown in Figure 11 for the Basic achievement level. Information in the first four columns of the form, as shown in Figure 11, were output by the computer program. The computer program used the Round 4 median cut scores as input. Items were associated with achievement levels as explained in the Psychometric Procedures section of this report. The probabilities shown in the fourth column were conditional on the midpoint (Basic and Proficient) or median (Advanced) of the level, as defined in the Psychometric Procedures section.

RELIABILITY ESTIMATES

The term "reliability" is used here to represent the notion that cut scores from two different achievement level setting meetings should not differ if the same method is used and the meetings are the same in all key respects, such as using the same assessment and achievement level descriptions. Cut score reliability was evaluated by examining the standard error of the cut score. More reliable cut scores have smaller standard errors.

Single Meeting Estimate

ACT has traditionally used an estimate based on the division of the panelists in the ALS meeting into two groups, A and B, where each group works with equivalent but overlapping item pools. The formula for this estimate is:

$$SE_1 \quad \frac{|\,Cut_A - Cut_B\,|}{2}, \qquad \textit{(17)}$$

where Cut_A and $Cut_{B.}$ are final cut scores from Groups A and B, respectively, within the same meeting or study.

Two drawbacks of this estimate are that: 1) it is unstable since it is based on only two observations and can, therefore, be unaccountably very large or very small, and 2) the observations (Cut_A and Cut_B) are not really independent, and so they do not truly represent the hypothetical results of two separate ALS meetings.

Table 9 provides evidence of the instability of SE_1 estimates. SE_1 estimates in this table are based on the studies performed in the current mathematics achievement level setting project. SE_1 estimates from the 1992 ALS meeting are also shown. There are large differences across achievement levels within studies, and across studies for the same achievement level. Mapmark estimates range from 0.0 to 9.5. Item Rating estimates range from 2.40 to 5.72.

As much as the SE_1 estimates in Table 9 vary over levels and methods, it difficult to say whether the reliability of cut scores tends to vary across levels or methods. There is no consistent evidence that Basic cut scores tend to be more reliable than Advanced cut scores in general, for example.

Without convincing evidence that SE_1 is larger for one level than another, a reasonable strategy is to average estimates over levels. This is done in the last row of Table 9. Even these averages should not be compared across methods, however, if the number of panelists per group is not the same. For example, the average SE_1 estimate for Field Trial 1 (7.5) may be large because there were only five panelists per group.

Table 9. Within-Occasion Estimates of Standard Error of Cut Scores (SE_1)

Level	Mapmark					Item Rating	
	Field Trial 1	Field Trial 2	Grade 8 Study	Pilot Study	ALS Meeting	1992 ALS Meeting*	Pilot Study
Basic	7.0	1.5	5.25	3.00	5.50	2.40	4.75
Proficient	9.5	2.5	4.75	0.75	1.00	5.72	4.00
Advanced	6.0	2.0	0.75	0.00	0.25	4.84	2.81
Average	7.5	2.0	3.58	1.25	2.25	4.32	3.85

*These values were reported for Grade 12 in the final report for the 1992 ALS Meeting (ACT, 1993).

Estimates Based on Two Meetings

An ideal estimate of the standard error of a cut score resulting from a given method would involve actual replications of the method. A formula for estimating the standard error of a method, when the method is performed on two separate occasions is:

$$SE_2 \quad \frac{|Cut_1 - Cut_2|}{2}, \qquad (18)$$

where *Cut₁* and *Cut₂* are final cut scores from the two meetings in which the same method was used. Unfortunately, this particular estimate—being based on just two observations—is also unstable. However, since each observation for SE_2 (Cut_1 and Cut_2) is based on all the panelists within a given replication, while those for SE_1 are each based on only half of the panelists within a replication, SE_2 may be more stable than SE_1.

Another problem with SE_2 is that true replications of a method are almost impossible to come by. In the present project, one could treat Field Trial 1 and the Grade 8 Study as replications because they involved the same RP criterion, achievement level descriptions, and used items from a common source (Grade 8). But Field Trial 1 was a two-day procedure involving only a two-round Mapmark method, while the Grade 8 Study was a five-day, four-round procedure. The Pilot Study and ALS meeting in this project are more nearly exact replications of Mapmark, but even these methods differed in potentially important ways.

Table 10 shows SE_2 estimates for the Mapmark method based on the replications described in the previous paragraph. Despite the potentially significant differences between replications, the average of all the SE_2 estimates in Table 10 is small in comparison to most of the Mapmark SE_1 estimates in Table 9. This may be due in part to the large sample sizes per observation, as noted above.

Table 10. Across-Occasion Estimates of the Standard Error of Mapmark Cut Scores (SE_2)

	Grade 8 Data (R2 Field Trial vs. R3 Grade 8 Study)	Grade 12 Data (Pilot Study vs. ALS Meeting)
Basic	0.50	1.00
Proficient	0.00	2.50
Advanced	0.50	1.00
Average	0.33	1.50

To provide another estimate of SE_2, differences between the Pilot Study and ALS meeting cut scores were computed from ACT's previous standard setting work for NAGB. The average of 42 estimates pooled across three grades (4, 8, and 12), 3 achievement levels (Basic, Proficient, and Advanced), and seven subjects (mathematics, reading, civics, science, writing, geography and U.S. history) was 5.4. There was no clear evidence in this historical data that the reliability of cut scores differed across the grades, levels, or subjects involved. Angoff-based methodologies were used in all cases.

SPECIAL STUDIES

Table 11 shows special studies that were conducted in this project. All studies involved a Mapmark method similar to that used in the ALS meeting. The technical procedures and materials for performing the Mapmark method in the special studies were substantially the same as those described in this Technical Report with the exception that items and data from the 2003 NAEP in Grade 8 mathematics were used for Field Trials 1 and 2 and for the Grade 8 study.

Table 11. Special Studies

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

The Special Studies report provides more specific information about the methods and materials used in the special studies. Because the studies using Grade 8 data were not essential to the ALS outcomes for Grade 12, detailed technical information concerning such matters as the item statistics, item handles, and item maps produced with the Grade 8 data, similar to what is being provided in this Technical Report, are not presented in the Special Studies Report for the project.

The item statistics, materials, transformation constants, and all technical procedures used for the Mapmark method in the Pilot Study are exactly as described here in this technical report. A general description of the materials and procedures used in the Item Rating method in the Pilot Study are contained in the Special Studies Report. Psychometric and technical procedures for this method are detailed in the reports produced for the 1998 NAEP Civics ALS project (Loomis & Hanick, 2000). More detail concerning the application of the technical and psychometric procedures to the Grade 12 data used in the Pilot Study are not presented in the reports for this project because these are not directly related to the ALS outcomes of this project.

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), 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 C.)

Four output files are created:

- *naepg12out1.out* contains, for each score point in the assessment (N=237), the cumulative probability given in Equation 6 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.
- *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=180), the expected score (item true score), as defined in Equation 13 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.
- *naepg12out4.out* contains, for each item in the assessment, the expected item true score as defined in Equation 12 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:

- *final4.prn* is a file that contains assignments of items to Teacher Domains.
- *code_seq* contains item codes created in the domain development component of the project.
- *g12_math_irt* is a file of item statistics received from the test development contractor, essentially like the Preliminary Item Statistics table in Appendix C.
- *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, CROIB, and DOIB. 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, Constructed Response Ordered Item Book, and Domain 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_dm_txt* is a list for assembling the Group A DOIB.
- *groupa_dm.task1* is a list for assembling the Domain Task 1 Form for Group A
- *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 programs domain_map.sas and primary_map.sas also use Set9 from Mapmark1.sas. They produce output files *domain.map* and *primary.map*, respectively. These files are incorporated into excel spreadsheets to create the item maps. These programs include the code that reassigns some items to item pools A and B in Step 2 of the item pool division process described earlier.

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 14, 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.

mapmark-examplar.out is the output file. The contents of this file are copied into an Excel spreadsheet containing the formatting for the Exemplar Item Rating Form.

Programs for Item Rating

The program theta12.for is a FORTRAN program used to compute cut scores for each panelist. Six input files are needed:

- *naep12.cc* contains mean and standard deviation of student achievement on the composite score scale, transformation values (Table 2), and subscale correlations (Table 3).
- *g12_irt_info* contains each item's NAEPID, block, sequence, subscale classification, item type, and IRT item parameter estimates (see Item Statistics table in Appendix C).
- *FDIMPG12_2.prn* is the frequency distribution of student achievement received from the test development contractor.
- *judgeid.txt* contains panelist ID, secret ID, and group assignment (A or B).
- *R*GroupARatings.txt* is a multiple file designation where "*" is a place holder for an integer representing the Round number (1, 2, 3). Another two input files contain Group B ratings by round.

Four output files are produced:

- *estTheta-R*.out* contains cut scores for each panelist for Round *.
- *estTheta-R*a.out* contains panelist ID, cut scores for basic, proficient, and advanced for each panelist.
- *estTheta-R*b.out* contains secret ID, cut scores for basic, proficient, and advanced for each panelist.
- *raterloc-R*.txt* for creating rater location plot.

The program reck2a12-fornape.for is a FORTRAN program used to obtain the Reckase charts. Four input files are needed:

- *naep12.cc* contains mean and standard deviation of student achievement on the composite score scale, transformation values (Table 2), and subscale correlations (Table 3).
- *g12_irt_info* contains NAEP ID, block, sequence, scale, type, and item parameter estimates.

- *raterid!.txt* contains raters' ID for group !.
- *R*Group!Ratings.txt* is a multiple file designation where * represents round, and ! represents group. The file name needs to be changed for each group and each round.

The output file is *reckase-R*!.out,* where * and ! are round and group. This text file can be formatted to produce the Reckase chart for each panelist.

The program plot12.for is a FORTRAN program for creating rater location plots. Three input files are needed:

- *raterloc-R*.txt* is one of the output files for theta12.for. It contains cut scores for each panelist for each achievement level for round *.
- *FDIMPG12_2.prn* contains the student distribution.
- *raterid12.txt* contains the raters' secret ID for both groups.

The output file is *locPlot-R*.out,* where * stands for round.

The FORTRAN program exmp-ir.for is the program used to generate exemplar items for the Item Rating (IR) procedure in the Pilot Study using the method described with reference to Equation 14, using an RP of 0.5. Four input files are needed:

- *angoff.out* contains an output file generated using a SAS program (item, level, block, sequence, and sorted probabilities conditional on midpoints or median).
- *label.txt* contains a brief descriptor for each item.
- *pctatabove.txt* contains the percent of students at or above each scale score.
- *cutpts.txt* contains final cut scores using the IR procedure.

exemplar-ir.out is the output file.

TECHNICAL ADVICE

ACT relied extensively on the advice of its Technical Advisory Committee on Standard Setting (TACSS) in all phases of the project. The TACSS is a five-member group that collectively represents expertise in psychometrics, standard setting, and mathematics education. ACT met six times with its TACSS over the course of the project and held one TACSS conference call. Minutes of the TACSS meetings are presented in Appendix A. A member of the TACSS also served as a member of ACT's internal Technical Advisory Team (TAT). Key technical advice from persons external to ACT, principally TACSS, but also through the participation of a TACSS member on ACT's TAT, is summarized below.

RP Criterion

TACSS initially agreed to the use of an RP of 0.65 for Field Trial 1 and 0.50 for Field Trial 2, but through email correspondence agreed to the use of a 0.67 RP in Field Trial 1. After a review of results from the field trials, TACSS recommended the use of a 0.67 RP for the Grade 8 Study. This recommendation carried over to the Pilot Study. With TACSS input, ACT recommended a 0.67 RP to NAGB for use with Mapmark in the ALS meeting.

Mapmark Procedures

In the first meeting, TACSS approved the basic plan to use a bookmark-based procedure supplemented with the Primary Item Map in Round 1 of Mapmark, and to introduce domains and domain score feedback in Round 2. TACSS advice also encouraged the exclusive use of domain score feedback by panelists to select scale values for cut score recommendations in Round 2, rather than allowing panelists to continue using their Ordered Item Books in this round.

Process Evaluation Questionnaires and Data

TACSS provided extensive review and input concerning the process evaluation questionnaires used in the Pilot Study and ALS meeting and on the analyses performed on the process evaluation data. The input led to greater comparability of data across methods (Item Rating and Mapmark) and with previous ALS processes ACT had conducted for NAGB.

Design and Panelist Effects

TACSS advised against relying on parametric statistical analyses to identify effects of design (panelists and tables) and panelists (type, ethnicity, gender, and region). One reason was that these procedures do not support inferences about the median, which is used in Mapmark. TACSS advised that the presentation of effects be primarily descriptive. From their inspection of the data, they did not feel that Mapmark was susceptible to design and panelist effects, except that table effects might be greater in Mapmark than in Angoff-based procedures.

Methodology for ALS Meeting

In a report presented to the Committee on Standards, Design, and Methodology (COSDAM) in a special meeting to select Item Rating or Mapmark for the ALS meeting, TACSS and ACT jointly concluded that both procedures had procedural validity, were reliable, and were likely to produce results that would be considered reasonable. However, TACSS indicated a slight preference for the Mapmark procedure. This preference was based on a number of considerations including: 1) ACT's and TACSS's assertion that the Mapmark procedure could be conducted in four days without compromising the integrity of the process, 2) the perceived value of the item KSA (knowledge, skills, and abilities) review to the standard setting process, and 3) the potential that domains have for use in describing achievement levels in NAEP reports.

Recommendations for Future Studies

In the last meeting, TACSS recommended the following studies as follow-up activities to the project.

- Conduct a small study in which teachers independently classify items into the domains that were used in the Pilot Study and ALS meeting.
- Develop a prototype report that uses domains to describe the achievement levels.
- Explore the feasibility of using domains in other content areas, such as Reading.
- Explore the feasibility of incorporating domains into the assessment framework and test development process.

REFERENCES

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

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.

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