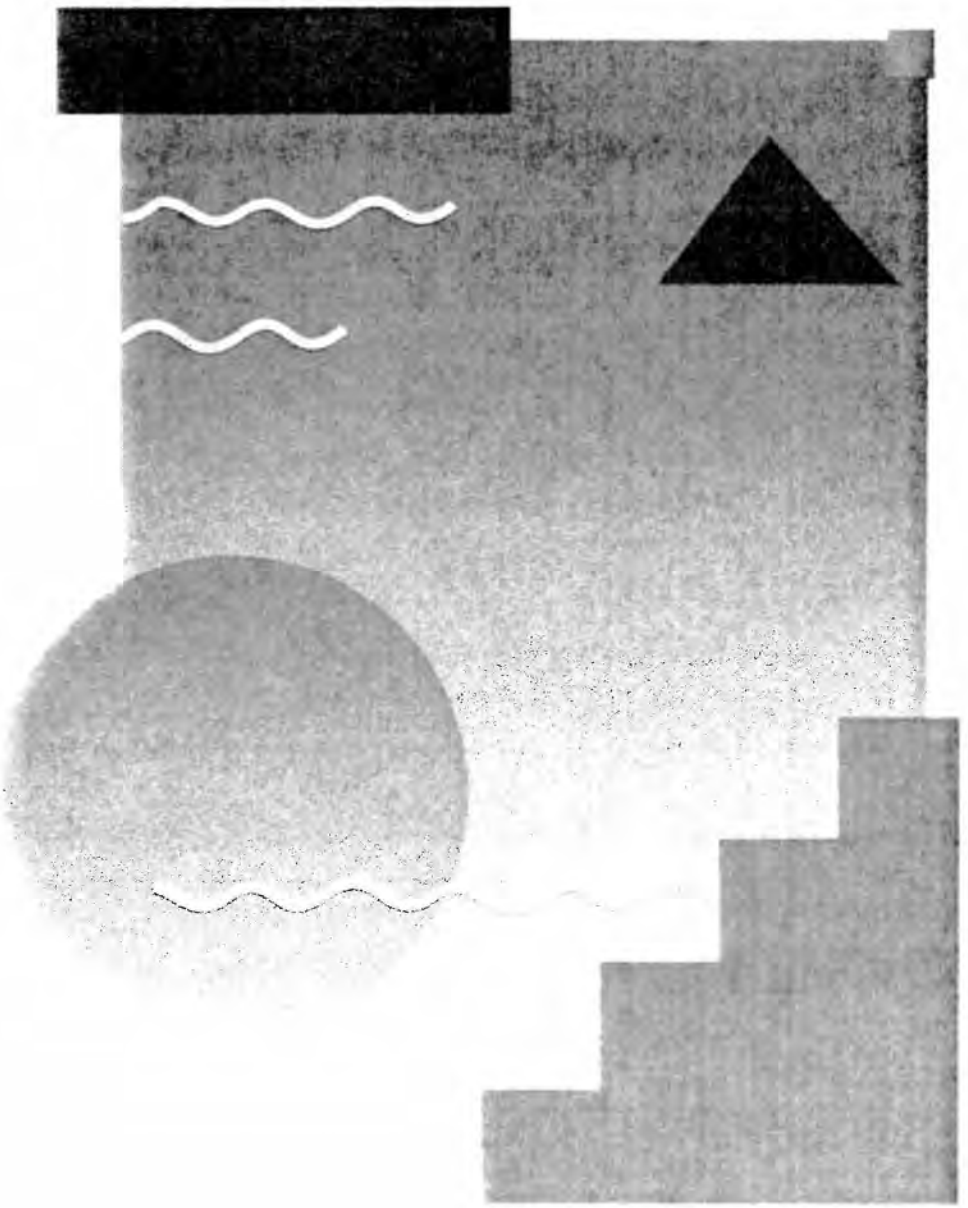
ACT

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

Final Report

Volume II
Pilot Study 2

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

List of Tables

List of Figures

Introduction

A second pilot study of the 1996 NAEP Science Achievement Levels-Setting (ALS) process was conducted August 18-22, 1996, to test procedures planned for the ALS panel to be convened in September, 1996, and to determine whether any additional modifications or adjustments were necessary.

For the most part, procedures planned for the science ALS were those used in the 1994 NAEP Geography and U.S. History ALS processes. Some adjustments were made to accommodate the inclusion of hands-on tasks in the science assessment.

A pilot study to focus on the inclusion of hands-on tasks in the science assessment was implemented as a first step in the ALS process for science. Findings from that study were incorporated in the design and implementation of this pilot study. A full report on the Pilot Study 1 (PS1) of the 1996 NAEP Science ALS is included in Volume I.

Three new procedures were tested in Pilot Study 2 (PS2) for the first time. The inclusion of these procedures in the actual ALS meeting was dependent on their success in this pilot study. The three procedures are as follows.

1. As a first step in determining whether modifications to the preliminary achievement levels descriptions were necessary, panelists used charts to determine the alignment of descriptors across the three achievement levels.

2. Panelists participated in a "preview" round of ratings prior to the training in the rating process.

3. The Achievement Level Description (ALD) Location Exercise was implemented after the third round of ratings. The purpose of the exercise was to ascertain the opinions of panelists regarding the location of the cutpoints with respect to the ALDs, and with respect to the items located within the ranges of the cutscores defining the levels.

Achievement Levels Descriptions Alignment Charts

A special meeting was held in July to discuss the Framework and the preliminary achievement levels descriptions. That meeting was attended by the five content persons who worked with the process, plus one additional person who was on the Framework panel. In addition, input from another member of the Framework panel was communicated in writing and telephone discussions.

That panel was asked to determine whether the preliminary achievement levels descriptions could be used as the starting point for the second pilot study and the ALS process. Concern had been raised by the results of the first pilot study conducted in March. In particular, the very low cutscore for grade 4 Basic, relative

to those for the other levels, suggested that the description of grade 4 Basic was too low.

This panel spent most of two days discussing the process to be implemented and the question of whether the preliminary descriptions were appropriately calibrated to serve as a starting point for the process. The clear answer from those content persons was that the preliminary descriptions were properly calibrated. That is, they were a reasonable interpretation of the Framework, in terms of the policy definitions. So, the sorts of things that students should know and be able to do in science, given the Framework, were properly aligned with NAGB's policy statements of what students at the three levels of achievement should know and be able to do.

The content group acknowledged that the preliminary descriptions were in need of modifications, however. The content staff from PS1 had recommended that the descriptions developed by panelists be taken as the starting point for panelists in PS2. TACSS had rejected that recommendation, however.

The content panel felt that some sort of matrix, similar to those developed before, during, and after PS1 would be helpful for panelists to use in incorporating the preliminary descriptions and in assessing the need for modifications.

The process facilitators discussed this and the ALD Alignment Charts were agreed upon. ACT's publications department worked with project staff to design charts with movable components of the descriptions.

Each sentence of each description was printed on a separate "card." Each card was color coded for an achievement level and the sentence sequence was identified. The descriptors had a small velcro tab to be used to attach it to the chart. The goal was for panelists to settle on an alignment criteria or scheme and to align descriptors that represented the same dimension at different levels of achievement. If there were no mention of a particular knowledge/skill dimension at a level, the space was left blank to note that.

Working in table groups of approximately five panelists each, a chart was developed. Each table group worked together for about one hour to complete the chart. They then presented their charts to the grade group.

The grade group as a whole reached agreement on an alignment chart to serve as their working chart for future sessions in internalizing ALDs and modifying them.

Panelists were instructed that some "missing pieces" or empty cells might be appropriate. This was one aspect of the alignment chart to study.

Panelists generally seemed to gain understanding of and to form a good working knowledge of the ALDs as a result of this exercise. Its success led ACT to recommend that this be implemented in the ALS process.

Preview Round of Rating

During the debriefing session with PS1 panelists, they suggested that an extra "first round of ratings" be added, and that it be very early in the training process. Training for days before doing the "real task" created frustrations and anxieties in the beginning of the five-day process. After discussions with TACSS, a "preview round" of rating was added to the agenda for the second pilot. This preview rating session was scheduled before panelists were trained to be raters. A copy of the agenda is in Appendix A.

The purpose of the preview rating session was to give panelists a more concrete notion of the item rating procedure for which they were training. Panelists were given the opportunity to rate the items included in their NAEP exam (i.e., the test form they were administered on the first day). This session gave the panelists firsthand experience in the process for which they were training, and it was deemed to make panelists more aware of what they needed to know.

After the panelists had been given an orientation to the NAEP Science Frameworks and Achievement Levels Descriptions (ALDs), they began working to develop familiarity with and understanding of the ALDs using the ALD charts. They were introduced to the concept of borderline performance, and were given a brief description of the item ratings. For the preview rating session, they were instructed to use their understanding of the preliminary ALDs, their concept of borderline performance, and their judgment of test items. No results or feedback were provided from their ratings.

During a follow-up general session, the purpose of the preview round of ratings was reiterated. They were reminded that they were not yet trained for the rating process and that this was to have made them aware of the aspects in the training with which they were still less confident.

ALD Location Exercise

This exercise involved panelists' evaluation of the match or correspondence between the ALDs and the items within the range of scores denoting each achievement level on the ACT NAEP-like scale. The purpose of the exercise was to have panelists evaluate the ALDs with respect to the cutscores they had set. This exercise was implemented after all steps in the ALS process, per se, were completed. This meant that the final round of ratings had been completed and the exemplar items had been selected prior to this exercise.

Each item in each grade level item pool was mapped to the ACT NAEP-like scale using two different mapping criteria. The first mapping used a 50% probability criterion. That is, using the item characteristic curve (ICC), each dichotomously scored item was located on the scale such that the probability of a correct response by a student performing at that level was 50%. The second mapping used a 75% probability criterion. Each polytomously scored item was dichotomized at each partial or full credit score level. The ICC for each of those levels was used to map the "item" on the scale using either the 50% or the 75% probability criterion. Thus, each polytomously scored item with *n* score levels was mapped on the scale *n-1* times using each probability.

The item maps using the 50% probability were presented to panelists in group A in each grade level, and the item maps using the 75% probability were presented to group B. The group A panelists from the three grade levels were trained in the exercise together, and the group B panelists in each grade level were trained together. The training was in the general session. The script used in training the panelists in this exercise is in Appendix B. Panelists were instructed not to discuss the ALD location feedback with panelists from the other group. Panelists performed the exercise in the grade groups.

To evaluate the items mapped within the performance domain of each achievement level, panelists were asked to consider two conceptualizations of improving the correspondence of the ALDs and the items. The first conceptualization involved moving the cutpoints but keeping the item locations. The second conceptualization involved moving the items on the scale and leaving the cutpoints where they were based on round 3 ratings. Panelists were asked to consider each conceptualization separately, for each achievement level, and overall. They were, of course, made aware that in reality, moving either the items relative to a cutpoint for an achievement level or moving the cutpoint relative to the item locations, would affect more than a single level. Panelists were asked to complete a questionnaire regarding the ALD Location Exercise. A copy of the questionnaire is also in Appendix B.

Sampling Error

The intention was to implement the selection process described in the *Design Document*. An error in the process of sampling school districts was discovered after the samples were drawn. Please see Appendix C for documentation on the impact of the error. An error occurred when selecting records from the "master" Market Data Retrieval (MDR) data file. It was necessary to create a file with only district records before sampling districts according to the stratification criteria in the sampling design. The error occurred in specifying the logical record length and the output file. The result was a file of only about 3,000 records rather than nearly 15,000. (Please see Table 1.)

The proportion of districts that were included in the sample for each state was fairly close to the number of districts per state in the total sample. The problem in representativeness appeared with respect to the proportion of districts having 25% or more of their population living in poverty and in the proportion of districts having enrollments of 50,000 or more students. The effect on enrollment representativeness was greater. (Please see Table 2.)

The investigation was prompted by concerns regarding the samples for Pilot Study 2. Notification letters had already been sent to all nominators in the Pilot Study 2 sample, and the nomination packets and letters were ready to be mailed. We drew 30 more districts to increase representation from the southeastern states and to increase the districts with larger enrollments.

This problem in no way affected the ALS samples of nominator districts.

In summary, the samples for the two pilot studies were drawn from a subset of only about 3,000 school districts. The sample of districts in Pilot Study 2 was augmented to increase the number of districts in the southeastern region and to increase the number of districts with large enrollments.

One hundred ninety-six districts were included in PS2: 24% were from the central region, 21.4% from the northeast, 21.9% from the southeast, and 32.7% from the west. (Please see Table 3.) All of the samples were drawn without replacement **except** for the districts with enrollments of 50,000 or more. Since there were relatively few of those districts, and since we had to draw seven samples for the 1994-96 NAEP ALS processes (4 pilot studies and 3 ALS meetings), we were short of districts in the largest enrollment category. We judged that it would be acceptable to have replication of the largest enrollment districts, provided that the replication did not occur for the same category of panelists. That means that a district in the largest enrollment category drawn in the "teacher nominator" sample for U.S. history, for example, could be drawn in the "general public" nominator sample for science.

Panelists

A total of 313 persons were nominated to be panelists for Pilot Study 2. (Please see Table 4.) Grade 12 had the largest number of panelists nominated and grade 8 the smallest. Thirty-nine percent of the nominees were representatives of the general public. That was a significant improvement from PS1, for which only 9% of the nominees were representatives of the general public. This increase was a result of adding a new set of nominators for general public panelists. The new nominators were employers. For each school district drawn in the "general public" sample, ACT identified a company likely to employ persons with training in science. The CEO, director of personnel, or other such persons were identified as a nominator.

In addition, ACT identified several industries in the Standard and Poor's Index that were likely to require science training. A rather diverse set of industries was included. Two companies were identified for each state having a school district included in the general public sample. A list of titles was used as the sample frame, and persons in those positions were asked to serve as nominators.

More persons were nominated from the southeast region than any other region and fewer from the northeast region. This is a rather common pattern. The percentages of male and female nominees were about equal overall, but the numbers differed substantially at each grade level. There were more male nominees for each of grades 8 and 12, and there were more female nominees for grade 4. Again, this was a familiar pattern. About 20% of the nominees were minorities. Figure 1 indicates the content area of expertise or interest of the nominees as reported by the nominators. There was an indication of increasing specialization across grade levels, as one would expect.

The PS2 panel was composed of 54% teachers, 13% nonteacher educators, and 30% general public. (Please see Table 5.) The target distribution was 55%, 15%, and 30%, respectively. There was an overrepresentation of the central region, and the northeast was underrepresented. Minorities were somewhat overrepresented. Although 54% of the panelists were female, females were underrepresented in grade 12.

For the first pilot study most grade 12 panelists reported physical science as their area of specialization. The conjecture was that the lack of grade 12 panelists from other content areas was due to the guidelines for teacher nominees. That is, in the letter to nominators and in the guidelines for nominations, it was specified that teachers must "teach 12th grade science," that stipulation would exclude teachers of subjects other than physics. This was remedied by requesting nominations of teachers that "teach high school science." Thus, for PS2, there was a more diverse grade 12 panel relative to content area specialization. Please see Figure 2 for the distribution of the panelists by content area specialty or special interest.

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 of expertise. Each group consisted of two table groups that were also as equivalent as possible with respect to the above variables.

The item pool for each grade level was divided into two parts so that group A panelists rated one set of items and group B rated the other set of items. Included in each set were items that both groups of panelists rated. This design feature provided the opportunity to examine ratings from each group as a replication of the ratings from the other half.[1]

[1] Details about the item rating pools are discussed in the *Item Pools* section of Volume III of this report.

Results

Achievement Levels Descriptions

Advance materials were sent to the panelists to help them learn about the Framework, the policy descriptions of the three achievement levels, and the preliminary science ALDs prior to arriving for the pilot study. Nonetheless, the first part of the process included sessions for training panelists in the Frameworks and preliminary ALDs. Panelists were engaged in activities focusing on the ALDs to determine whether modifications were needed. If modifications were to be made, panelists had to reach agreement on changes. They were also asked to develop borderline descriptions.

For the first time in ACT's experience in setting achievement levels for the NAEP, training in the ALDs began by having panelists examine the alignment of the preliminary descriptions of Basic, Proficient, and Advanced achievement levels for their respective grade levels. This exercise to study the alignment of the preliminary descriptions was described earlier in this report. The alignment charts resulting from this exercise are in Figures 3-5.

Panelists were given opportunities to revisit the descriptions and make modifications before each round of ratings. Modifications were reviewed by content staff. Panelists were advised that their last chance to modify the ALDs was prior to the final round of ratings. 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.

Cutpoints

Overall Cutpoints

In Table 6 are the overall cutpoints and the standard deviations of the rater locations resulting from each round of ratings. All cutpoints are on the ACT NAEP-like scale. It has been ACT's experience that overall cutpoints do not change very much across rounds of ratings, and that the standard deviations decreased from round 1 to round 2 to round 3. In PS2 there were no substantial changes in the cutpoints across rounds. However, the standard deviations of rater locations *increased* from round 2 to round 3 for grade 4 for all achievement levels.

Although the cutpoints set by panelists did not change much from round to round, Figures 8-10 indicate that panelists did change their ratings. The directions and magnitudes of the changes balanced out, however, so that the cutpoints did not seem to change.

The cutpoints and standard deviations were presented to the panelists graphically as feedback after each round of ratings. Those graphs are included in Appendix D.

The results related to the cutpoints presented here were from round 3 ratings.

Analyses of Effects Associated with Panelists
Because considerable emphasis was placed on the selection of panelists and their assignment to rating groups, analyses of "panelist effect" on the cutpoints were conducted. Tables 7-15 contain data resulting from these analyses.

Rating Groups. The cutpoints set by rating groups are presented in Table 7. For each grade level, within each achievement level, none of the cutpoints set by groups A and B were significantly different. In grades 4 and 8, group A's cutpoints were consistently higher than those of group B. There was no consistent pattern in grade 12.

Common Blocks. Since the item rating pools were not exactly equivalent, it was possible that the consistent patterns of difference in ratings by rating groups in grades 4 and 12 were due to the differences in item rating pools. That is, it was possible that ratings by panelists in grade 4 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 the common blocks differed this would indicate rater difference because the items were the same. The results are in Table 8.

There was no consistent pattern found in any of the grade levels. Moreover, none of the cutpoints from groups A and B were significantly different at the 0.05 level.

Table Groups. Panelists were assigned to table groups (approximately five panelists in each group) according to the same criteria used to form item rating groups. Since most of the activities and discussions were within table groups, differences in the cutpoints set were investigated. The only significant differences were found in grade 4 at the Proficient and Advanced levels. (Please see Table 9.) These differences were further investigated using the common blocks. (Please see Table 10.) In each case, the cutpoints set by table group 3 was significantly lower than that of table group 4. (Table groups 3 and 4 were both in rating group B.) At all three achievement levels, the standard deviations for group 3 was substantially higher than for any other group.

It was determined that a grade 4 panelist had mistaken her "rater location" secret code.[2] For example, she thought that she was rater "B." Rater B had set very high cutpoints relative to others in grade 4. The mistaken rater apparently decided to bring her ratings down. She lowered her ratings rather drastically for round 2. Seeing no real change, for round 3 she again lowered her ratings.

[2] The grade 4 process facilitator related the information regarding the raters "mistaken identity." Ratings by that panelist were examined to surmise the remainder of this account.

Cutpoints were recomputed for grade 4 excluding the "outlier" rater. Results are in Table 11. Cutpoints set by groups A and B were closer when the outlier was excluded.

Demographic Characteristics of Panelists. The cutpoints set by panelists from different regions of the country were compared. In grade 12, at least one significant difference was found at each of the Proficient and Advanced level. There were no other significant differences found.

For both grades 8 and 12, panelists from the west region consistently set higher cutpoints. Please see Table 12.

Comparisons of cutscores by panelists' ethnicity (minority vs. nonminority) are in Table 13. For grades 4 and 12, minority panelists set consistently lower cutscores. The differences were statistically significant at the 0.05 level at the Basic and Proficient levels for grade 12, and at the Advanced level for grade 4. There was no consistent pattern nor were there statistically significant differences for grade 8.

The results of the comparisons of average cutpoints set by different subgroups of panelists with respect to panelist type and sex are presented in Tables 14 and 15. No statistically significant differences were found.

Analyses of Effects Associated with Items

Hands-On Tasks. Cutpoints set on the different types of item blocks were computed. Table 16 shows that grade 4 panelists set higher cutpoints on the Hands-On (HO) blocks than they did on the other types of blocks. Grade 12 panelists, however, set lower cutpoints on the HO blocks than on the other blocks. There was no pattern in grade 8.

Block Types. Cutpoints set using items in the three different types of blocks were also computed. Please see Table 17. For grade 4, the lowest cutpoints were for Theme-Based (TB) blocks and the highest cutpoints were for HO blocks. For grade 8, the lowest cutpoints were for TB blocks. There was no consistent pattern nor were there statistically significant differences for grade 12.

Item Type. In Table 18, comparisons of cutpoints from different item types (dichotomous and polytomous) are presented. Notice that polytomous cutpoints were consistently higher than dichotomous cutpoints. For grade 12, the differences at the Proficient and Advanced levels were statistically significant. For grade 8, the polytomous cutpoints were lower than the dichotomous cutpoints at the Basic and Proficient levels. This was an unusual result, and it was further investigated by computing the cutpoints by item type, by rating group. (Please see Table 19.)

Grade 8 group B ratings for polytomous items resulted in a lower cutpoint than that for dichotomous items for each achievement level, and those for group A were the

same for Basic and Proficient. Grade 12 group A also set the Basic cutpoint lower for polytomous items than for dichotomous items.

Scatter plots (Figures 11-19) showed that for grade 8, about half of the panelists set higher dichotomous cutpoints at each of the Basic and Proficient levels.

Content Area. Table 20 has the cutpoints set for each field of science. 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 cutscores highest. Earth Science was the field of science for which cutpoints for grade 12 were set the lowest.

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 a special interest. Results of the analyses to test this hypothesis are in Tables 21-23. At grade 4, ratings by Life Science panelists (versus all others) were consistently higher, and the differences in cutpoints were statistically significant at the Basic and Advanced levels. In grade 8, Physical Science panelists set consistently higher cutpoints, and the difference was statistically significant at the Basic level. The grade 8 Earth Science panelists consistently set higher cutpoints, and the differences were statistically significant at the Basic and Proficient levels. Significantly lower cutpoints were set by Life Science panelists, and the differences were statistically significant at the Proficient and Advanced levels. In grade 12, Physical Science panelists set consistently lower cutpoints, and the difference was statistically significant at the Basic level. The grade 8 Earth Science panelists consistently set higher cutpoints, and the difference was statistically significant at the Basic level.

Exemplar Items

The lists of items in the released blocks that were presented to the panelists for their review and recommendations are in Figures 20-22. Figures 23-25 report the lists of items that panelists recommended as exemplars. Tables 24-26 provide a count of items by type: the entire item pool, those presented to panelists, and those recommended by panelists. They also provide statistical information about the items that passed the statistical criteria. The items passing 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. All other items having an average probability for correct response across the range ≥50% were included in the secondary list. Panelists were instructed to recommend items that would serve to illustrate performance of students at each level. If no or few items from the primary list were accepted, panelists could make recommendations from the secondary list.

Other Results

ALD Location Exercise

Each panelist was presented with an item map or chart. Charts used by group A panelists used a 50% probability correct (p-value) to map the items to the ACT NAEP-like scale, and charts used by group B panelists used a 75% p-value. All items for a grade level were included on each chart. Although instructions for this exercise were given in two general sessions, one for group A and one for group B panelists, the exercise was performed in grade group sessions. Each panelist had pens to mark the cutscores in different colors, and item lists with the items ordered for each mapping.

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.

Panelists were given time to examine the charts. They were instructed not to discuss their charts with other groups, but they were allowed to discuss with their table mates. At the end of that time period, they were told that discussion with table mates must end. Questionnaires were distributed. Summaries of panelists' responses to the questionnaires are presented in Tables 27-30.

The specific design of the ALD Location Exercise implemented in PS2 had been recommended by TACSS. The instructions to be used to train panelists were reviewed by TACSS prior to implementation in PS2. The questionnaire was also provided to TACSS for review and modification prior to PS2.

A central goal of the procedure was to collect data on "the" mapping criteria to use for locating items with respect to the score scale.

The ALD Location Exercise was far too complex and difficult. Most panelists failed to focus on the fact that this was "another perspective on their work" and a "research activity." Instead, they responded that they felt that the correspondence between the ALDs and the items, with respect to the cutscores, was pretty good; to change them "arbitrarily" would make little sense after the process they had used to set the cutscores.

Also, the grade 4 facilitator believed that the two conceptualizations that panelists were asked to consider were "as different as apples and oranges." He apparently did not limit the focus to the purpose of the exercise and discussed the more obvious implications of moving cutscores and changing item locations. He apparently explained that moving the cutscores would change the distribution of students

scoring at or above each achievement level. ACT suggested that the results for grade 4 be disregarded because of these departures from the design.

Panelists did not have very much time to do the exercise, but they took even less time than they could have taken. Several did comment that the exercise was good or helpful in that it provided them a look at what they had done.

The results showed that panelists in group B using the 75% probability mapping were more likely to make changes in cutscores or item locations than panelists in group A. The 50% criterion would cause relatively harder items to be mapped within the levels.

Consequences Questionnaire Data

After the panelists had completed the final round of ratings and selected exemplar items, they were shown consequences data. (Please see Figure 26-28.) These data were the percentages of students scoring at or above each cutscore for their grade level. The percentages were computed as estimates based on a normal distribution, and not based on actual distributions of student performance. Ratings for grade 8 resulted in a very low percentage of students scoring at or above the Advanced cutpoint.

The panelists were not shown these data together in general session. Thus, panelists at each grade level did not see the consequences data for the other two grade levels. Panelists were asked to complete a questionnaire regarding the consequences data. The tallied responses from the Consequences Questionnaires for the ALS are in Appendix E.

About one half of the panelists indicated that the results were not what they expected. Of those panelists, one third indicated that they would recommend one or more of the achievement levels set if they could. The recommendation was to lower the cutscore in order to increase the percentages of students scoring at or above the Proficient and Advanced levels, which is not the case at the Basic level.

Process Evaluation

Seven evaluation questionnaires were completed by panelists, one at the end of each day or major procedure. These questionnaires were the same as the ones used in 1994, except for questions regarding ratings of HO blocks that were added. Detailed results of the analyses are included in Appendix F.

Some results of the process evaluations are included in Figures 29-34. Panelists were asked to indicate the clarity of their understanding of the ALDs at each level. At each achievement level, they reported that their understanding of the ALDs was clearer prior to round 1. Their level of understanding increased across rounds after an initial decline in round 1.

Panelists were also asked to indicate how well formed their concept of borderline student performance was during each round of rating. Figure 30 shows that their conceptualization of borderline student performance became more well formed across rounds.

Regarding the rating methods that they used in setting the cutpoints, the clarity and ease of applications of the methods increased across rounds. The mean estimation method, however, was always less clear and less easy to apply than the modified Angoff method. (Please see Figures 31 and 32.) When asked about the clarity and ease of applications of the rating methods for items in HO blocks, panelists indicated that methods were neither as clear nor as easy to apply when rating the HO blocks. (Please see Figures 33 and 34.)

Debriefing Session

A few minutes after the pilot study was adjourned, a debriefing session was held. Process facilitators, content facilitators, NAS evaluators, NAGB staff and 12 panelists (four for each grade) were present. The panelists were selected and invited to participate three weeks prior to coming to Phoenix. Panelists invited to the debriefing were representatives of the demographic attributes of the PS2 panels. Issues and concerns about the process were discussed. Questions and transcripts of the debriefing are included in Appendix G.

Panelists found the work to be too intense: days were too long, there was not nearly enough time to do the tasks, breaks were too short. Based on recommendations by panelists at the debriefing session, meetings were to start at 7:30 A.M. for the ALS panels in September. Furthermore, longer breaks were to be inserted, and meetings were to be adjourned earlier on most days.

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.

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.