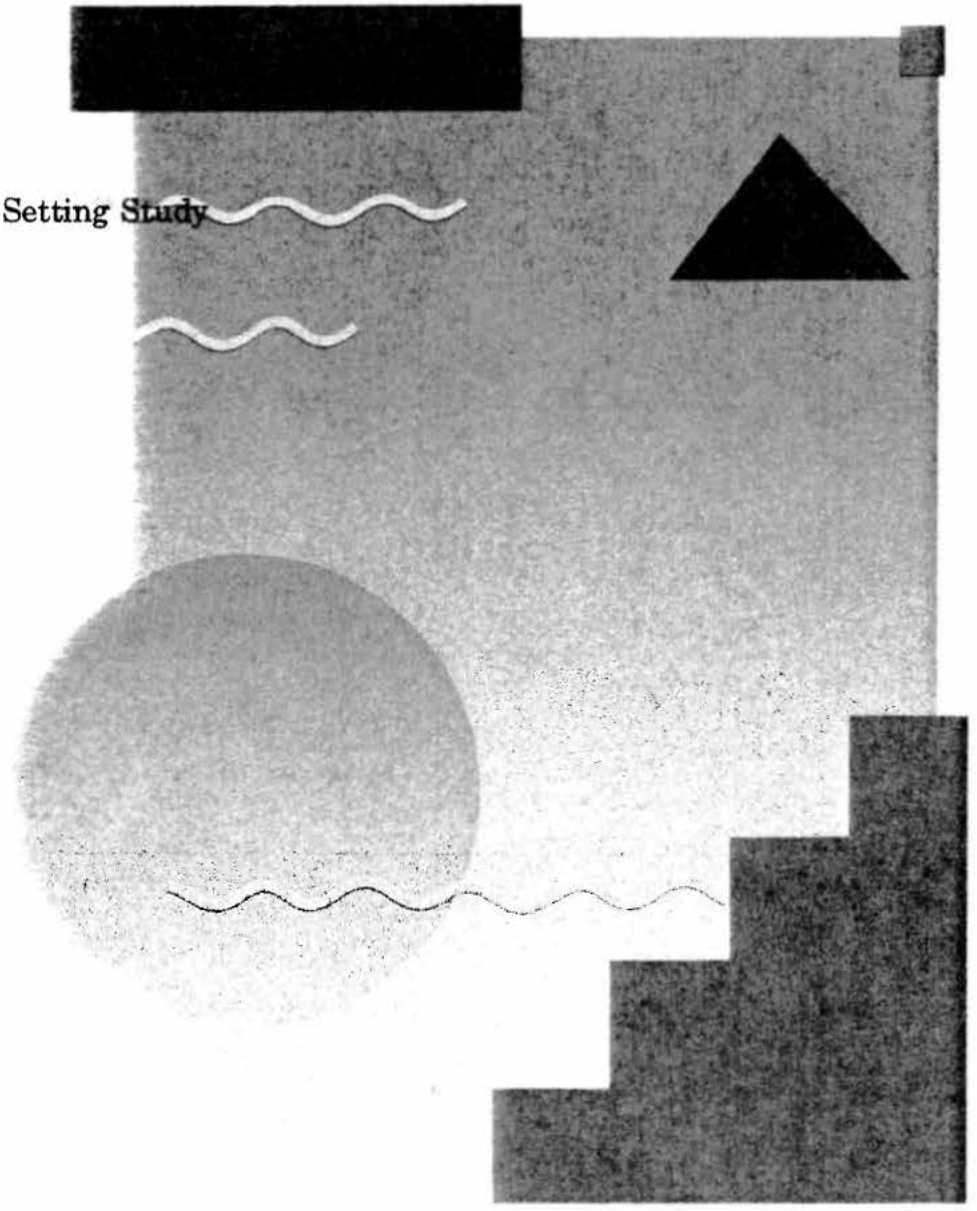
ACT

Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science

Final Report



Volume III
Achievement Levels Setting Study

Presented by ACT
May 1997

The National Assessment Governing Board

Honorable William T. Randall, Chair
Commissioner of Education
State of Colorado
Denver, Colorado

Mary R. Blanton, Vice Chair
Attorney
Salisbury, North Carolina

Patsy Cavazos
Principal
W.G. Love Accelerated Elementary School
Houston, Texas

Catherine L. Davidson
Secondary Education Director
Central Kitsap School District
Silverdale, Washington

Edward Donley
Former Chairman
Air Products & Chemical, Inc.
Allentown, Pennsylvania

Honorable James Edgar (Designate)
Governor of Illinois
Springfield, Illinois

James E. Ellingson
Fourth-grade Classroom Teacher
Probstfield Elementary School
Moorhead, Minnesota

Thomas Fisher
Director of Student Assessment
State of Florida
Tallahassee, Florida

Michael J. Guerra
Executive Director
Secondary School Department
National Catholic Education Association
Washington, DC

Edward H. Haertel
Professor
School of Education
Stanford University
Stanford, California

Jan B. Loveless
District Communications Specialist
Midland Public Schools
Midland, Michigan

Marilyn McConachie
Former School Board Member
Glenbrook High Schools
Glenview, Illinois

William J. Moloney
Superintendent of Schools
Calvert County Public Schools
Prince Frederick, Maryland

Honorable Annette Morgan
Former Member
Missouri House of Representatives
Jefferson City, Missouri

Mark D. Musick
President
Southern Regional Education Board
Atlanta, Georgia

Mitsugi Nakashima
Former President
Hawaii State Board of Education
Honolulu, Hawaii

Michael T. Nettles
Professor of Education & Public Policy
University of Michigan
Ann Arbor, Michigan
and Director
Frederick D. Patterson Research Institute
United Negro College Fund

Honorable Norma Paulus
Superintendent of Public Instruction
State of Oregon
Salem, Oregon

Honorable Roy Romer
Governor of Colorado
Denver, Colorado

Honorable Edgar D. Ross
Former (State) Senator
Christiansted, St. Croix
U.S. Virgin Islands

Fannie L. Simmons
Mathematics Coordinator
District 5 of Lexington/Richland County
Ballentine, South Carolina

Adam Urbanski
President
Rochester Teachers Association
Rochester, New York

Deborah Voltz
Assistant Professor
Department of Special Education
University of Louisville
Louisville, Kentucky

Marilyn A. Whirry
Twelfth-grade English Teacher
Mira Costa High School
Manhattan Beach, California

Dennie Palmer Wolf
Senior Research Associate
Harvard Graduate School of Education
Cambridge, Massachusetts

Ramon C. Cortines (Ex-Officio)
Acting Assistant Secretary of Education
Office of Educational Research
and Improvement
U.S. Department of Education
Washington, DC

Roy Truby
Executive Director, NAGB
Washington, DC

Daniel B. Taylor
Contracting Officer
Washington, DC

Mary Lyn Bourque
Contracting Officer's Technical
Representative
Washington, DC

Achievement Levels Committee
Michael T. Nettles, Chair
James E. Ellingson
Thomas Fisher
Norma Paulus
Honorable Roy Romer
Deborah Voltz

Table of Contents

Introduction 1

Panelist Selection 3

Item Pools 4

The Achievement Levels-Setting Process 5
Orientation 5
Training in the Frameworks and Achievement Levels Descriptions 5
Training in Rating Procedures 7
Item Rating Methods 7
Feedback 7
Selection of Exemplar Items 8

Results 9
Achievement Levels Descriptions 9
Cutpoints 10
Overall Cutpoints 10
Analyses of Effects Associated with Panelists 11
Rating Groups 11
Common Blocks 11
Cross-Grade Blocks 11
Demographic Characteristics of Panelists 12
Analyses of Effects Associated with Items 12
Hands-On Tasks 12
Block Types 12
Content Area 12
Content Area and Panelists' Expertise or Special Interest 13
Exemplar Items 13

Other Results 13
ALD Location Exercise 13
Consequences Questionnaire Data 14
Process Evaluation Data 15

Grade 8 Reconvention 18

Public Commentary 19
Numerical Cutscores Represented by Percentage of Total Points Scored 19
Achievement Levels Descriptions 20
Exemplar Items 20

References 21

List of Appendices

Appendix A: List of Panelists
Appendix B: Information About Assessment Items
Appendix C: Division of Item Pools
Appendix D: Agenda
Appendix E: Briefing Booklet
Appendix F: List of Staff and Observers
Appendix G: Facilitators' Outline
Appendix H: Feedback and Other Information
Appendix I: Achievement Levels Descriptions
Appendix J: Exemplar Items - ALS
Appendix K: Consequences Questionnaire Report - ALS
Appendix L: Process Evaluation Questionnaire Data - ALS
Appendix M: Exemplar Items - Reconvention
Appendix N: Consequences Questionnaire Report - Reconvention
Appendix O: Process Evaluation Questionnaire Data - Reconvention
Appendix P: Public Comment Forum Materials

List of Tables

Table		Page
1	Distribution of Nominee Pool	23
2	Distribution of Panelists	24
3	1996 NAEP Science Items	25
4	Cutscores and Standard Error	26
5	Averages and Standard Deviations of Rater's Locations	27
6	Cutpoints and Percentages of Students At or Above: Grade 4	28
7	Cutpoints and Percentages of Students At or Above: Grade 8	29
8	Cutpoints and Percentages of Students At or Above: Grade 12	30
9	Cutscores and Percent Correct Data	31
10	Comparisons of Cutscores and SDs by Rating Group	32
11	Comparisons of Cutscores and SDs of Common Blocks	33
12	Comparisons of Cutscores and SDs with Cross Grade Blocks	34
13	Comparisons of Average Ratings with Cross Grade Blocks	35
14	Comparisons of Rater Locations by Regions	36
15	Comparisons of Rater Locations by Types	37
16	Comparisons of Rater Locations by Gender	38
17	Comparisons of Rater Locations by Panelist Ethnicity	39
18	Comparisons of Cutscores and SDs with Hands-on vs. Non-Hands-On	40
19	Comparisons of Cutscores and SDs with HO, TB, and CP Blocks	41
20	Cutscores and Standard Deviations by Item Type	42
21	Percent of Students' Performance At or Above Each Achievement Level	43
22	Comparisons of Rater Locations for Item Content Area: Grade 4	44
23	Comparisons of Rater Locations for Item Content Area: Grade 8	45
24	Comparisons of Rater Locations for Item Content Area: Grade 12	46
25	Exemplar Items: Grade 4	47
26	Exemplar Items: Grade 8	48
27	Exemplar Items: Grade 12	49
28	ALD Location Exercise	50
29	Cutscores Deviations of Panelists Recommending Lower Cutscores	51
30	Knowledge and Skills Assessed by Different Item Types	52
31	Understanding of Student Performance at Each Achievement Level	53
32	Conception of Student Performance at the Borderline	55
33	Grade 8 Reconvention Cutpoints	56
34	Cutpoints and Percent of Students At or Above: Reconvention	57
35	Results from Grade 8 Reconvention	58
36	Average Difference Between Cutpoints from ALS & Reconvention	59
37	Exemplar Items: Reconvention	60
38	Comparison of Responses to Selected Evaluation Questions	61

List of Figures

Figure		Page
1	Distribution of Nominee by Content Area Specialty/Interest	63
2	Distribution of Panelists by Content Area Specialty/Interest	64
3	Achievement Levels Descriptions Chart: Grade 4 Report	65
4	Achievement Levels Descriptions Chart: Grade 8 Report	67
5	Achievement Levels Descriptions Chart: Grade 12 Report	69
6	Final Achievement Levels Descriptions	71
7	Final Borderline Descriptions	75
8	Exemplar Items Presented to Panelists: Grade 4	80
9	Exemplar Items Presented to Panelists: Grade 8	81
10	Exemplar Items Presented to Panelists: Grade 12	83
11	Exemplar Items Selected by Panelists: Grade 4	85
12	Exemplar Items Selected by Panelists: Grade 8	86
13	Exemplar Items Selected by Panelists: Grade 12	87
14	Percentage of Students At or Above Each Level: Grade 4	88
15	Percentage of Students At or Above Each Level: Grade 8	89
16	Percentage of Students At or Above Each Level: Grade 12	90
17	Understanding of Student Performance	91
18	Conception of Student Performance	91
19	Clarity of Rating Methods	92
20	Ease of Applying Rating Methods	92
21	Clarity of Rating Methods (Hands-On)	93
22	Ease of Applying Rating Methods (Hands-On)	93
23	Final Reconvention Descriptions	94
24	Final Reconvention Borderline Descriptions	95
25	Exemplar Items Presented to Reconvention Panelists	97
26	Exemplar Items Selected by Reconvention Panelists	99
27	Percentage of Students At or Above Each Level: Reconvention	100

Introduction

On September 21-26, 1996, ACT convened a panel to set achievement levels for the 1996 National Assessment of Educational Progress (NAEP) in Science. The achievement levels-setting (ALS) process was held at the Ritz-Carlton Hotel in Phoenix, AZ. The ALS process and the outcomes of that process are described in this volume.

Prior to the ALS meeting in September, two pilot studies for the 1996 Science NAEP ALS had been conducted. The first pilot study was convened in St. Louis in March 1996 with 29 panelists at grade 12, 10 at grade 8, and 18 and grade 4. The primary purpose of the first pilot study was to determine the interactions among panelists, the process, and the science assessment. The Science NAEP included an experiment for each student tested. Each test booklet included a set of questions (a block of test items) related to the performance of an experiment. There were four different experiments for each of the three grades tested. Students were provided all the equipment needed for conducting the experiment. The "hands-on task" was the final set of items in the test booklet taken by each student. Panelists generally found those hands-on tasks to be very interesting and a positive aspect of the assessment. They rated them as relatively easy, and indicated that they had no special difficulty with rating the hands-on task items.

The second pilot study was conducted in August 1996. That study was conducted to test all procedures to be implemented in the ALS process, and to determine whether any additional modifications or adjustments in the process would be needed before convening the ALS panels in September. Twenty panelists at grade 4, 18 at grade 8, and 19 at grade 12 were convened for the second pilot study. Again, no significant difficulties with the process emerged, and no major modifications were made to the process. Most of the changes that were introduced were in response to panelists' needs for more time in the process. The number of items included in various training exercises and procedures was reduced, and special instructions were developed to facilitate completion of tasks by panelists.

The full reports for the two pilot studies are included in Volumes I and II of this report.

Ninety-four panelists participated in the ALS meeting: 29 at grade 4, 33 at grade 8, and 32 at grade 12. These 94 panelists were selected from a pool of 755 nominees. The panelists at each grade level were divided into item rating groups and items were divided into item rating pools. Half the panelists at each grade level rated over half the items in the item pool for a grade. Panelists were divided into the two groups so that panelist type (teachers, other educators, and general public), race/ethnicity, sex, and region were represented as equally as possible in the two groups. Similarly, item blocks (units averaging approximately 9 items for grade 4 and 13 for grades 8 and 12) were divided into two sets so that item characteristics such as difficulty (probability of correct response), content area, number of hands-on

tasks, and item format (multiple choice, short constructed response, extended constructed response) were as equivalent as possible.

Item-by-item rating methods were used to rate all the items. Panelists were engaged in training exercises for approximately three full days prior to the first of three item rating sessions. The training was designed to assure that panelists understood the contents of the assessment framework and shared a common understanding of the definitions of achievement at each level, to help panelists reach a common agreement on "borderline performance," to become familiar with the scoring rubrics and their use for each of the items in the grade-level assessment, and to give panelists a "reality check" with respect to student performance on constructed response items.

Panelists completed a process evaluation at the end of each day or after completing a major step in the process. Those evaluations were reviewed each day to monitor any potential problems or difficulties for panelists. A summary of panelists' responses to the evaluation items is discussed in the "Other Results" section of this report.

Following the final round of ratings, panelists were provided "consequences data" that showed estimates of the percentages of students scoring at or above each level. A questionnaire was designed to collect opinions of panelists regarding the correspondence between these estimates and their own expectations as to the consequences of the levels they had set.

As a result of the comments shared on the "consequences" questionnaire, the grade 8 panel was reconvened November 23-24, 1996 in St. Louis. Forty-five percent of the grade 8 panelists had recommended changes in the cutscores set. In particular, they recommended that the cutscores for the Proficient and Advanced achievement levels be lowered so that the percentage of students scoring at or above each level would be larger. In addition, the grade 8 panelists indicated on their final process evaluation questionnaire that they were somewhat dissatisfied with the selection of exemplar items because they felt they had not had enough time to consider each item carefully before making their selections.

Twenty-five of the original 33 grade 8 panelists were reconvened and allowed to make modifications to the achievement levels descriptions, the numerical cutscores, and the exemplar items. Results of the reconvention were shared with the remaining eight panelists and they were asked to comment on those results. The results produced by the reconvened grade 8 panel are reported in the latter part of this volume.

Panelist Selection

The panelist selection process is described more completely in the ***Design Document***.[1] Principles of sampling were used for drawing stratified random samples of school districts from a national database. The samples were stratified to represent the four NAEP regions approximately equally, districts with enrollments of at least 50,000 students, and districts with at least 25% of the population below the poverty level. Three samples without replacement were drawn for identifying ALS nominators: one sample each from which nominators of panelists of the three types (teacher, other educators, and general public representatives) were identified.

Nominators of teachers were district superintendents, leaders of teacher organizations, or head persons of private schools. In addition, state science curriculum directors were invited to nominate teachers from any district within their state.[2] Nominators of the other educators included various persons who were career educators but not K-12 classroom teachers. Nominators of the general public included the mayor (or other chief elected official of the primary city in the school district), the school board president, and the chair of the education committee of the local Chamber of Commerce. Additionally, employers who employ persons with science backgrounds were identified for the districts drawn for general public and were asked to submit nominations.

A total of 755 persons were nominated as panelists. Panelists were selected to serve at a specific grade level, although they were occasionally nominated at more than one grade level. Panelists were nominated to represent a specific perspective: grade-level classroom teacher, other educators, and general public. Table 1 shows the distribution of the nominee pool.

The specialty of panelists with respect to the content areas of the assessment was considered in the selection process. For the science assessment the content areas are the fields of science that were specified in the framework: Life, Earth, and Physical Sciences. Nomination forms requested information regarding the field of expertise or special interest. The nominator was asked to check all that applied to each nominee[3]. A Venn diagram of the nominees' expertise or special interest is in Figure 1.

---

[1] ACT, Inc., *Design Document: Setting Achievement Levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science* (Iowa City, IA: Author, 1993).

[2] Only science curriculum directors from states for which one or more school districts had been drawn in the "teacher nominator sample" were invited to nominate teachers in their state as panelists. Unlike other nominators of teachers, these persons could nominate teachers from the state "at large," rather than specific districts drawn in the sample.

[3] When selected nominees were interviewed on the telephone, their content area specialty or interest was verified.

ACT project staff developed a computerized process for selecting panelists from the pool of nominees. Panelists were rated according to their qualifications based on information provided on the nomination form. Panelists with the highest qualifications ratings had the highest probability of being selected, other things being equal. The selection program was set to yield panels with 55% of the members representing grade-level classroom teachers, 15% representing other (nonteacher) educators, and 30% representing the general public. The distribution by panelist type was only one requirement to be met, however. The selection was set to yield panels with 20% "minority" racial/ethnic representation, up to 50% males, and regional representation approximately equal to the proportion of the U.S. population in each of the four NAEP regions.

Some of the persons who were selected could not serve at the time required. Data for the panelists who actually participated in the ALS process are presented in Table 2 and Figure 2. The 94 panelists represented 39 states and the Virgin Islands. A list of ALS panelists is in Appendix A. This appendix includes summary information about "outstanding" credentials for teacher panelists, and a list of occupations of general public panelists.

Panelists for each grade were divided into two rating groups (A and B) that were equivalent with respect to panelist type, sex, region, race/ethnicity, and area of interest or expertise. Each rating group consisted of three table groups that were also as equivalent as possible with respect to the above variables.

Item Pools

Three types of item blocks were included in the 1996 NAEP Science Assessment: Concept/Problem Solving (CP), Theme-Based (TB), and Hands-On (HO). The first two types were paper-and-pencil blocks with both multiple-choice (MC) and constructed response (CR) items. Each CP block included items from the three fields of science (Earth, Physical, and Life) identified in the Framework. TB blocks included items related to one theme identified in the Framework, and most of these blocks included items from a single content area. HO blocks involved performing a task and responding to items related to the task or the results of the experiment performed to complete the task. For each grade level, there were eight CP blocks, three TB blocks, and four HO blocks, for a total of 15 blocks. Summary information about items in each block and other item classification information are in Appendix B. Other properties of the 1996 NAEP Science Assessment items are discussed in Bay, Chen, and Loomis (1997b). This paper is included in Volume IV of this report.

The item pool for each grade level was divided into two parts so that one half of the panelists rated one set of the items and the other half rated the other set of items, with some blocks of items rated by both groups. This design feature provided the opportunity to examine ratings from each group as a replication of the ratings of the other half.

The item pools were constructed so that they were as nearly like one another as possible with respect to average p-value and distribution of items across content area and item type. Because the block types were very different, the division of each grade level item pool was first done within block type. And because there were to be blocks common to both rating groups within block type, each item rating pool had about two-thirds of the items with one-third of the items common to both rating pools. Summary statistics of the item pools are in Table 3. Information about the blocks assigned to each rating pool is in Appendix B. The division of item pools is discussed in more detail in Appendix C.

The Achievement Levels-Setting Process

Orientation

Panelists selected through the previously described process were convened in Phoenix for a six-day ALS process[4]. During the opening sessions, panelists were provided a complete overview, *via* a computerized presentation, of the process and the procedures to be followed. Each panelist was given a *Briefing Booklet* that described each task to be performed during each session, the purpose for the task, and how to perform the task.[5] In addition, a "glossary" of terms was included to help panelists with new and jargon terms.

During the first day, a presentation by NAGB staff described the NAEP program, including an explanation of the government agencies involved, the contractors, and the interrelationships of each. That presentation traced the development of the NAEP for the specific subject. During the first day, panelists were administered the NAEP exam and allowed to self-score the exam.

Training in the Frameworks and Achievement Levels Descriptions

Each grade-level panel was led by a process facilitator and a content facilitator. All staff participated in both the second pilot study and the ALS meetings[6]. Content facilitators provided training in the Frameworks and the preliminary achievement levels descriptions. They presented the preliminary achievement levels descriptions, along with a description of the conceptual and philosophical foundations of the assessment Framework.

---

[4] The meeting agenda is in Appendix D.

[5] The ***Briefing Booklet*** was distributed as part of the advance materials sent to each panelist approximately two weeks prior to the meetings. Other advance materials sent to all panelists included a ***Summary Design Document*** that had been recommended as a "user friendly description of the process" during public comment sessions held during the spring of 1994. A copy of the ***Briefing Booklet*** is in Appendix E.

[6] A list of ALS staff and observers is in Appendix F.

Training in the Framework and preliminary descriptions was concentrated during the first three days of the process. Panelists were engaged in exercises designed to give them experience in working with the achievement levels descriptions (ALDs). Sessions were scheduled to discuss the ALDs, and an opportunity was provided for panelists to modify the descriptions before each round of ratings.

The first ALD exercise involved using a chart that helped the panelists examine the alignment of descriptors across the three achievement levels. This was the first step in determining whether modifications and refinements to the preliminary descriptions were necessary, in addition to gaining a common understanding of the ALDs.

There were two more exercises aimed at helping panelists continue to gain a common understanding of the ALDs. Both of these exercises used assessment items as tools. In the first exercise, panelists used their understanding of the ALDs to estimate student performance for each item in a block. The items used in this exercise were not included in their rating pool. In the second exercise, panelists applied their understanding of the ALDs more holistically. They were given a sample of student booklets to review. Then they were asked to determine whether the performance exhibited in each test booklet should be classified as Basic, Proficient, or Advanced based on their understanding of the ALDs. The exercises were performed individually, and a group discussion followed.

Since the main focus in setting achievement levels was on performance at the lower boundary of each achievement level, it was important that panelists become familiar with and clear about the concept of borderline performance. The concept of borderline performance was introduced in a general session, and panelists in each grade level developed operational descriptors of borderline performance for each achievement level.

The exercises and training provided to panelists were designed to help them become more familiar with the items in the NAEP for their grade level and the scoring rubrics for the items, as well as to become familiar with the Frameworks and descriptions. Prior to the first round of ratings, panelists were asked to review student responses written to extended constructed response questions. This exercise was especially designed to provide panelists with a "reality" check on student performance, to allow them to become more familiar with the scoring rubrics for these items, and to help them form a concept of borderline performance for each achievement level.

Training in Rating Procedures

Instructions for each exercise and task were provided in general sessions so that all panelists were given the same instructions. Process facilitators reviewed the instructions and answered questions from panelists in the grade-level sessions. All

tasks were performed by panelists in grade-level sessions. To ensure that grade level facilitators provided uniform instructions, a very detailed outline of the achievement levels-setting process was provided to them. (Please see Appendix G.) Instructions were also provided using overhead transparencies so panelists could check to see if they were doing the right thing during each part of the procedure.

Item Rating Methods

Two different rating methods were used in the item-by-item rating process. For the first round of ratings, panelists were instructed to think about how they would answer each item and to check the scoring guides to determine the correct answers before they decided on the rating to give an item for each achievement level. The modified-Angoff method was used to rate dichotomous items. Panelists were instructed to think of the description of what students should know and be able to do at each achievement level (Basic/Proficient/Advanced) and to focus on the minimum level of performance required to reach that level of achievement. Then they were instructed to estimate the percentage of students at the lower borderline of the achievement level who would answer the item correctly. To rate constructed response items, panelists were instructed to estimate the average score on each item for students performing at the borderline of the achievement level.

Three rounds of ratings were collected. The item ratings were entered, verified, and analyzed on site. Ratings on the first two rounds were provided by panelists independently. They were allowed to modify their ratings during each round after being provided with feedback information based on their previous round of ratings. They were also given the chance to modify or refine their ALDs and borderline descriptors prior to each round of ratings.

For the third round of ratings, panelists were allowed to discuss their ratings for particularly troublesome items with other panelists. Examples of rating forms are included in Appendix H.

Feedback

Feedback information provided to the panelists after the first round of ratings were the following:

Cutpoints and Standard Deviations

The *cutpoints* were the combined ratings over all raters and all items for each achievement level for each grade. The *standard deviation* was the indicator of how different the cutpoint for each rater was with respect to the overall grade level cutpoint. This feedback was provided to the panelists graphically.

P-Values

A list reporting the percentage of students who gave the correct answer for each dichotomous item and the mean score for each polytomous item. The percentage of students scoring at each score point was also reported for each polytomous item.

Rater Location

The location of each panelist's "cutpoint" relative to the "cutpoints" of the other panelists in his/her grade group. The rater location data was provided graphically.

Whole Booklet

The *whole booklet feedback* reported to the panelists the expected score (on the test booklet form that they took) for students performing at the borderline of each achievement level based on the cutpoints they set. This feedback was provided graphically. The *whole booklet exercise* was an extension of this feedback. Panelists were shown copies of booklets with scores around each achievement level cutpoint. They were then asked to examine the responses in those booklets and determine if that was what they expected from students performing at each achievement level. Although the feedback was updated after each round, panelists were only shown student booklets after the first round of ratings.

All feedback provided to the panelists during the process is in Appendix H.

Each piece of feedback information was presented to the panelists and they were instructed in the meaning of the feedback and how to use it in subsequent rounds of ratings. Time was provided for panelists to ask questions and discuss the feedback before beginning another round of ratings. Feedback from the previous round was updated after each round of ratings and presented to panelists along with new information. Following round 3, all feedback was again updated relative to the ratings for this, the final round.

Selection of Exemplar Items

On the final day of the ALS process, panelists reviewed items from the CP and TB blocks scheduled for release to the public. These items were those from which the panelists could select items to serve as illustrations of the achievement levels for each grade.[7]

Items presented to the panelists had already met the statistical criteria. First, an item for consideration as an exemplar for an achievement level must have an average conditional p-value of at least 50% across that level. Second, the difference between its average conditional p-value at that level and its average conditional p-value at the next lower level should be in the 60th percentile.

Two lists were presented to the panelists for each achievement level. The primary list contained items that satisfied both criteria, and the secondary list contained

[7] Four blocks of items were released for each grade level: two CP, one TB and one HO. Panelists did not select exemplar items from HO blocks.

items that satisfied only the first criterion. Panelists were instructed to consider items from the secondary list only if they rejected all the items in the primary list.

Results

Results of the ALS process are presented here, and they are organized on the basis of the three outcomes: (1) achievement levels descriptions; (2) cutpoints; and (3) exemplar items. Results of process evaluations and consequences data questionnaires are presented in the next section.

Achievement Levels Descriptions

Training in ALDs began by having panelists examine the alignment of the preliminary descriptions of Basic, Proficient, and Advanced achievement for their grade level.[8] The alignment charts resulting from this exercise are in Figures 3-5.

In addition to a thorough study of the preliminary descriptions, panelists were given opportunities to revisit the descriptions and make modifications before each round of ratings. After each session of review and proposed modification, the content staff met to discuss the suggestions and to determine jointly whether the suggestions seemed in keeping with the Framework, and whether the changes represented significant substantive changes. All changes for all grades were reviewed by all content staff. If the content staff agreed that the changes were consistent with the Framework, the changes were incorporated into the achievement levels descriptions and then taken back to the panelists for final agreement before each round of ratings. Panelists were advised that once the third (final) round of item ratings had begun, no further changes in achievement levels descriptions could be made. *Relatively* few changes were recommended.

The achievement levels descriptions developed by panelists during the ALS meetings in Phoenix were edited for form and uniformity across grades.[9] The descriptions of the three achievement levels at each of the three grades are in Figure 6. The final versions of the borderline descriptors used in the item rating process are in Figure 7. The different versions of the ALDs and borderline

[8] In July 1996, a panel was convened including seven persons who had been members of the NAEP Science Framework Panel, and/or Science NAEP item writers and reviewers, and/or are members of the continuing review panel for the Science NAEP. This two-day meeting was for the purpose of reviewing the preliminary achievement levels descriptions and determining whether they were appropriate and adequate. In order to make this determination, it was necessary to review the context in which the preliminary descriptions were to be used and their role in the ALS process. The members of this panel concluded that the preliminary descriptions could be used without modifications prior to the ALS process. This exercise to study the alignment of the preliminary descriptions was suggested by the panel, and the details worked out by the ACT Project Staff.

[9] The editing was performed by the Publications Department at ACT.

descriptors resulting from different stages of the process and the preliminary ALDs are included in Appendix I for comparisons.

Panelists recommended that the ALDs be made more "user friendly" in format. They recommended that some highlighted or "bulleted" text be extracted from the descriptions to make key aspects clear to readers.

Cutpoints

Overall Cutpoints

The cutpoints in Table 4 are on the ACT NAEP-like scale. The standard error for each cutpoint is computed as half the difference between the cutpoints set by the two rating groups. The *cutpoints* reported to the panelists during the process are in Table 4. There were no substantial changes in the cutpoints from one round to another, except for grade 8 Advanced. This is not meant to say that individual item ratings were not changed substantially from round to round; only that the average across all items and raters did not change substantially.

Table 5 shows the averages and standard deviations of the rater locations. The *standard deviations* presented to the panelists during the process are those found in Table 5. The averages in Table 5 are only slightly different from the cutpoints in Table 4.

Only the cutpoints from the third round of ratings are recommended to NAGB, and only round 3 results will be presented here.

The percentages of students scoring at or above each achievement level cutpoint are presented in Tables 6-8. The performance distributions relative to the cutpoints by demographic variables of the students are also included. The distribution information found in Table 6 was based on the plausible values distributions provided by the Educational Testing Service (ETS).

Using the test characteristic curve (TCC) for each grade level item pool, the expected proportion of score points (percent correct score) of students performing at each achievement level cutpoint was estimated. The data are presented in Table 9. The expected percent correct score for students scoring at the mean of the distribution for each grade level was also included.

Analyses of Effects Associated with Panelists

Since considerable emphasis was placed on the selection of panelists and their assignment to rating groups, analyses of panelist effects on the cutpoints were conducted. Tables 10-17 contain data resulting from these analyses.

Rating Groups. The cutpoints set by rating groups are presented in Table 10. For each grade level, within each achievement level, none of the cutpoints set by groups A and B were significantly different. In grade 4, group A's cutpoints were

consistently higher than those of group B. In grade 12, group B's cutpoints were consistently higher than those of group A. There was no consistent pattern in grade 8.

Common Blocks. Since the item rating pools were not exactly equivalent, it was possible that the consistent patterns of differences in ratings by rating groups in grades 4 and 12 were due to the differences in item rating pools. That is, it is possible that ratings by panelists in grade 12 group A produced lower cutpoints because the items in their rating pool were relatively harder. Thus, cutpoints set by the rating groups were computed using only the blocks that were common to the item rating pools. If the cutpoints for common blocks differed, this would indicate rater differences because they were based on the same items. The results are in Table 11.

In grade 4, group A did set consistently higher cutpoints than group B on the items that both groups rated. In grade 12, group B did set higher cutpoints than group A at the Basic and Proficient levels on the common items. The Advanced cutpoints were virtually the same for both rating groups at grade 12. None of the cutpoints in Table 11 were significantly different at the 0.05 level.

Cross-Grade Blocks. Some blocks of items in the assessment were administered to more than one grade level. There were two CP, one TB, and two HO blocks that were common to grades 4 and 8, and similarly for grades 8 and 12. On the blocks common to grade 4 and 8, grade 8 set consistently higher cutpoints than grade 4. On the blocks common to grade 8 and 12, grade 8 set consistently higher cutpoints than grade 12. (Please see Table 12.) Since the scaling of the items was *within* grade, these results do not necessarily indicate that grade 8 panelists set higher achievement levels than grade 12 panelists. Thus, instead of comparing the cutpoints set on the cross grade blocks, the ratings provided by the grade level panels to the items in those blocks were compared. (Please see Table 13).

The higher average ratings provided by the grade 8 panel to the items rated by both grades 4 and 8 panels indicate that students performing at an achievement level, e.g., Basic, in grade 8 were expected to perform better on those items than students performing at the Basic level in grade 4. That is the logical expectation. Similarly for the blocks common to grades 8 and 12, ratings indicated that students at grade 12 were expected to perform better than students at grade 8. Notice that the differences in average ratings between grades are all statistically significant at the 0.05 level.

Demographic Characteristics of Panelists. The cutpoints set by panelists from different regions of the country were also compared. (Please see Table 14.) The only statistically significant difference was found in grade 4 at the Advanced level where cutpoints set by panelists from the central region were higher than the cutpoints set by panelists from the west. There was no general pattern observed.

Tables 15-17 show the results of the comparisons of average cutpoints set by different subgroups of panelists with respect to panelist type, sex, and ethnicity. There were no statistically significant differences found.

Analyses of Effects Associated with Items

Hands-On Tasks. Cutpoints set on the different types of item blocks were compared. Table 18 shows that grade 4 panelists set higher cutpoints on the HO blocks than they did on the rest of the items. Grade 12 panelists, however, set lower cutpoints on the HO blocks than they did on the rest of the items, and the difference between the cutpoints at the Proficient level is statistically significant. There was no pattern in grade 8. The same results were found in Pilot Study 2.

Block Types. Cutpoints set using items in the three different types of blocks were also computed. Pairwise comparisons were performed, and some statistical differences were found. However, there were no striking patterns to warrant further analyses. Please see Table 19.

Item Type. In Table 20, the cutpoints from different item types (dichotomous and polytomous) are presented. Notice that polytomous cutpoints were consistently higher than dichotomous cutpoints. Comparing the cutpoints in Table 20 to those in Table 4 indicated that combined cutpoints were dominated by the polytomous item ratings. Information weighting was used to combine dichotomous and polytomous item ratings. This generally means that polytomous items "weighed" more than dichotomous items. That fact, plus the relatively large number of constructed response items, means that the overall cutpoints for science were largely determined by polytomous items.[10]

Content Area. Table 21 reports the cutpoints set for each field of science, and student performance relative to those cutpoints are also reported. The panelists at grade 4 varied the fields of science for which they set the cutscores highest or lowest. For grade 8, Life Science was consistently the field of science for which they set the highest cutscores. Life Science was also the field of science for which both the Proficient and Advanced cutscores for grade 12 were set highest.

Most 7th and 8th graders study general science. Of the three fields of science covered in the Framework, Life Science was the most frequently included in the curriculum. Further, of the students in high school who study science, most study biology. Perhaps this was why panelists had higher expectations of student performance for items assessing the Life Sciences.

Content Area and Panelists' Expertise or Special Interest. It was hypothesized that panelists would set higher cutpoints in the content area in which they had expertise or special interests. Results of the analyses to test this

[10] A study comparing NAEP student performance on different item formats relative to the achievement levels is included in Volume IV of this report.

hypothesis are in Tables 22-24. At grade 4, ratings by Physical Science panelists (versus all others) were significantly higher at the Advanced level, and ratings by Earth Science panelists (versus all others) were significantly higher at the Basic level. In grade 8, Earth Science panelists set significantly lower cutpoints at the Proficient and Advanced level. The grade 8 Life Science panelists consistently set higher cutpoints in Life Science compared to other content areas, but the differences in the cutpoints were not statistically significant. Grade 12 Earth Science panelists set significantly higher cutpoints at the Basic level. However, Earth Science panelists set consistently lower cutpoints in Earth Science than in other content areas, although the differences were not significant.

Exemplar Items

Figures 8-10 present the lists of items in the released blocks that were presented to the panelists for their review and recommendations. Figures 11-13 are the lists of items that panelists recommended as exemplars. Tables 25-27 provide a count of items by type: the entire item pool, those presented to panelists, and those recommended by panelists. These tables also provide statistical information about the items that passed the statistical criteria. The items meeting the difficulty criterion (average percent correct across the level ≥50%) were rank ordered by their discrimination index (the difference in the average percent correct across the level in question and that for the next lower level). Items in the 60th percentile (considering all the items, not just the released items) were included in the primary list.

Exemplar items recommended by panelists and the scoring rubrics are in Appendix J.

Other Results

ALD Location Exercise

Each panelist was presented with two item maps or charts. ALL items for a grade level were included on each chart. One chart used a 50% p-value to map the items, and the other used a 65% p-value. Half of the panelists had the 50% p-value map as Chart 1, and the other half had the 50% p-value map as Chart 2. Each panelist had pens to mark the cutscores in different colors, and item lists with the items ordered for each mapping (see Appendix H for sample materials). The panelists were not told the p-value used to map the items, just that these charts represented two ways of locating the items on the score scale with respect to the cutscores.

This exercise was implemented in the general session. Panelists' materials were at their place, and they were seated in alphabetical order by grade group. Panelists' materials were arranged so panelists sitting side-by-side had different charts.[11]

[11] Chart 1 for panelist 1 used a 50% p-value to make the items, whereas a 65% p-value was used to map items on Chart 1 for panelist 2, and so forth. Every-other panelist had the

Panelists were instructed to examine the charts and item lists. They were given the cutscores from round 3 and told how to mark those on the charts and item lists. They were instructed to evaluate the correspondence of items with descriptions. They were told that all items to the left of the lower borderline of each achievement level were "can do" items and that those within the cutscore boundaries of each level were "challenging" for students at that level. It was explained to them that "can do" and "challenging" items change from one mapping (chart) to the other. They were instructed to compare those items in each chart to the ALDs to identify the chart that would best represent student performance to the public for reporting.

Panelists were given about 10 minutes to examine the charts, during which time they were allowed to discuss the items and charts with table mates. At the end of that time, they were told that discussion with table mates must end. A questionnaire was distributed. The questionnaire asked which mapping best corresponded to the ALDs, in an overall sense.

A summary of the panelists' responses to the questionnaire is in Table 28. About half of the panelists for each grade level indicated that the 50% mapping best corresponded to the ALDs.

Consequences Questionnaire Data

After the panelists had completed the final round of ratings and selected exemplar items, they were shown consequences data. (Please see Figures 14-16.) These data were the percentages of students scoring at or above each cutscore for their grade level. All panelists were shown these data together in the general session. Thus, panelists at each grade level saw the consequences data for all three grade levels. Panelists were then asked to complete a questionnaire regarding the consequences data.

The tallied responses from the Consequences Questionnaires for the ALS are in Appendix K. A summary sheet is also included for each panelist who recommended that at least one achievement level be changed. That summary sheet reports the panelists' responses to the Consequences Questionnaire (including comments), the rater location (that rater's cutscore), the grade level cutscore and student distribution data.

ACT provided consequences feedback in both U.S. history and geography, as well as for both pilot studies for science. We typically found that about half of the panelists indicated that the results were not what they expected, but few indicated that they would recommend any changes in the levels presented to NAGB. For the science ALS process, however, nearly half of the grade 8 panelists indicated that they would make changes in one or more cutscores in order to change the consequences

---

same chart.

data. With few exceptions, the recommendations were to lower the cutscore in order to increase the percentage of students scoring at or above the level.

ACT analyzed the relationship between the location of each individual panelist's cutscore relative to the grade level cutscore and their recommendations regarding raising or lowering cutscores. In general, those panelists who recommended changes were those who were within 1 standard deviation (SD) of the grade-level cutscore. Please Table 29. The results of this investigation indicated that the "extreme" raters were generally satisfied with the consequences of their ratings.

At grade 4, two of the three people who recommended lowering the Proficient cutscore were within 1 SD of the Proficient cutscore. One person who recommended the grade 4 Advanced cutscore be lowered was within 1 SD and one was +1 SD (but less than 2). At grade 8, three of the five panelists who recommended that the Basic cutscore be lowered were within 1 SD, one was -1 SD and one was -2 SDs. Of the 16 who recommended the Proficient cutscore be lowered, 13 were within 1 SD, two were -1 SD, and one was -2 SDs. Of the 17 who recommended the Advanced level be lowered, 14 were within 1 SD, and the counts for -1 SD and -2 SDs were the same as the Proficient level. For grade 12, two people recommended that the Basic cutscore be lowered, and they were both within 1 SD of the grade-level cutscore. Four of the five who recommended that the Proficient cutscore be lowered were within 1 SD of the Proficient cutscore, and seven of the eight who recommended the Advanced cutscore be lowered were within 1 SD of the Advanced cutscore.

Process Evaluation Data

Panelists completed seven process evaluations; there was usually one at the end of each day. The first process evaluation was completed at the end of Day 2, the first full day of the ALS process. At that time, panelists had completed taking the NAEP exam, working with all the hands-on tasks, and working with the preliminary descriptions to examine their "alignment" across the three levels. Panelists were generally positive with respect to their experiences, and grade 8 panelists were not an exception to this. Summaries of panelists' responses to all questions are included in Appendix L. Figures 17-22 and Tables 30-32 provide summaries of some key questions.

Data reported in Table 30 and Figure 17 show the average responses (1="most positive" and 5="most negative") with respect to their understanding of the achievement levels descriptions (ALDs). Panelists at grade 8 reported the highest level at the end of Day 2. Of the three grades, grade 8 responses averaged highest for levels of personal satisfaction with the achievement levels descriptions for their grade and their understanding of the descriptions.

At the end of Day 3, panelists completed a questionnaire again, and some of the questions were repeated. At that time, the level of understanding of the ALDs had increased for panelists at each of the grade levels and at each of the achievement

levels, except for grade 8 Basic. Their work with developing borderline descriptions had provided most panelists with a fairly well-formed concept of borderline performance. Responses of grade 8 panelists indicated that their conceptualization of borderline performance was "most" well-formed. (See Table 28.)

Following the paper selection process and following round 1 ratings, panelists completed process evaluations. Grade 8 panelists were most positive about the paper selection process and the purposes it served.

Panelists have always indicated a lower level of understanding in ALDs following round 1 ratings than had been indicated prior to the first round of ratings. (See Figure 17.) That pattern was found for the science ALS panelists, too. Again, grade 8 panelists were most positive regarding their level of understanding of the ALDs. As can be seen in Table 29 and Figure 18, they were also most positive in terms of their concept of borderline performance (how well formed).

Panelists in grade 8 were observed to spend very little time in review of the Whole Booklet Exercise material following round 1. Indeed, they spent little time reviewing the feedback from round 1. This was surprising, because panelists had always been quite eager to have feedback, and they had been especially interested in the Whole Booklet Exercise. In response to questions regarding the feedback, grade 8 panelists indicated little impact from the Whole Booklet Exercise, and they had the lowest average level expecting to use all the feedback data to adjust ratings in round 1. With respect to *individual* pieces of feedback, however, their average responses were highest in terms of plans to use feedback to adjust cutscores.

Of particular interest is the finding that grade 8 panelists had the highest average response to a question indicating that they planned to change their cutscore to be lower than the average for the grade level and the lowest average response to the question indicating that they planned to change their cutscore to be higher than the grade average.

Following round 2, all panelists were again rather clear on their understanding of the achievement levels descriptions and felt that their concept of borderline performance was rather well formed. The grade 8 panelists were still most confident about the ratings they had provided. They slipped below grade 4 panelists in terms of the level of understanding of the ALDs, but they still had the highest average with respect to their concept of borderline performance.

After completing the final round of ratings, all panelists in grade 8 were more than *Somewhat Confident* about the ratings they had provided in round 3. A total of 61% of grade 8 panelists responded that they were *Totally Confident* about their round 3 ratings. Average responses were very high to questions regarding the panelists' understanding of the ALDs and their concept of borderline performance.

Grade 8 panelists reported less discussion of the ratings during round 3, but their evaluation of the "helpfulness" of the discussions was most positive. They also indicated the *lowest* agreement with the statement that they had all the information they needed for round 3 ratings.

Responses in the final process evaluation indicated that panelists were generally satisfied with the process. The exception to this was their responses to the questions regarding the selection of exemplar items. Some panelists at grade 4 were very negative in their comments, and many panelists at grade 8 were dissatisfied with the items and the process.

Grade 8 panelists were generally a rather contentious group. They were almost never able to reach amicable agreement on any task. Many of them needed to express their opinion on every single comment made in the group. They rarely had time to complete the tasks they were given because they needed so much time to discuss the task that they were about to do. This was equally true of the exemplar item selection process.

When the actual "vote" was being taken, they were very insistent that a count of the vote be taken. If there was not a clear majority, they wanted an extensive debate of the decision to present the pros and cons of approving the item. While there was more time scheduled for selection of exemplar items than usual, the grade 8 panelists hardly had any time left for the selection of items at the Advanced level. The fact that not all panelists had rated all items was seen as a major problem for these panelists. That was, perhaps, due to the lack of trust for and confidence in the recommendations made by most other panelists.

In general, no particular problems were revealed during the ALS process. Grade 8 panelists were observed to be struggling *as a group*. However, their responses to the questionnaires indicated that they were confident and comfortable, *as individual panelists*.

Grade 8 Reconvention

ACT was rather alarmed by the grade 8 responses to the Consequences Data Questionnaire and their responses to the final process questionnaire. This information was presented to both TAT and TACSS. TACSS recommended that the grade 8 panelists be reconvened. The NAGB Achievement Levels Committee concurred with that recommendation, and the panel was reconvened on November 23-24, 1996.

Twenty-five of the 33 panelists participated in the Grade 8 Reconvention.[12] Most of the panelists who were unable to participate were representatives of the general

[12] Three additional panelists had planned to participate but family emergencies and similar unexpected events prevented their participation.

public, and only two representatives of the general public were present. Some comments on the results of the Reconvention were provided by those persons who could not attend and they were in agreement with the recommendations of their colleagues.

During the Reconvention, panelists were given the opportunity to make revisions in each of the three outcomes of the ALS process: descriptions, cutpoints, and exemplar items. Some modifications were made to each. The ALDs resulting from the Reconvention are in Figure 23, and the borderline descriptors are in Figure 24.

An item mapping procedure was used for evaluating the cutpoints. Panelists examined maps of items located along the ACT NAEP-like scale. Items were located at the ACT NAEP-like score for which the probability of correct response was 65% (with a correction for guessing). Panelists were asked to identify each cutpoint in such a way that items representing the domain of knowledge and skills included in the description of the achievement level would be associated with the cutpoints. If they were unable to identify a single score point, they were allowed to give a range of scores within which their cutpoint would be located.

Table 33 reports the means and standard deviations of the minima, the maxima, and the midpoints of the ranges. The minimum, maximum, and midpoint *were* the cutpoints given by panelists who identified a cutscore. The means of the midpoints were the achievement levels cutpoints. The percentages of students performing at or above each achievement level are in Table 34.

The new grade 8 cutpoints for different subgroups of panelists are presented in Table 35. Table 36 reports the means of the differences between individual cutpoints from the ALS and the reconvention (i.e., computed for each panelist then averaged). Table 37 presents information about the exemplar items. The list of items presented to the panelists and those that the panelists selected are in Figures 25 and 26. The items selected by the panelists are in Appendix M.

The percentage of students scoring at or above each achievement level based on the new cutpoints were presented to the panelists. (Please see Figure 27.) They were asked to fill out the consequences data questionnaire. Panelists' recommendations were included in the summary of their responses to the questionnaire. (Please see Appendix N.)

There were three evaluation questionnaires. Statistical summaries of panelists' responses are in Appendix O. Panelists responses to selected questions were compared to responses from the ALS meeting in Table 38. Results indicate that panelists were more satisfied during the reconvention.

A report on the grade 8 Reconvention was prepared for TACSS review. TACSS recommended that these be used instead of the ALS cutpoints developed in Phoenix.

Results of the Reconvention were recommended to NAGB as the achievement levels for grade 8 science. NAGB adopted those as the official cutpoints for grade 8.

Public Commentary

Approximately 1,000 packets of information were sent to members of the Association for the Education of Teachers in Science (domestic addresses only), to selected organizations on our Stakeholders list, and to state officials in selected positions who served as Science Panel Nominators.

ACT received responses from 95 persons. About a quarter of the responses were item reviews and failed to address the questions for which we asked for commentary. Of those responses that did provide substantive comments about the three components of the achievement levels (descriptions, numerical cutscores reported as the percentage of total possible points, and three exemplar items for each level), the reviews were generally favorable. Most of these respondents did note that the information was too limited to be of much use in making an evaluation of the correspondence between performance and the descriptions, and they also noted that the percentage of total possible points is of little use for such an evaluation with no more supporting information. With those reservations noted, here are tabulations of their comments. The coding categories are very general, but they provide a sense of the general support or opposition to the preliminary results presented in the review material. A sample packet of material sent for public comment is included in Appendix P.

Numerical Cutscores Represented by Percentage of Total Points Scored

Three people expressed some sort of general concern about the numerical data representing the achievement levels. Twenty-six people provided "generally positive" comments, and there were comments that specifically mentioned grade levels. Eighteen people addressed comments to levels that appeared to be too low. Three addressed comments to levels that appeared to be too high.

Achievement Levels Descriptions

Forty persons commented on the descriptions in a generally positive manner. Seven comments addressed the descriptions of specific grade(s) as being good. One positive comment was specifically on the language of the descriptions. Ten people commented on the descriptions in a generally negative way. Most of these were comments about how the NAEP descriptions did not match other standards. Seven comments addressed negative aspects of descriptions at particular grades and levels of achievement. The grade 4 descriptors had the most negative comments. Many of these were related to the reference to "mass" and "density." One comment addressed negative aspects of the language used in the descriptions, in general.

Exemplar Items

Many people had very positive comments to make about the fact that the assessment included so many open-ended items. There were 18 general comments about positive aspects of the items. In addition, there were 8 comments about the items at specific levels and grades.

Several people had negative comments to make about the exemplar items, however. Twelve people provided generally negative comments, and 11 comments addressed specific grades and levels of achievement. The negative comments tended to be about wording, bias, scoring, and that sort of thing. ACT learned, for example, that mayonnaise is a very controversial subject, and that references to potato salad introduce cultural bias.

Fifteen people had negative comments to make about the scoring rubrics and their applications.

In general, the criticisms of the lack of adequate information for comment were warranted.[13] Nonetheless, the general tone of the comments about the achievement levels descriptions was positive, the numerical cutscores were generally thought to be about right or too low (in terms of the percent correct associated with each level), and the items included as exemplars were valued as good examples of what students should know and be able to do.

[13] Data on actual cutscores and the distribution of student performance relative to those cutscores could not be released. Thus, there was little for which reviewers could provide comments.

References

ACT (1994). *Design document for setting achievement levels on the 1994 National Assessment of Educational Progress in Geography and in U.S. History and the 1996 National Assessment of Educational Progress in Science.* Iowa City: Author.

Bay, L. Chen, W.H., & Loomis, S.C. (1997a). Comparing Student Performance on Different Item Formats Relative to Achievement Levels Cutpoints. In ACT, *Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science Final Report* (Vol. IV). Iowa City, IA: ACT.

Bay, L. Chen, W.H., & Loomis, S.C. (1997b). Observations About the 1996 NAEP Science Assessment. In ACT, *Setting Achievement Levels on the 1996 National Assessment of Educational Progress in Science Final Report* (Vol. IV). Iowa City, IA: ACT.

Council for Chief State School Officers (n.d.). *Science Framework for the 1996 National Assessment of Educational Progress*. Washington, D.C.: National Assessment Governing Board.