Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science

Technical Report

Presented by ACT, Inc.
August 2010

Redacted by:
National Assessment Governing Board

National Assessment Governing Board
2009–2010

Honorable David P. Driscoll, Chair
Former Commissioner of Education
Melrose, Massachusetts

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

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

Carol A. D'Amico
President and Chief Executive Officer
Conexus Indiana
Indianapolis, Indiana

Louis M. Fabrizio
Director, Accountability Policy and Communications
North Carolina Department of Public Instruction
Raleigh, North Carolina

Honorable Anitere Flores
Member
Florida House of Representatives
Miami, Florida

Alan J. Friedman
Consultant
Museum Development and Science Communication
New York, New York

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

Doris R. Hicks
Principal and Chief Executive Officer
Dr. Martin Luther King, Jr. Charter School for Science and Technology
New Orleans, Louisiana

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

Kim Kozbial-Hess
Fourth-Grade Teacher and Educational Technology Trainer
Toledo, Ohio

Henry Kranendonk
Mathematics Consultant
Milwaukee Public Schools
Milwaukee, Wisconsin

Tonya Miles
General Public Representative
Mitchellville, Maryland

Honorable Steven L. Paine
State Superintendent of Schools
West Virginia Department of Education
Charleston, West Virginia

Honorable Sonny Perdue
Governor of Georgia
Atlanta, Georgia

Susan Pimentel
Educational Consultant
Hanover, New Hampshire

W. James Popham
Professor Emeritus
Graduate School of Education and Information Studies
University of California, Los Angeles
Wilsonville, Oregon

Andrew C. Porter
Dean
Graduate School of Education
University of Pennsylvania
Philadelphia, Pennsylvania

Warren T. Smith
Vice President
Washington State Board of Education
Olympia, Washington

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

Oscar A. Troncoso
Principal
Anthony High School
Anthony Independent School District
Anthony, Texas

Honorable Leticia Van de Putte
Senator
Texas State Senate
San Antonio, Texas

Eileen L. Weiser
General Public Representative
Ann Arbor, Michigan

Darvin M. Winick
President
Winick & Associates
Austin, Texas

Ex-officio Member
John Q. Easton
Director
Institute of Education Sciences
U.S. Department of Education
Washington, D.C.

Staff
Cornelia Orr
Executive Director

Susan Loomis
Assistant Director, Psychometrics

Committee on Standards, Design, and Methodology

Louis Fabrizio, Chair
Steven Paine, Vice Chair
Carol D'Amico
Tonya Miles
James Popham
Andrew Porter
Darvin Winick
Susan Loomis (staff)

Developing Achievement Levels on the 2009 National Assessment of Educational Progress in Science

Technical Report

The work for this report was conducted by ACT, Inc., under contract ED-08-CO-0143 with the National Assessment Governing Board.

Technical Report
Table of Contents

INTRODUCTION **1**

PSYCHOMETRIC PROCEDURES **2**

Description of Item Pool 2
Computation of Item Scale Values for a Response Probability of 0.67 7
Item Handles 8
Item Map Values 10
Whole Booklet Feedback 10
Consequences Feedback 11
Mapping Potential Exemplar Items to Achievement Levels 11
Reliability Estimates 12
Process Evaluations 14
Scale Transformation Error 14

MATERIALS **16**

Division of Panelists and Item Pools into Rater-Groups/Pools A and B 16
Ordered Item Book 20
Constructed Response Ordered Item Book 21
Cut Score Recommendation Form and Computation of Cut Scores 22
Item Map 23
Cut Score Distribution Chart 23
Scale Value to OIB Page Lookup Table 24
Item Score Table, Booklet Score Chart, and Booklet Score Plot 24
Consequences Feedback and Questionnaire 29
Exemplar Item Rating Form 30

PILOT STUDY **31**

TACSS INPUT **31**

COMPUTER PROGRAMS **33**

OIB and CROIB 34
Item Maps 34
Whole Booklet Feedback 34
Exemplar Item Charts 35

REFERENCES **36**

APPENDICES

A. Technical Advisory Committee on Standard Setting
B. Scale Value to OIB Page
C. Item Maps Used in ALS
D. Item Maps Using Correct Scale Transformations
E. Item Order for OIB by Group
F. Constructed-Response Ordered Item Book Contents by Group
G. Range of Uncertainty

List of Tables

Page

Table 1: Collapsed constructed-response items for grade 4 2
Table 2: Summary of grade 4 item pool by block 3
Table 3: Collapsed constructed-response items for grade 8 4
Table 4: Summary of grade 8 item pool by block 5
Table 5: Collapsed constructed-response items for grade 12 6
Table 6: Summary of grade 12 item pool by block 7
Table 7: Example of item handles, scale values, and map values for hardest and easiest items within item type (grade 12) 9
Table 8: Estimates of standard error of cut scores 13
Table 9: ANOVA results for rater group (item pool) and table comparisons of cut scores 14
Table 10: Differences in scale values for items due to incorrect transformation 15
Table 11: Consequences data presented in ALS versus that based on the correct scale transformations 16
Table 12: Summary of item pools A and B (grade 4) 18
Table 13: Summary of item pools A and B (grade 8) 19
Table 14: Summary of item pools A and B (grade 12) 20

List of Figures

Page

Figure 1: Numbers of items reviewed by group A, group B, and both groups A and B at each grade 17
Figure 2: Illustration of the information on an OIB page 21
Figure 3: Panelist Cut Score Recommendation Form 22
Figure 4: Cut Score Distribution Chart showing the distribution of cut scores by achievement level after round 1 for grade 4 24
Figure 5: Item Score Table for grade 8, form C 25
Figure 6: Proficient Booklet Score Chart for grade 4, group A 28
Figure 7: Booklet Score Plot for grade 4, form C 29
Figure 8: Sample consequences data 30
Figure 9: Exemplar Item Rating Form for the Basic achievement level for grade 4 31

DEVELOPING ACHIEVEMENT LEVELS FOR THE 2009 NAEP IN SCIENCE FOR GRADES 4, 8, AND 12: TECHNICAL REPORT

INTRODUCTION

In September 2008, the National Assessment Governing Board contracted with ACT to conduct activities for setting achievement levels on the 2009 National Assessment of Educational Progress (NAEP) in science for grades 4, 8, and 12. The contract called for two reports, including a Technical Report documenting the technical aspects of ACT's contract activities.

This Technical Report provides technical information related to the materials and process used for the achievement level setting (ALS) meeting that was held in January 2010 to set achievement levels for the 2009 NAEP in science for grades 4, 8, and 12. The data used in the meeting consisted of items, item statistics, and estimates of student achievement from the 2009 NAEP administration of science. The methodology used to set the achievement levels was Mapmark with Whole Booklet Feedback, a bookmark-based procedure that includes item maps and student test booklets.

This report also provides information for the technical aspects of the Pilot Study held in October 2009. In addition to this Technical Report, the Process Report (ACT, 2010) provides an overview of the Pilot Study and a detailed description of the ALS meeting process and results.

The Technical Report also accounts for technical advice ACT received throughout this project. ACT relied on the advice of its Technical Advisory Committee on Standard Setting (TACSS), the Contract Officer's Representative (COR), and the Committee on Standards, Design, and Methodology (COSDAM). The TACSS was a six-member group that collectively represents expertise in standard setting, science education, and experience with the NAEP. The Governing Board's COR was Dr. Susan Loomis, Assistant Director of Psychometrics. COSDAM is a committee of the Governing Board. The Project Director gave a progress report to COSDAM at each Board meeting during the time frame of the contract. One meeting was also held with COSDAM on May 7, 2010, to discuss the results of the ALS meeting and possible options.

This document is divided into five primary sections: *Psychometric Procedures*, *Materials*, the *Pilot Study*, *TACSS Input*, and *Computer Programs*.

The *Psychometric Procedures* section deals with the statistical characteristics and calculations used during the ALS process. This includes descriptive information for the items used, the description of the statistics used in the Pilot Study and ALS meetings, and the subsequent analysis of the results. The method for calculating the statistic is also given in cases where it is not straightforward.

The *Materials* section provides technical information needed to prepare some of the meeting materials that were given to the panelists during the ALS meeting. A description is given, along with an example of the material.

The *Pilot Study* section describes the technical information and data needed for a preliminary study conducted in October 2009 as part of the contract. The purpose of the Pilot Study was to tryout the Mapmark with Whole Booklet Feedback method as planned for the ALS meeting. Results from the Pilot Study were used to provide recommendations for enhancements to the process that were used for the ALS meeting.

Technical input from TACSS is summarized in the *TACSS Input* section. The TACSS met six times over the course of the project and provided technical advice concerning all aspects of the project. Input from these meetings was used to guide implementation of the Mapmark process used in the ALS meeting. Members of the committee and the minutes for each meeting are presented in Appendix A.

The final section, *Computer Programs*, lists the computer programs used for the Pilot Study and ALS meeting. The name of the program, along with a brief description of what the program does and the inputs and outputs are given.

PSYCHOMETRIC PROCEDURES

Description of Item Pool

The ALS meeting used items, item statistics, and student performance data from the 2009 NAEP in science.

For grade 4, the assessment originally had 144 items in nine blocks. Only 141 items were actually used for the ALS. The following three items on the assessment were not used in the scaling of the items, or in the estimation of student ability parameters:

Block SD, Item 13
Block SH, Item 4
Block SI, Item 1

Table 1 lists the constructed-response items that had score points that were "collapsed." The score levels for an item were collapsed if, in the item scaling process, the results did not support the original number of scoring categories.

Table 1: Collapsed constructed-response items for grade 4

Block	Item Number	Score levels prior to collapse	Score levels after collapse
SC	13	4	2
SD	10	4	3
SE	10	5	4
SG	5	3	2
SJ	3	5	4
SK	14	3	2

After these adjustments, there were 14 to 17 items scored in each block for grade 4. Of the 141 scored items, 95 were multiple choice (MC) and 46 were constructed response (CR). The CR items represented a total of 101 score points, or 51% of the points in the

item pool, and MC items represented 49% of the points. The total number of points was 196. Table 2 shows how the items were distributed by block, content area, and item type.

Table 2: Summary of grade 4 item pool by block

Block	All Items	Number of Items with Item-Statistics					CR Points[c]	Total Points	Ave Scale Score[d]
		Content Area[a]			Item Type[b]				
		E & S	Life	Phys	MC	CR			
SC	17	5	6	6	12	5	10	22	405.6
SD	14	4	7	3	9	5	8	17	384.4
SE	17	4	6	7	11	6	13	24	412.3
SF	16	7	4	5	11	5	12	23	391.1
SG[e]	16	6	5	5	10	6	12	22	395.0
SH	15	3	6	6	10	5	13	23	408.2
SI	14	6	5	3	10	4	11	21	390.6
SJ	16	5	5	6	11	5	12	23	405.0
SK[e]	16	6	3	7	11	5	10	21	401.0
Total	**141**	**46**	**47**	**48**	**95**	**46**	**101**	**196**	**400.9**
		33%	33%	34%	67%	33%	51%		

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] MC = Multiple Choice; CR = Constructed Response
[c] CR Points = the number of score points represented by constructed response items
[d] Score scale for grade 4 had a mean of 364 and standard deviation of 33
[e] Common blocks

For grade 8, the assessment originally had 166 items organized into ten blocks. Only 162 items were used in the ALS. The following four items on the assessment were not used in the scaling of the items, or in the estimation of student ability parameters:

Block C, Item 9
Block C, Item 10
Block G, Item 11
Block I, Item 11

Again, there were some extended CR items (i.e., items with more than 2 score categories) with scoring levels collapsed. These items are listed in Table 3.

Table 3: Collapsed constructed-response items for grade 8

Block	Item Number	Score levels prior to collapse	Score levels after collapse
SC	14	5	4
SD	11	4	3
SE	15	5	4
SF	7	5	4
SG	4	5	4
SG	13	5	3
SH	13	4	3
SI	5	4	3
SI	12	4	3
SJ	7	4	2
SJ	12	3	2
SK	10	4	3
SK	14	4	3
SL	7	3	2
SL	9	3	2
SL	12	5	3

After these adjustments, there were 10 blocks with 13 to 18 items in each block for grade 8. Of the 162 scored items, 104 were MC and 58 were CR. The CR items represented a total of 145 score points, or 58% of the points in the item pool, and MC items represented 42% of the points. The total number of points was 249. Table 4 summarizes the items by block, item type, and content area.

Table 4: Summary of grade 8 item pool by block

Block	All Items	Number of Items with Item-Statistics					CR Points[c]	Total Points	Ave Scale Score[d]
		Content Area[a]			Item Type[b]				
		E & S	Life	Phys	MC	CR			
SC	15	6	4	5	10	5	15	25	602.3
SD	18	6	7	5	13	5	14	27	613.4
SE	18	7	6	5	13	5	13	26	623.4
SF	17	6	4	7	11	6	14	25	620.3
SG	15	6	5	4	9	6	17	26	619.2
SH	18	5	6	7	13	5	12	25	626.9
SI	13	6	4	3	7	6	15	22	631.1
SJ[e]	17	8	4	5	10	7	17	27	613.0
SK[e]	16	9	5	2	9	7	17	26	622.2
SL	15	9	2	4	9	6	11	20	615.3
Total	**162**	**68**	**47**	**47**	**104**	**58**	**145**	**249**	**619.5**
		42%	29%	29%	64%	36%	58%		

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] MC = Multiple Choice; CR = Constructed Response
[c] CR Points = the number of score points represented by constructed response items
[d] Score scale for grade 8 had a mean of 579 and standard deviation of 33
[e] Common blocks

For grade 12, the assessment had 185 items, of which 179 were used for the ALS. The following six items on the assessment were not used in the scaling of the items, or in the estimation of student ability parameters:

Block D, Item 5
Block E, Item 9
Block G, Item 1
Block L, Item 6
Block L, Item 13
Block M, Item 3

Table 5 shows the items adjusted for collapsing of score levels.

Table 5: Collapsed constructed-response items for grade 12

Block	Item Number	Score levels prior to collapse	Score levels after collapse
SC	4	5	3
SC	5	4	3
SC	11	5	4
SC	12	3	2
SD	3	5	3
SD	12	3	2
SD	15	4	3
SE	7	4	2
SE	10	3	2
SE	15	4	2
SF	2	4	2
SF	6	3	2
SF	15	3	2
SG	5	3	2
SG	6	4	2
SH	11	4	3
SH	12	5	2
SI	4	5	4
SI	12	3	2
SI	14	4	3
SJ	9	5	3
SJ	12	4	2
SK	5	4	3
SK	13	4	2
SL	10	4	3

Table 6 shows how the items for grade 12 were distributed by content area and item type. The items were organized into eleven blocks. There were 14 to 18 items in each block. Of the 179 scored items, 120 were MC and 59 were CR. The CR items represented a total of 125 score points, or 51% of the points in the item pool, and MC items represented 49% of the points. The total number of points was 245.

Table 6: Summary of grade 12 item pool by block

Block	All Items	Number of Items with Item-Statistics					CR Points[c]	Total Points	Ave Scale Score[d]
		Content Area[a]			Item Type[b]				
		E & S	Life	Phys	MC	CR			
SC	17	2	6	9	11	6	12	23	828.7
SD	17	5	5	7	12	5	11	23	831.9
SE	16	4	7	5	11	5	7	18	837.5
SF	17	6	6	5	12	5	9	21	842.3
SG	16	5	8	3	10	6	11	21	829.1
SH	17	5	7	5	11	6	14	25	820.2
SI[e]	18	6	6	6	12	6	13	25	835.8
SJ[e]	16	3	6	7	11	5	12	23	837.9
SK	16	4	5	7	11	5	9	20	828.3
SL	14	3	7	4	9	5	13	22	844.2
SM	15	3	7	5	10	5	14	24	841.1
Total	**179**	**46**	**70**	**63**	**120**	**59**	**125**	**245**	**834.6**
		26%	39%	35%	67%	33%	51%		

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] MC = Multiple Choice; CR = Constructed Response
[c] CR Points = the number of score points represented by constructed response items
[d] Score scale for grade 12 had a mean of 793 and standard deviation of 33
[e] Common blocks

Computation of Item Scale Values for a Response Probability of 0.67

For each grade, all items in the assessment were calibrated together on an overall score scale, even though the items were classified into three content areas and four science practices. As noted earlier, the three content areas are Physical Science, Life Science, and Earth and Space Sciences. The four science practices are Identifying Science Principles, Using Science Principles, Using Scientific Inquiry, and Using Technological Design.

For each grade, the computation of item scale values in the Mapmark method begins with the computation of score probabilities. Let U_i represent the random score on item i and let θ represent student achievement on the overall scale. For MC and short CR (i.e., dichotomously-scored) items, the following item response theory model was used:

$$P(U_i = 1 \mid \theta) = p_i = c_i + \frac{1 - c_i}{1 + \exp\left[-Da_i(\theta - b_i)\right]}, \quad (1)$$

where D is 1.7, a_i is the item discrimination parameter, b_i is the item difficulty parameter, c_i is the pseudo-guessing parameter for multiple choice items or $c_i = 0$ for dichotomously scored constructed response items. For extended CR (i.e., polytomously-scored) items, the following item response theory model was used:

$$P(U_i = h \mid \theta) = p_{ih} = \frac{\exp\left[\sum_{r=0}^{h} Da_i(\theta - b_i + d_{ir})\right]}{\sum_{k=0}^{m_i} \exp\left[\sum_{r=0}^{k} Da_i(\theta - b_i + d_{ir})\right]}, \quad (2)$$

where m_i is the maximum score on the item, and d_{ir} is the threshold parameter on item i for score r, $r=0,1,\ldots,m_i$, and $d_{i0} = 0$.

The values on the theta scale were transformed to a NAEP-like scale by multiplying each theta value by 35, and then adding 150. This gives a mean of student achievement on the scale (μ) for grades 4, 8, and 12 of approximately 150, with a standard deviation of approximately 35.[1]

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

$$P(U_i \geq h \mid \eta) \geq RP, \quad (3)$$

where *RP* stands for the response probability criterion. For the ALS meeting, an RP of 0.67 was used.

In the science ALS process, for grades 4, 8, and 12, respectively, 214, 428, and 642 was added to the item scale value obtained as described with reference to Equation 3. 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 or who may have inappropriately compared the cut scores across grades. This addition produced item scale values starting at 215, 429, and 643 for grades 4, 8, and 12, respectively. Item scale values on the Mapmark scale are shown in the Scale Value to OIB Page Lookup Tables in Appendix B. Panelists used these tables when setting cut scores in round 2 to locate items in their OIBs corresponding to scale values associated with student booklets.

Item Handles

An item handle is a short character string that represents the item on the item map. Multiple choice and short CR items had a single item handle for the single score point. Extended CR 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 MC and “C” if the item is a CR item. For MC 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 MC item handles at grade 4 were numbered M1 to M95.

[1] Note that the theta scale transformation used for the ALS is not the same as the one used for the NAEP reporting scale. See subsection *Scale Transformation Error* for details and implications of the error.

Table 7 shows the handles, scale values, and map values (defined in the next section) for the easiest and most difficult items within each type for grade 12. (Similar tables could be constructed for grades 4 and 8.) Some of these items have scale values outside the range of the score intervals displayed on the item map—718 to 939—and are, therefore, located in the rows or categories on the item map labeled "above" or "below."

Table 7: Example of item handles, scale values, and map values for hardest and easiest items within item type (grade 12)

Item Type	Item Handle	Scale Value	Map Value
Multiple Choice	M120	898	899
	M119	895	896
	M118	873	872
	M117	871	872
	.	.	.
	.	.	.
	M4	740	740
	M3	733	734
	M2	728	728
	M1	716	below
Constructed Response	C59_3	1012	above
	C58_3	971	above
	C57_2	937	938
	C56_2	924	923
	.	.	.
	.	.	.
	C37_2	774	773
	C8_1	769	770
	C35_1	755	755
	C37_1	741	740

The item handle for a score on an extended CR 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 C37_2 represents a score of "2" on item C37. More precisely, as can be seen in the Grade 12 Item Map in Appendix C, item C37 is the 37th most difficult extended CR item in terms of having a 0.67 probability of earning full credit on the item (a score of 4). As shown in Table 7, each score level of item C37, as well as each score level of every other extended CR 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 (OIB) corresponding to their respective scale values or map values.

Item Map Values

An item's map value (see Table 7) was the midpoint of the score interval in which the item was located on the item map. The map for each grade was divided into 74 score intervals, plus two extreme catch-all categories labeled "above" and "below." The score intervals were three units wide. For grade 4, the score intervals represented scale scores ranging from 290 to 511. (The interval midpoints ranged from 291 to 510 in steps of 3.) For grades 8 and 12, the scale score ranges represented scores from 504 to 725 and from 718 to 939, respectively. (The interval midpoints ranged from 505 to 724 and from 719 to 938 in steps of 3, respectively.) Items with scale values outside this range were represented in the "above" or "below" category. Of the 196 item scale values for grade 4, seven were represented as "above" 511, and three were represented as "below" 290. Of the 249 for grade 8, seven were represented as "above" 725 and one was represented as "below" 504; of the 245 for grade 12, two were "above" 939 and one was "below" 718.

Whole Booklet Feedback

In round 2 of the Mapmark method, feedback was given to the panelists in the form of student test booklets. Booklets were selected to represent specific ranges of performance on each test form (three per grade – a common form and one unique to each group) and given to the panelists to review. For each test form, two booklets were selected close to each cut score (i.e., Basic, Proficient, and Advanced), and one booklet was selected at the midpoint of each level, including Below Basic. The level of a booklet was determined using an expected number correct (ENC) score. The ENC score for a given scale value is given as:

$$ENC = \sum_{ik} P\left(I_{ik} = 1 \mid \eta\right), \tag{4}$$

where η is the scale value and I_{ik} is an indicator function for a score of at least k on item i. The index k will equal 1 for all MC items and range from 1 to m_i for CR items, where m_i is the total number of score points possible on the item.

The ENC is calculated for each possible scale value. Panelists are shown booklets that would be representative of students classified as being at the borderline of an achievement level. To that end, booklets were chosen so that the number of points earned is close to the ENC for the cut score for that achievement level. To identify these booklets, the ENC associated with the cut score for an achievement level was rounded to the nearest multiple of half a score scale point. If the rounded value was an integer, the two booklets were chosen at that score point. If the rounded value was not an integer, then one booklet was selected at each number correct above and below the ENC at the cut score. To demonstrate performance of a student within an achievement level, a similar method was followed, using the scale score that was at the midpoint of the achievement level. At each level, one booklet was selected with a number of points correct equal to the rounded value of the ENC associated with the scale score for 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.

The ENC score is also used to produce the Booklet Score Charts and the Booklet Score Plots. These are described in the *Materials* section of this document. Item Score Tables are based on the chosen booklets. The tables show correct/incorrect responses for the chosen booklets for each possible score point. For item *k* in student booklet *i*, the Item Score Table has the value I_{ik}, where I_{ik} is defined as in equation 4. The procedure for generating the Item Score Table and an example table are given in the *Materials* section of this report.

Consequences Feedback

Consequences feedback consists of the percentage of students scoring at or above the cut score for each achievement level for a grade. The purpose of presenting these percentages to the panelists is to give them a chance to consider whether the values seem reasonable, given what they know about the population of students at the grade level, and the achievement level descriptions (ALDs). After reviewing the consequences data in round 3 of the ALS, the panelists are given an opportunity to change their cut scores. The empirical distributions of student achievement based on the 2009 NAEP science assessment for grades 4, 8, and 12 were provided to ACT by the Design, Analysis, and Reporting (DAR) contractor in the form of relative frequency distributions. The data provided included the percentage of students at each score point, and the percentage at or below each score point on the NAEP 1 to 300 reporting scale.

Mapping Potential Exemplar Items to Achievement Levels

Potential exemplar items in the ALS meeting were drawn from two blocks for each grade level (blocks SG and SK for grade 4, blocks SJ and SK for grade 8, and blocks SI and SJ for grade 12) that had been selected for possible release to the public. Each score level above zero on an extended CR item was treated as a separate item in mapping potential exemplars to achievement levels. An item/score point was presented as a possible exemplar for an achievement level if the scale value for that item/score point fell within the range of scale values associated with that achievement level.

Panelists used an Exemplar Item Rating Form when rating the potential exemplar items. For each potential exemplar item, the form provides the item handle, the OIB page number for the item, the content area for the item, the average probability of a correct response for the item for students within the achievement level, and the probability of a correct response for the item at the Basic, Proficient, and Advanced cut scores. An example rating form is given in the *Materials* section of this report.

The average probability of a correct response for students in an achievement level for item *i* and score level *h* is calculated as:

$$\text{Average probability} = \frac{\sum_{j=C_L}^{C_{L+1}-1} \Pr(U_i \geq h \mid j) * f_j}{\sum_{j=C_L}^{C_{L+1}-1} f_j}, \qquad (5)$$

where *j* is a scale value, C_L represents the cut score for the achievement level, C_{L+1} is the cut score for the next higher achievement level, and f_j is the number of students scoring at scale value *j*. For the advanced level, C_{L+1} was set to the highest possible scale value plus

1. The values in the numerator and denominator of equation 5 can be calculated as a function of the cumulative expected probability and the cumulative distribution function. The cumulative expected frequency (CEF) at scale point *k* is defined as

$$CEF(\mathrm{k}) = \sum_{j \le k} \Pr(U_i \ge h \mid j) * f_j \,, \tag{6}$$

while the cumulative frequency is just the cumulative distribution function of the student estimates,

$$F(k) = \sum_{j \le k} f_j \,. \tag{7}$$

If, say, the Proficient cut score is at scale point *P*, and the Basic cut score is at scale point *B*, then the average expected probability for items in the Basic achievement level range is

$$\text{Average probability} = \frac{CEF(P\text{-}1)\text{-}CEF(B\text{-}1)}{F(P\text{-}1)\text{-}F(B\text{-}1)}. \tag{8}$$

The probability of a correct response for the item at the Basic, Proficient and Advanced cut scores can be calculated using the probability as given in equation 3, where η is the cut score at the achievement level.

Reliability Estimates

The term "reliability" is used here to represent the extent to which the cut scores could be reproduced if the ALS process was repeated. Cut score reliability was evaluated by examining the standard error of the cut score. More reliable cut scores have smaller standard errors.

The group median is used as the cut score in the Mapmark method, 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:

$$E[median]^k = \int x^k \binom{n}{n/2 - 1} \binom{n/2 + 1}{1} (F(x))^{n/2 - 1} (1 - F(x))^{n/2} f(x) dx \tag{9}$$

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:

$$\sum_{i=1}^{n}\left(i/n\right)^{k}\left\{F_{\beta}\left(i/n, n/2, n/2+1\right)-F_{\beta}\left(i-1/n, n/2, n/2+1\right)\right\} \quad (10)$$

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 8 below shows the standard errors for both estimators for round 1 and the final round, round 3.

Table 8: Estimates of standard error of cut scores

		Basic		Proficient		Advanced	
Grade	Statistical Method	Round 1	Final	Round 1	Final	Round 1	Final
4	Maritz-Jarrett SE	5.2	2.0	2.2	2.0	3.8	3.8
	Bootstrap SE	4.9	2.0	2.1	1.6	4.0	3.7
8	Maritz-Jarrett SE	2.2	1.5	3.3	0.8	6.0	3.6
	Bootstrap SE	2.2	1.4	3.1	0.8	5.6	3.5
12	Maritz-Jarrett SE	2.5	0.7	2.2	1.5	1.9	2.0
	Bootstrap SE	2.1	0.6	2.2	1.4	1.6	2.0

Additional analyses were conducted on the stability of cut scores across groups and panelist types. Cut scores were analyzed for each characteristic of interest, including gender, race/ethnicity, panelist type, table, and rater-group, using an ANOVA procedure. The results of the analyses showed that some of the table and rater-group (or item pool) effects showed greater differences than expected by chance. Table 9 shows the F value and the associated p-value for rounds 1 and 3 for each of the achievement levels and each of the grades for rater-groups (A or B) and tables (1-6). Note that since tables are embedded within rater groups, significance at the rater group level will often imply significance at the table level.

Table 9: ANOVA results for rater group (item pool) and table comparisons of cut scores

Source	Achievement Level	Grade	df	Round 1 F value	Round 1 p	Round 3 F value	Round 3 p
Rater Group (Item Pool)	Basic	4	1	3.35	0.08	0.00	0.95
		8	1	0.00	0.96	3.29	0.08
		12	1	15.65	< 0.01	4.82	0.04
	Proficient	4	1	12.57	< 0.01	0.66	0.42
		8	1	0.26	0.62	1.07	0.31
		12	1	0.10	0.76	0.03	0.86
	Advanced	4	1	2.65	0.11	0.00	0.98
		8	1	2.38	0.14	9.41	< 0.01
		12	1	7.75	< 0.01	1.19	0.29
Table	Basic	4	5	5.89	< 0.01	10.30	< 0.01
		8	5	0.24	0.94	2.60	0.06
		12	5	5.27	< 0.01	3.17	0.03
	Proficient	4	5	7.74	< 0.01	21.50	< 0.01
		8	5	0.59	0.71	1.52	0.23
		12	5	0.91	0.49	4.51	0.01
	Advanced	4	5	1.14	0.37	9.10	<0.01
		8	5	1.42	0.26	4.55	0.01
		12	5	6.46	< 0.01	3.77	0.01

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 typically on a 5-point Likert scale. For each question, the mean value for the responses and the standard deviation were calculated. The process evaluation questionnaires are shown in Appendix L of the Process Report (ACT, 2010).

Scale Transformation Error

As mentioned previously, the scale transformation used by ACT in the ALS differed from the transformation used for the NAEP reporting scale. The transformation from the theta scale to the NAEP-like scale used for the ALS started with the transformation

$$\text{Scale Score} = 150 + 35 * \theta$$

for all three grades, and then a different constant was added to the scale for each of the different grades (214 for grade 4, 428 for grade 8, and 642 for grade 12). The correct transformations from the theta scale to the NAEP reporting scale are given by

Grade 4 scale score = 149.664 + 36.798 * θ
Grade 8 scale score = 149.182 + 37.182 * θ
Grade 12 scale score = 149.192 + 37.397 * θ.

This error was due to a miscommunication between ACT and the DAR contractor, and was discovered when the DAR contractor and ACT were trying to reconcile results from an item classification study done by the DAR after the ALS. The three location parameters are all quite close to 150, but the scale parameters are larger than 35. Consequently, when calculating the scale value required for a probability of a correct response of 0.67, the values used for the items could be quite different from what was actually used in the ALS. Table 10 shows the differences between the scale value that was used, and the corrected scale value using the appropriate transformations.

Table 10: Differences in scale values for items due to incorrect transformation

Difference in scale values	Grade 4	Grade 8	Grade 12
	Number of items	Number of items	Number of items
0	29	39	23
1	37	73	53
2	49	45	57
3	39	37	40
4	13	16	25
5 or more	29	39	47

From the table, we see that more than half the items, in each grade, had a difference in scale values of two points or less. The items with a difference of 5 points or more were at the extremes of the items in terms of difficulty; either very easy or very hard items. In grade 12, where the proportion of items with a scale difference of 5 points or more was the highest, the items in this category were primarily the score points associated with full credit on the constructed response items.

The differences in the transformations have an effect in three primary ways. These are each described below and the implications discussed.

1. The scale value associated with item mastery that was presented to the panelists differed from the correct scale value for mastery of that item. As was seen in Table 10, this difference can be large. However, this was unlikely to have made a difference in the cut scores chosen by the panelists. The scale shown to the panelists (the NAEP-like scales) and the NAEP reporting scale are both arbitrary linear translations of the underlying theta scale. The panelists were concentrating on items, and the probability of a correct response for that item. The scale value that is associated to the item is not relevant to that judgment.

2. The item maps used in the ALS, which were supposed to be a visual representation of the differences between item difficulties, are not to scale. In order to print the item maps on a single sheet of paper, the items were grouped into score intervals that were three units wide with two extreme catch-all categories labeled “above” and “below.” The choice of where to start the grouping and the

width of the score intervals is arbitrary. The item maps as they might have appeared using score intervals three units wide and the correct transformations are shown in Appendix D. Comparing these maps to those used in the ALS (Appendix C), the items are more spread out. This is consistent with the higher values for the scale parameters in the correct transformations.

3. The consequences data shown to the panelists were incorrect. The distributions of scores provided by the DAR were based on the correct transformations. Table 11 gives the percent at or above each achievement level provided to panelists after round 3 in the ALS and the corresponding values based on the correct transformations. The differences are very small so the error is unlikely to have impacted panelists' decisions regarding reasonableness of the group cut scores.

Table 11: Consequences data presented in ALS versus that based on the correct scale transformations

Achievement Level	Percent at or Above					
	Grade 4		Grade 8		Grade 12	
	ALS	Correct Transformation	ALS	Correct Transformation	ALS	Correct Transformation
Basic	84.7	85.9	62.4	63.5	59.1	60.2
Proficient	39.5	39.5	30.3	30.3	21.8	20.9
Advanced	0.2	0.1	1.0	0.7	1.2	0.8

MATERIALS

Information on materials used in the ALS is provided in this section. For each, a brief description of the material is given, along with an illustrative example. In this document, only materials that are constructed using some type of technical information (e.g., an item handle), require some calculation, or were constructed on site at the ALS meeting are included. Additional information and descriptions of other materials can be found in the Process Report (ACT, 2010), including:

- Agenda
- Briefing Booklet
- General Contents of Ordered Item Book (OIB)
- General Contents of Constructed Response Ordered Item Book (CROIB)
- Consequences Questionnaire
- Process Evaluation Questionnaires

Division of Panelists and Item Pools into Rater-Groups/Pools A and B

The panelists and the item pools were divided into two sets, A and B, in order to minimize the fatigue effect and reduce the amount of time necessary if each panelist was required to review every item (141, 162, and 179 items for grades 4, 8, and 12, respectively). The division also creates a design that allows the reliability of the process to be evaluated (see *Reliability Estimates* subsection). At grade 4, there were 30 panelists in the ALS meeting. Fifteen panelists were assigned to group A and 15 to group B. At grade 8, there were 27 panelists; 14 panelists were assigned to group A and 13 to group B. At grade 12, there

were 28 panelists; 14 panelists were assigned to each group. Each rater group was further divided into three tables of four or five panelists each. The demographic attributes and content expertise of panelists were considered when assigning members to rater groups and tables; otherwise the assignments were random. The goal was to have rater groups and tables as equal as possible with respect to panelist type, gender, region, race/ethnicity, and content expertise.

For each grade, the item pool was divided into two similar overlapping pools. Each pool contained about 60% of the items in the grade level assessment. Items included in both pools are referred to as *common items*. Equivalence was monitored with regard to: (a) item difficulty, (b) content area representation, (c) science practices representation, and (d) item type representation. Figure 1 illustrates the division of items into two equivalent overlapping item pools for each grade.

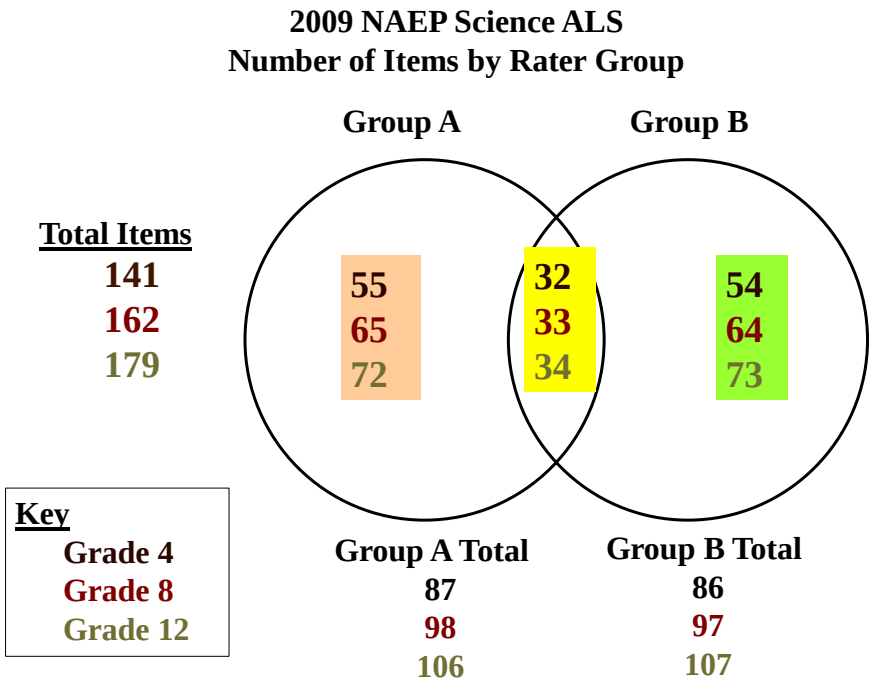

Figure 1: Numbers of items reviewed by group A, group B, and both groups A and B at each grade

The two pools were created by assigning approximately six blocks of items to each pool with two blocks in common. The common blocks are ones that were selected for possible release to the public. These were blocks SG and SK for grade 4, SJ and SK for grade 8, and SI and SJ for grade 12. The remaining blocks are assigned to groups to achieve the desired equivalence between pools. Dividing the item blocks to get similar pools is not too difficult because the blocks are generally constructed to be similar in terms of scale representation and difficulty (see Tables 2, 4, and 6).

In grade 4, item pool A consisted of complete blocks SC, SF, SG, SH, and SK, while item pool B consisted of complete blocks SE, SG, SI, SJ, and SK. The items in block SD were split into two sets. One set of items went to pool A, while the other was put into pool B.

The items in block SD were split so as to create the equivalence discussed above. In grade 8, item pool A consisted of blocks SC, SE, SF, SJ, SK, and SL and item pool B consisted of the blocks SE, SG, SH, SI, SJ, and SK. For grade 12, item pool A consisted of complete blocks SE, SH, SI, SJ, SK, and SM, while item pool B consisted of complete blocks SC, SF, SG, SI, SJ, and SL. The items in block SD were split into two sets, with half the block going into each of the two item pools. Tables 12, 13, and 14 present summaries of the grade level item pools by group and then overall. It can be seen that the item pools for groups A and B are very similar with respect to content area, science practice, item type, and item difficulty, as intended.

Table 12: Summary of item pools A and B (grade 4)

Group	ALL Items	CR Items	Points by Content Area[a]			Points by Science Practice[b]			
			E & S	Life	Phys	UP	IP	UI	UT
A	87	29	40	38	42	39	31	39	11
B	86	28	32	45	42	44	29	37	9
Total	141	46	58	69	69	66	51	62	17

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] UP = Using Principles, IP = Identifying Principles, UI = Using Inquiry, UT = Using Technology

Group	ALL Items	CR Items	Points by Item Type[a]		No. of CR Items by No. of Score Points			
			MC	CR	1	2	3	4
A	87	29	58	62	5	17	5	2
B	86	28	58	61	4	17	5	2
Total	141	46	95	101	6	28	9	3

[a] MC = Multiple choice; CR = Constructed Response

Group	Items	Item Difficulty						
		Points	Mean	SD	Min	Max	1st Quartile	3rd Quartile
A	87	120	399.2	49.3	287	597	364	421.5
B	86	119	399.8	53.2	242	556	364	431.0
Total	141	196	399.8	52.2	242	597	364	431.0

Table 13: Summary of item pools A and B (grade 8)

Group	ALL Items	CR Items	Points by Content Area[a]			Points by Science Practice[b]			
			E & S	Life	Phys	UP	IP	UI	UT
A	98	36	68	44	37	53	42	42	12
B	97	36	63	50	40	67	41	32	13
Total	162	58	102	78	69	102	67	61	19

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] UP = Using Principles, IP = Identifying Principles, UI = Using Inquiry, UT = Using Technology

Group	ALL Items	CR Items	Points by Item Type[a]		No. of CR Items by No. of Score Points			
			MC	CR	1	2	3	4
A	98	36	62	87	4	17	11	4
B	97	36	61	92	3	17	9	7
Total	162	58	104	145	5	27	18	8

[a] MC = Multiple choice; CR = Constructed Response

Group	Items	Item Difficulty						
		Points	Mean	SD	Min	Max	1st Quartile	3rd Quartile
A	98	149	616.2	43.8	497	748	586	637
B	97	153	620.6	43.1	497	828	592	641
Total	162	249	618.6	42.5	497	828	590	640

Table 14: Summary of item pools A and B (grade 12)

Group	ALL Items	CR Items	Points by Content Area[a]			Points by Science Practice[b]			
			E & S	Life	Phys	UP	IP	UI	UT
A	106	35	36	57	53	60	36	40	10
B	107	35	34	60	53	61	34	40	12
Total	179	59	60	98	87	103	59	65	18

[a] E & S = Earth and Space, Life = Life, Phys = Physical
[b] UP = Using Principles, IP = Identifying Principles, UI = Using Inquiry, UT = Using Technology

Group	ALL Items	CR Items	Points by Item Type[a]		No. of CR Items by No. of Score Points			
			MC	CR	1	2	3	4
A	106	35	71	75	8	17	7	3
B	107	35	72	75	8	17	7	3
Total	179	59	120	125	14	29	11	5

[a] MC = Multiple choice; CR = Constructed Response

Group	Items	Item Difficulty						
		Points	Mean	SD	Min	Max	1st Quartile	3rd Quartile
A	106	146	833.3	39.9	728	1012	811	860
B	107	147	835.8	40.0	716	1012	807	862
Total	179	245	834.1	40.1	716	1012	810	860

Ordered Item Book

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

The items were provided by the DAR contractor in pdf format. The items were stored in a database using the accession number as the item identifier. An item information file (created by the program *handle_book_als.sas* described in the *OIB and CROIB* subsection) contained the accession number as well as other information needed for the OIB, including the item handle, the group identifier (A or B), the content area and content statement classification, the science practice classification, the answer key, block and sequence number, and the page number. The files are merged together using the accession number as the link, and the two groups are split into separate files, using the group identifier. The items are ordered by page number, and merged into an item template page. These pages are then printed. Figure 2 illustrates how an item and its associated information are presented in the OIB. The rubrics for the CR items are also printed and inserted by hand in the OIB after the corresponding item.

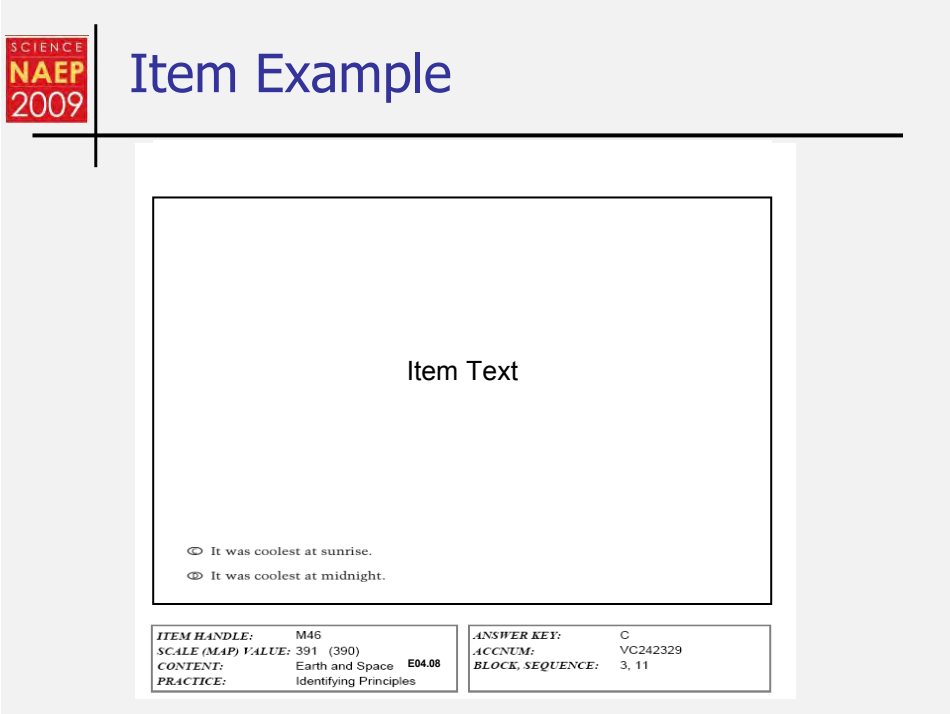

Figure 2: Illustration of the information on an OIB page

Constructed Response Ordered Item Book

The Constructed Response Ordered Item Books (CROIBs) contain the CR items in order of their scale values for a fully correct response, from easiest to hardest. Groups A and B have different CROIBs because they have different sets of items. Tables listing the item handles for the items contained in the group A and B CROIBs are included in Appendix F. The items are listed in the order they appeared in the CROIB. For each short CR item and each extended CR 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 an extended CR item contained within the same tab.

The items highlighted in yellow in the tables in Appendix F were *common items*. Four of these items were reviewed by the grade group (groups A and B combined) during stage 1 of the round 1 item review task (see Process Report, ACT, 2010), which was led by the Mapmark content and process facilitators. Subsequently in stage 2 of the round 1 item review task, the panelists reviewed the remaining items in their CROIB at the table group level.

To construct the CROIB, the items for each group were selected from the item database using the item information file to identify the CR items from the correct group. The item page was printed in the same manner as for the OIB. The rubrics and examples of student responses were included after the item. The student examples were taken from the anchor papers used for scoring of the items, which were provided by the DAR contractor. At least one example at each score point was chosen, if possible. Scoring of CR items took place prior to scaling so the student examples had item scores on the 1 to k scale and the

collapsing of score categories was not represented, making it necessary to recode and rescore student examples prior to selection.

Cut Score Recommendation Form and Computation of Cut Scores

Figure 3 shows the Cut Score Recommendation form that was used by panelists to record their cut scores. In round 1, panelists recorded their round 1 bookmark placements and associated ranges of uncertainty on this form. In rounds 2 and 3, panelists recorded their scale value selections for cut scores on this form.

Rater ID _____

2009 NAEP Science ALS
Panelist Cut Score Recommendation Form

Round 1

Basic Bookmark on Page #		Proficient Bookmark on Page #		Advanced Bookmark on Page #	
Range of Uncertainty					
Low	**High**	**Low**	**High**	**Low**	**High**

For office use only:

Basic Scale Value	Proficient Scale Value	Advanced Scale Value

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

Following 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 Tables shown in Appendix B. The scale values

corresponding to the bookmarked page numbers were entered by staff 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 a 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 grade group cut scores for that round.

In round 2 and subsequent rounds, panelists entered actual 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 a spreadsheet, and the median cut score across all panelists was computed, as in round 1.

Item Map

In the Primary Item Map for each grade, items were organized into columns corresponding to content areas of the assessment. The item maps are shown in Appendix C. In the ALS meeting, the maps were printed on 8 ½″ x 11″ paper.

Item handles on 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, *handle_book_als.sas*. The program *createmap.sas* was used to create three files, one for each content area. The file consisted of the item handles combined with the color code. Each item handle was in a row that corresponds to the item map value. A spreadsheet containing the template for the item map was used. The template had three sets of columns, one set for each content area. The possible item map values were in a column at the far left of the template. The output files from the program were pasted directly into the spreadsheet. Cells with a given color code (e.g., "G" for green) were highlighted and colored the appropriate color and the color code was removed.

Cut Score Distribution Chart

For each grade group, feedback from rounds 1, 2, and 3 included the distribution of panelists' cut scores from that round. Figure 4 shows the Cut Score Distribution Chart provided as feedback from round 1 to the grade 4 panelists. This chart was used to indicate the location of all grade 4 panelists' round 1 cut scores for each achievement level, the overlap (if any) in the distributions of cut scores for the different achievement levels, and the highest and lowest cut scores for each level. The overlap in ratings across achievement levels and the spread in ratings within each achievement level decreased across rounds.

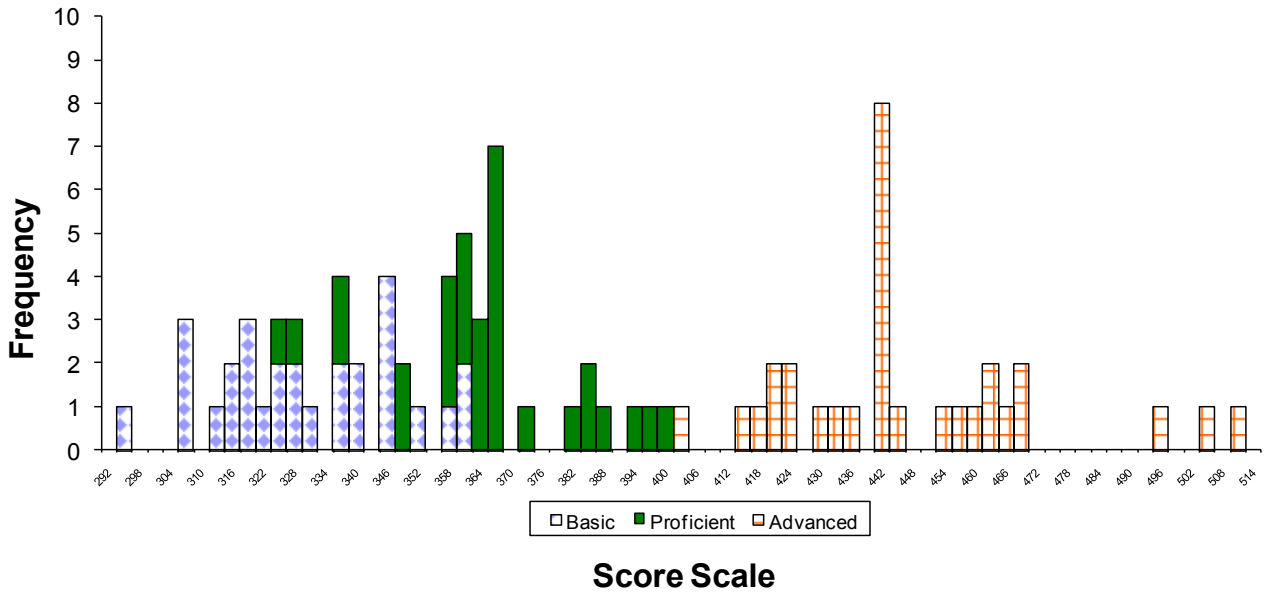

Figure 4: Cut Score Distribution Chart showing the distribution of cut scores by achievement level after round 1 for grade 4

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 10 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.

Item Score Table, Booklet Score Chart, and Booklet Score Plot

In addition to actual student test booklets in round 2, panelists received Item Score Tables, Booklet Score Charts, and Booklet Score Plots as part of the whole booklet feedback. The Item Score Table contains the score a student received (0 = incorrect, 1 = correct) for every item/score point on each student booklet included in the table. The items/score points are ordered from easiest to hardest, bottom to top, and the student booklets are ordered from lowest to highest scoring, left to right. An example of an Item Score Table is given in Figure 5.

2009 NAEP Science ALS
Item Score Table
Grade 8 - Form C

					Basic Cut			Proficient Cut			Advanced Cut		
Handle	Scale Value	OIB Page Grp A	Grp B	1C 12	2C 18	3C 19	4C 21	5C 23	6C 23	7C 31	8C 38	9C 39	10C 45
C55_3	748	148	151	0	0	0	0	0	0	0	0	0	0
C52_2	729	146	149	0	0	0	1	0	1	0	0	1	1
C50_4	716	144	148	0	0	0	0	0	0	0	0	0	0
C46_4	702	143	144	0	0	0	0	0	0	0	0	0	0
C37_3	683	135	141	0	0	0	0	0	0	1	0	0	1
C46_3	679	134	138	0	0	0	0	0	0	0	0	0	0
C31_2	672	132	134	0	0	0	0	0	0	0	1	0	1
C50_3	672	131	133	0	0	0	0	0	0	0	0	0	0
M103	667	130	132	0	0	0	0	0	0	0	1	0	1
C30_2	667	129	131	0	0	0	0	0	0	1	0	1	0
C29_2	665	128	130	0	0	0	0	0	0	0	1	0	0
C37_2	652	122	122	0	0	0	0	0	0	1	0	0	1
M101	643	118	118	0	0	0	0	1	0	0	0	1	1
C16_2	642	116	116	0	0	0	0	0	0	0	1	1	1
C46_2	636	109	107	0	0	0	0	0	0	0	1	0	1
M93	636	110	108	0	1	1	1	0	0	1	1	1	1
C31_1	632	102	101	0	0	0	0	0	0	0	1	1	1
C29_1	632	101	98	0	0	0	0	1	0	0	1	0	1
C12_2	631	100	96	0	1	0	0	0	1	0	1	1	1
C10_4	627	94	90	0	0	0	1	1	0	1	0	0	1
M81	627	95	91	0	0	1	0	1	0	0	0	1	1
M80	626	93	88	0	0	0	0	0	0	0	1	1	1
C37_1	625	92	87	0	0	1	1	0	0	1	1	1	1
C55_2	621	90	83	0	0	0	0	0	0	1	1	1	1
M73	618	87	80	0	1	0	0	0	0	0	0	1	1
M72	617	86	79	1	0	0	0	0	0	0	1	1	1
M63	610	78	70	0	0	1	0	1	1	0	1	1	1
C30_1	609	77	69	0	0	0	0	0	0	1	1	1	1
M60	609	72	68	0	0	1	1	1	0	1	1	1	1
C12_1	608	71	67	1	1	1	1	1	1	1	1	1	1
M58	607	69	63	0	0	1	0	0	1	1	0	1	1
M56	607	67	61	0	1	1	0	0	1	1	0	1	1
C16_1	601	61	51	0	1	0	0	0	0	0	1	1	1
C52_1	599	57	49	0	0	0	1	1	1	0	1	1	1
C50_2	598	55	45	0	0	1	1	1	1	1	1	1	1
C55_1	594	51	40	0	1	0	0	0	0	1	1	1	1
M35	592	48	39	0	1	0	0	0	1	1	0	1	1
C10_3	590	45	33	0	0	0	1	1	0	1	1	1	1
C10_2	585	36	30	0	0	0	1	1	1	1	1	1	1
C3	584	35	29	0	0	1	1	1	1	1	1	1	1
M25	583	30	27	0	0	0	0	1	1	1	1	1	1
C46_1	579	26	24	1	1	1	1	1	1	1	1	0	1
M19	576	22	22	1	0	0	0	1	1	1	1	1	1
C50_1	573	19	21	1	1	1	1	1	1	1	1	1	1
M15	571	17	17	1	0	1	1	1	1	1	1	1	1
M16	571	18	18	1	0	1	1	0	1	1	1	1	1
M10	566	14	13	0	1	1	1	1	1	1	1	1	1
M9	564	12	12	1	1	0	1	1	1	1	1	0	1
C2_2	562	11	10	0	1	1	1	1	1	1	1	1	1
C1	561	9	8	1	1	1	0	1	0	1	1	1	1
M2	551	6	4	1	1	0	1	0	1	1	1	1	1
C10_1	537	3	3	1	1	1	1	1	1	1	1	1	1
C2_1	497	1	1	1	1	1	1	1	1	1	1	1	1

Figure 5: Item Score Table for grade 8, form C

The values for the Item Score Table were derived from information sent by the DAR contractor. Item level scores were requested for five student booklets at each of the possible total score values on each form that was used (three forms for each grade). For some score points, at the top and bottom of the range, there were fewer than five student booklets receiving that total score. For each student booklet selected, a file was provided that had the student ID and the scored responses to each item. These item scores were 0/1 for MC items and short CR items. For extended CR items, the score was listed as a value from1 to k. A value of 8 meant the item response was missing, and a value of 9 meant the item was classified as "not reached." Because the total scores for the student booklets had been calculated prior to item calibration, each total score had to be recalculated to adjust for the items that were dropped and the items with collapsed score levels. All forms in all grades had at least one item that was affected, so the total score had to be recalculated for each student booklet for each form. Also at this stage, the response values for extended CR items were adjusted so that the response values ranged from 0 to k-1, rather than from 1 to k. This was consistent with what the panelists were shown for CR items in the CROIB and the OIB.

After the total scores were recalculated, two example student booklets at each resulting possible total score were chosen. These were chosen with the goal of minimizing the numbers of missing and not reached items, and ensuring the two chosen booklets were not too similar. The reason behind this choice was that, when reviewing the feedback, the panelists are asked to note that two students with the same score answered somewhat different items correctly and incorrectly.

For the booklets selected, the items scored as 8 or 9 are changed to scores of 0. In the Item Score Table, the item score is 0 or 1 for a MC or a short CR item. For an extended CR item, the item appears more than once in the item score table. For score point *j* associated with that item *i*, the item score table value is defined as

$$\text{item score table value} = \begin{cases} 1 & \text{if } U_i \geq j \\ 0 & \text{otherwise,} \end{cases} \tag{11}$$

where U_i is the score on item *i*.

To create the Item Score Table, a file is generated that has one line for each possible cut score. For each cut score, the two booklet scores satisfying the rule described on page 10 in the subsection *Whole Booklet Feedback* are listed, along with the item responses for the example student booklets with those two scores. Additionally, an indicator called the "tiebreaker" is included. This indicator is used in cases where only a single student booklet is required, representing performance in the middle of an achievement level range. The tiebreaker identifies which of the two student booklets should be selected when this occurs. If the two booklet scores are different, then the student booklet with the score closest to the ENC is selected. If the two booklet scores are the same, then the student booklet with the fewest number of omitted and not reached item responses is selected. The generated file is put into a database along with the item handles for the items and the OIB page numbers for those items. The Item Score Table is created by selecting the scale values associated with the cut scores for the achievement levels and keeping only the student booklet identifiers and item responses for the student booklets that are associated with the achievement level cut scores or the middle of achievement level ranges. Note that in some cases, there will not be a student booklet with the required score available. When

there is no student booklet with the required score, the column for that particular score will be blank in the Item Score Table.

The Booklet Score Charts relate performance on the items to scale values associated with the cut scores. Figure 6 shows the expected number of points correct on two booklets as a function of the achievement score scale. It also shows the location of the 20 student test booklets panelists review in round 2 in relation to the achievement scale. This chart is referred to when panelists select new cut scores in round 2, and it also provides cut score and rater location feedback. To create this document, the table shell was created prior to the ALS meeting. The shell contained all scale scores and the expected number correct (see equation 4), in multiples of 0.5. Columns were included for both the common form and the group specific form. The scale score associated with the given expected number correct in the chart was the score for which the expected number correct was closest to the given value.

The expected number correct for the whole booklets was taken from the Item Score Table. The booklet numbers were entered into the Booklet Score Chart at the appropriate scale values. For each achievement level, lines were entered at the lowest cut score from round 1 and the highest cut score. The row associated with the median cut score was shaded and labeled, and the portion of the chart starting at 10 points below the lowest cut score to 10 points above the highest cut score was printed.

Booklet Score Plots show all possible expected number correct scores on a test booklet as a function of the achievement score scale. The round 1 median cut scores are marked on this chart, as are the scale score locations of the ten student test booklets used in round 2, two at each cut score and one in the middle of each achievement level. The curve showing number correct as a function of scale value was created prior to the ALS. The labels indicating the location of the booklets are added after the cut scores are known, and are taken from the Item Score Table. An example of a Booklet Score Plot is given in Figure 7.

Booklet Score Chart - Grade 4 Group A

Proficient

		Common Form		Group A Only Form	
	Scale	Booklet	Expected No. of Points	Booklet	Expected No. of Points
	411		30.5		
	410				32.0
	409		30.0		
	408				
	407		29.5		31.5
	406				
	405	7C	29.0	7A	31.0
	404				
	403		28.5		30.5
	402		28.0		
High	401				30.0
	400		27.5		
	399				29.5
	398		27.0		
	397				29.0
	396		26.5		
	395				28.5
	394		26.0		
	393		25.5		28.0
	392				
	391		25.0		27.5
	390				
	389		24.5		27.0
	388				
	387		24.0		26.5
	386				26.0
	385		23.5		
	384		23.0		25.5
	383				
	382		22.5		25.0
	381				
	380		22.0		24.5
	379				
	378		21.5		24.0
	377				
	376		21.0		23.5
	375				
	374		20.5		23.0
	373				
	372		20.0		22.5
	371				
	370		19.5		22.0
	369				
	368	6C	19.0		21.5
	367				
Median -->	366		18.5	5A, 6A	21.0
	365				
	364	5C	18.0		20.5
	363				
	362				20.0
	361		17.5		
	360				
	359		17.0		19.5
	358				
	357		16.5		19.0
	356				
	355				18.5
	354		16.0		
	353				
	352		15.5		18.0
	351				
	350				17.5
	349	4C	15.0		
	348				
	347		14.5	4A	17.0
	346				
	345				16.5
	344		14.0		
	343				
	342				16.0
	341		13.5		
	340				
	339				15.5
	338		13.0		
	337				
	336				15.0
	335		12.5		
	334				
	333				14.5
	332	3C	12.0		
	331				
	330			3A	14.0
	329				
	328		11.5		
	327				13.5
	326				
Low	325				
	324	2C	11.0		
	323			2A	13.0
	322				
	321				
	320		10.5		12.5
	319				
	318				
	317				
	316		10.0		12.0
	315				

Figure 6: Proficient Booklet Score Chart for grade 4, group A

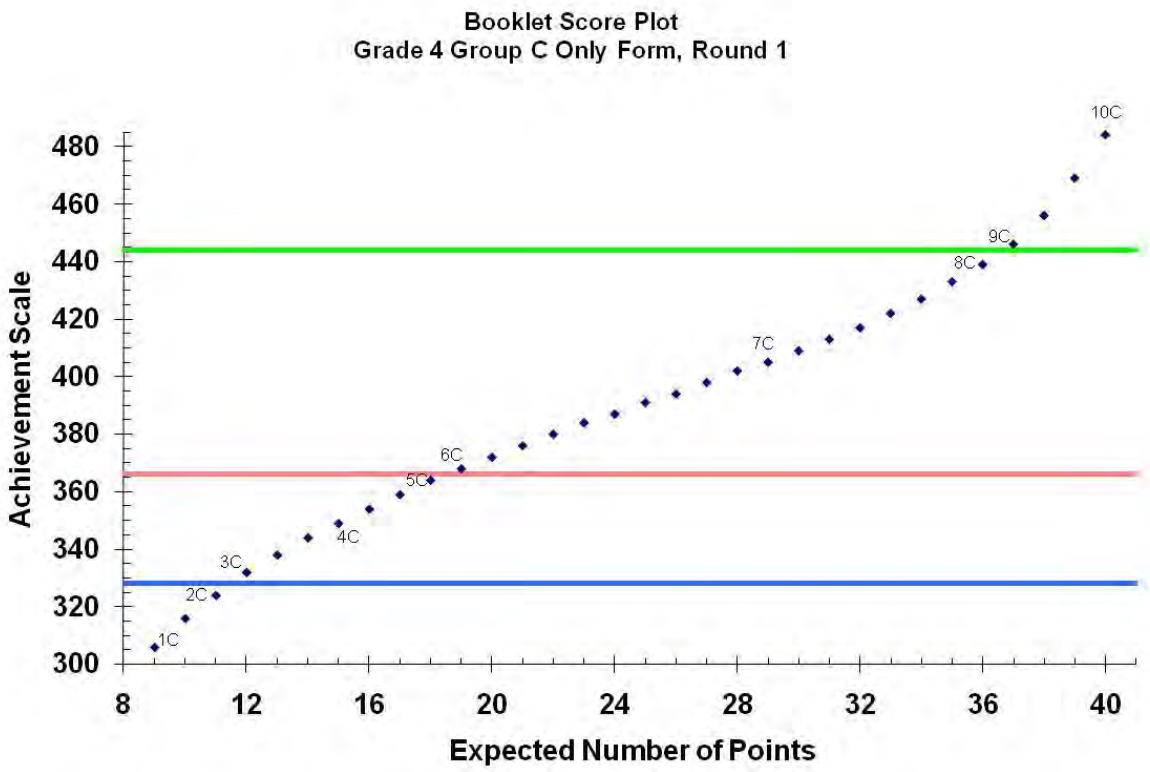

Figure 7: Booklet Score Plot for grade 4, form C

Consequences Feedback and Questionnaire

Consequences feedback was presented to panelists in the form of Figure 8. This data was displayed in a pie chart and a bar chart. The pie chart gave the percentage of students scoring within an achievement level, and the bar chart gave the percentage of students scoring at or above each achievement level. For each grade, the input data for the display were obtained from the relative frequency distributions of student performance tables provided by the DAR contractor.

After reviewing the final consequences data, a panelist 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 Appendix K of the Process Report (ACT, 2010).

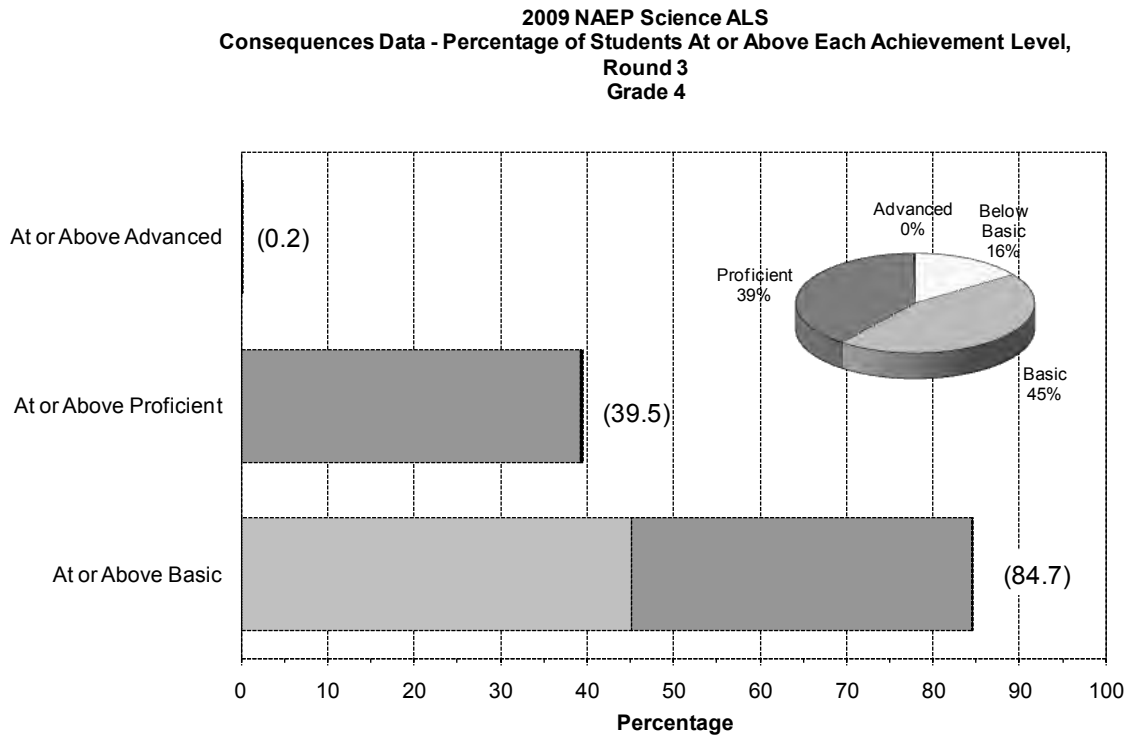

Figure 8: Sample consequences data

Exemplar Item Rating Form

An Exemplar Item Rating Form was produced for each achievement level for each grade. For each item/score point mapping to an achievement level, the form contained the item handle, the page numbers of the item in the OIBs for groups A and B, the content area for the item, the scale value of the item, the average probability of a correct response for the item for students in that achievement level, and the probability of a correct response for the item for students scoring at the Basic, Proficient, and Advanced cut scores. A database was created with this information included, and the input values were the cut scores for each of the three achievement levels. Figure 9 shows the output for the Basic achievement level for grade 4. The program that identified the achievement levels associated with each item used the round 3 median cut scores as input.

2009 NAEP Science ALS **Rater ID: _____**

Achievement Level: Grade 4 Basic

Item	OIB Page #		Science Content Area	Scale Value*	Avg Prob Correct for Basic	Probability at Cut Score			Rating as Exemplar			IF DO NOT USE -- Please Explain
	Group A	Group B				B	P	A	Very Good	OK	Do Not Use	
M14	13	14	Life	337	0.79	0.59	0.91	1				
M17	14	18	Physical	347	0.73	0.5	0.89	1				
C1	17	19	Physical	348	0.7	0.53	0.82	0.97				
C3_1	19	21	Physical	351	0.7	0.45	0.87	1				
M19	18	20	Life	351	0.69	0.53	0.82	0.98				
M20	20	22	Earth and Space	353	0.68	0.51	0.82	0.99				
M21	22	24	Life	358	0.65	0.51	0.76	0.96				
M22	24	26	Life	359	0.65	0.52	0.78	0.99				
M29	32	31	Life	364	0.62	0.48	0.73	0.95				
M27	30	30	Earth and Space	364	0.61	0.43	0.76	0.98				
M34	34	35	Physical	369	0.59	0.46	0.71	0.94				
M35	35	36	Physical	370	0.56	0.41	0.72	0.98				

*Scale value where RP = 0.67

Figure 9: Exemplar Item Rating Form for the Basic achievement level for grade 4

The results of the exemplar item rating task are given in Appendix D and Tables 49-50 of the Process Report (ACT, 2010). Table 49 summarizes the number of MC items and CR score points identified as potential exemplars and Table 40 summarizes the number of MC items and CR score points recommended for exemplars.

PILOT STUDY

The Pilot Study was conducted in October 2009. The Pilot Study used a Mapmark with Whole Booklet Feedback process similar to that used in the ALS meeting, and its purpose was to try out the procedures that were planned for the ALS meeting. Changes in the agenda and processes for the ALS meeting are documented in Appendix A and in Appendices A, F, and G of the Process Report (ACT, 2010).

The item parameters, transformation constants, and all technical procedures used for the Mapmark with Whole Booklet Feedback method in the Pilot Study are exactly as described previously in this technical report.

TACSS INPUT

Throughout the contract, TACSS provided technical advice and information. Meetings with TACSS were held at key points throughout the process to discuss plans and results, and consider next steps. A complete set of minutes for each meeting can be found in Appendix A. Some of the key technical decisions that were reached are listed here.

- ACT was asked to investigate how the panelists understand and use the RP criterion as part of the ALS process. The three research questions about panelists understanding of the RP values were discussed by the committee. COSDAM requested that information related to the three research questions be collected throughout the ALS process. Consequently, it was decided that information related to the three research questions would be gathered via the process evaluation questionnaires administered during the Pilot Study and the ALS meeting. For each research question, the type of question asked, the wording of the question, and the

timing were discussed, modified as necessary, and approved by TACSS. Results of this investigation are reported in the Process Report (ACT, 2010).

- Some items in the science assessment had multiple parts, with scores on each part combined into a single item score. The TACSS decision was to show the rubric for each part, followed by the scoring guide giving the translation from total score on all parts to item score. Panelists were to be shown examples of student work at different levels of the item score.

- Many items had score levels that were collapsed when the items were scaled. However, the rubrics and examples of student work had the information and score levels associated with the original scale, prior to this change. The decision was made to show the panelists the complete rubric, but to change the score levels to be consistent with the post-collapse scoring. Student work was to be shown for score levels corresponding to minimal changes in the item score after collapse, based on the scoring prior to collapse. As an example, if the original item was to be scored (4,3,2,1,0), and this was collapsed to (2,1,1,1,0), an example of student work at the rescored levels of 2 and 0 would be chosen (corresponding the to the original score levels of 4 and 0 respectively), and two examples of student work at the rescored levels of 1 would be chosen. One of these would represent student work at the original score level of 3, and was used to show the minimum skills required to go from a rescaled value of 1 to a rescaled value of 2. The second example at the rescored level of 1 would represent student work at the original score level of 1, and was designed to show the minimum skills required to go from a rescaled level of 0 to a rescaled level of 1.

- Part of the instructions for choosing the bookmark in the first round involved identifying a set of items that might plausibly be chosen as the bookmarked item for the proficiency level. This was referred to as the “range of uncertainty.” In order to better understand what panelists were thinking about when considering this concept, the TACSS asked that each panelist record their range of uncertainty at the same time they recorded their cut score in each round.

- Following the Pilot Study, TACSS decided that the information gathered for the range of uncertainty (see previous bullet) in rounds 2 and 3 was not useful, and it was decided to only ask for the information in round 1. Figure 3 shows the form used in the ALS meeting. Range of uncertainty results for the ALS meeting are given in Appendix G.

- After the pilot, the TACSS endorsed several changes to the process with the intent of increasing the allocation of time allotted to the CR item review task. These changes included eliminating the cross grade training sessions in rounds 1-3, reducing the time for the discussion of the framework, and spiraling the MC items among panelists within a table. Additionally, following the pilot, it was decided to change the training for CR item review to emphasize the skills required to achieve full credit down to zero credit, rather than the reverse, which had traditionally been used. It was felt that the structure of the rubrics for many items made the panelists’ task easier when done in this way.

- TACSS was consulted about the scale transformation error when it was discovered. Their recommendation was to report the scale values for the cut scores as they had been chosen by the panelists in the final round, and to adjust the percents at or above each achievement level to make them consistent with the correct scale.

COMPUTER PROGRAMS

A 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 used for meeting materials and data displays. All programs are written using the SAS software, and have the extension .sas.

To run these programs, certain input files are required. These consist of the item parameter files, the item information files, the student response data files, and the frequency distribution files. These files were provided by the DAR contractor. There are three programs that create most of the data needed for the materials used in the standard setting. These are:

- *Itemprob_1scale.sas*

  This programs calculates the probability of a score of *k* on item *i* as given in equations 1 and 2. The input file is the file of item parameters, and there are three output files, one for each grade. The output files contain for each scale point the probability of a score of *k* on item *i* conditional on the scale value, for *k=1,.. ,K*, where *K* is the maximum score on the item, matched with the item identifier, the type of item (MC or CR), the score point value, and the maximum number of score points for that item. The files are called *item_probs_equalto_compscale_GXX*, where *XX* is the grade level.

- *Itemscalevalue.sas*

  This program calculates the probability of achieving a score of *k* or greater on item *i*, for each score point associated with an item, along with the scale value for that item, as defined in equation 3. The input files are the output files from the program *itemprob_1scale.sas*. There are two types of output files, one of each type for each grade. The first files, called *items_probs_greaterorequalto_compscale_grXX,* where *XX* is the grade level, have the item identifier, the step value, the item type, the maximum score on the item, and for each scale point, the probability of a score of *k* or greater, for the *kth* step on each item, conditional on the scale value. The second set of output files, called *item_scale_values_GXX* have the item identifier, the step value, the item type, the maximum score on the item, and the item scale value.

- *Match.sas*

  This program matches the item information provided by the DAR contractor to the item scale value. There are two input files. One is the item level information provided by the DAR contractor, and includes the item identifier, the content area, the science practice associated with the item, the item type (MC or CR), the grade level for the item, the number of score points associated with the item, the block identifier in two formats, the item sequence number, the key (for MC items), and

the statement number that identifies the content from the framework. The other file is the scale value file created by the program *itemscalevalue.sas*. These two files are matched by item identifier, and some additional variables are created including the scale value on the NAEP-like scale that was used in the ALS, and the scale value associated with the location of the item on the item map. The merged file is output in a file called *item_characteristic_GXX*, where *XX* is grade level, with one file for each grade.

The output from these three programs is used as the basis for creating the materials needed for the ALS.

OIB and CROIB

The input for the item books (i.e., OIBs and CROIBs) is created in the program *handle_book_als.sas*. This program uses the output from the *match.sas* program as input. The program assigns an item handle to each item, breaks the items into two groups, representing the two different rater groups, identifies the color of each item for the item map, and determines the item pages that the item will appear on in the book. There are two output files created. One is a file that contains the item information with the new variables appended. These are called *item_information_file*. The other file is the same information, saved as a SAS data set. This file was transformed to a spreadsheet, and read into a database that creates the OIB and CROIB.

Item Maps

The template for the item maps already exists, and the items need to be placed onto the map in the correct location. This is done using the program *createmap.sas*. This program uses the *item_information_file* as the input. It takes the item handles with the color appended, and puts all items with the same map value onto the same row. This file is created separately for each content area. The file is output, read into a spreadsheet, and pasted directly onto the map. The color indicator is removed as the cell is shaded. A smaller program, *mapfit.sas*, is used to evaluate the number of columns needed for the map. It simply counts the number of items that will fall onto any row, for each of the content categories.

Whole Booklet Feedback

The whole booklet feedback requires a method for choosing the booklets and creating the Item Score Tables and Booklet Score Charts for these booklets. This is done using three separate programs. The first program is *modify_wholebookletscores_GRXX_BYYY.sas*, where *XX* represents grade level and *YYY* represent the booklet number. This program uses the item scores on the booklets sent by the DAR contractor. This data needs to be modified to adjust for the fact that CR items are scored from *1* to *n*, rather than from *0* to *n-1*. Additionally, the final score is based on the items before deletion of certain items, and collapsing of score points for other CR items. The program uses as input the files *grXX_BYYY_original.prn*. This file contains the booklet ID, the item responses, the number of items answered, the number omitted, and the total score. After the adjustments mentioned previously, a new total score is calculated. A similar file is output as *grXX_BYYY_revised* with the new item responses, and a new total score. Note that the items scored as *8* or *9*, representing omits and not reached respectively are left as is. If possible, booklets with large numbers of omitted or not reached items are not selected, so this information is necessary. Once this program has been run, the output is viewed and

two booklets per total score are selected and saved in a file called *grXX_BYYY_reduced*. In this file, the values *8* and *9* are replaced by *0*.

The whole booklets are chosen based on the expected number correct, as calculated in equation 4. The correspondence between a scale score and the expected number correct is calculated using the program *avgscore_byscrpt_booklet.sas.* The input files for this program are *item_information_file* and *item_probs_greaterorequalto_compscale_GXX*. The *item_information_file* is used to identify the items in the booklets, by using the blocks that make up those booklets. The data in the file *item_probs_greaterorequalto_compscale_GXX* is used to calculate the expected number correct for those items. The two raw scores that are closest to this expected number are also identified, using the rules given above. These values are output to the file *gradeXX_bookYYY_bookIDs*. This file is used to create the Booklet Score Chart, identifying the scale value where the expected number correct is closest to an integer value or the integer value plus 0.5.

The final program creates the data needed for the item score tables. This program, *create_wholebooklet_grXXbookYYY.sas,* has as its inputs, the *grXX_BYYY_revised* file containing the item scores for the chosen books, and the file *grXX_bookYYY_bookIDs*, containing the raw score values needed for each scale value. The program takes the item scores and creates a 0/1 variable for each item step. For MC items, this is equal to the value from the *grXX_BYYY*-revised file, and for CR items, it is 0 if the item score is below the step value, and 1 if the item score is greater than or equal to the step value. This 0/1 file is ordered by item difficulty, with the most difficult items first. Then, for each scale point, two of these 0/1 vectors are identified, corresponding to the raw scores needed for that scale value as given by the file *grXX_BookYYY_bookIDs*. These vectors are output in a single row. A header row is also created which has the item handles, starting with the most difficult item first. There are two output files created. The header information is in the file *GrX_bookYYY_wbf_header*, and the item response data are in the file *GrX_bookYYY_responsesforfeedback*. This file has the scale value, the two raw scores required, an indicator giving which of the two booklets is preferred if only one booklet is needed at that scale value, and the item responses. These two files are input into a spreadsheet file, and the header is inserted as the top row. This file is used as input into the database that creates the Item Score Tables.

Exemplar Item Charts

The final piece of feedback is the Exemplar Item Rating Form. The data for these forms were created by the program *exemplar_file_create_revised.sas*. This program uses the *item_information_file* and the *item_prob_greaterorequalto_compscale_grXX* file as input. The *item_information_file* is used to identify the items that will be used, based on the block. The *item_probs_greaterorequalto_compscale_GXX* file is used to store the probability of a correct response at the scale value, and the information used to get the expected proportion of students mastering an item within an achievement level. The output file, saved as a SAS data file, includes for each item, for every score scale from 1 to 300, the item handle, the page number in the OIB that the item appears on (for both groups), the content category, the scale value of the item, the probability of a correct response at that scale value, and the information used to calculate the expected proportion of students within an achievement level mastering the item. This file is then put into a spreadsheet and this file is used as input for the database from which the Exemplar Item Rating Form is created.

REFERENCES

ACT (2010). *Developing achievement levels on the 2009 National Assessment of Educational Progress in Science: Process report.* Iowa City, IA: Author.

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.