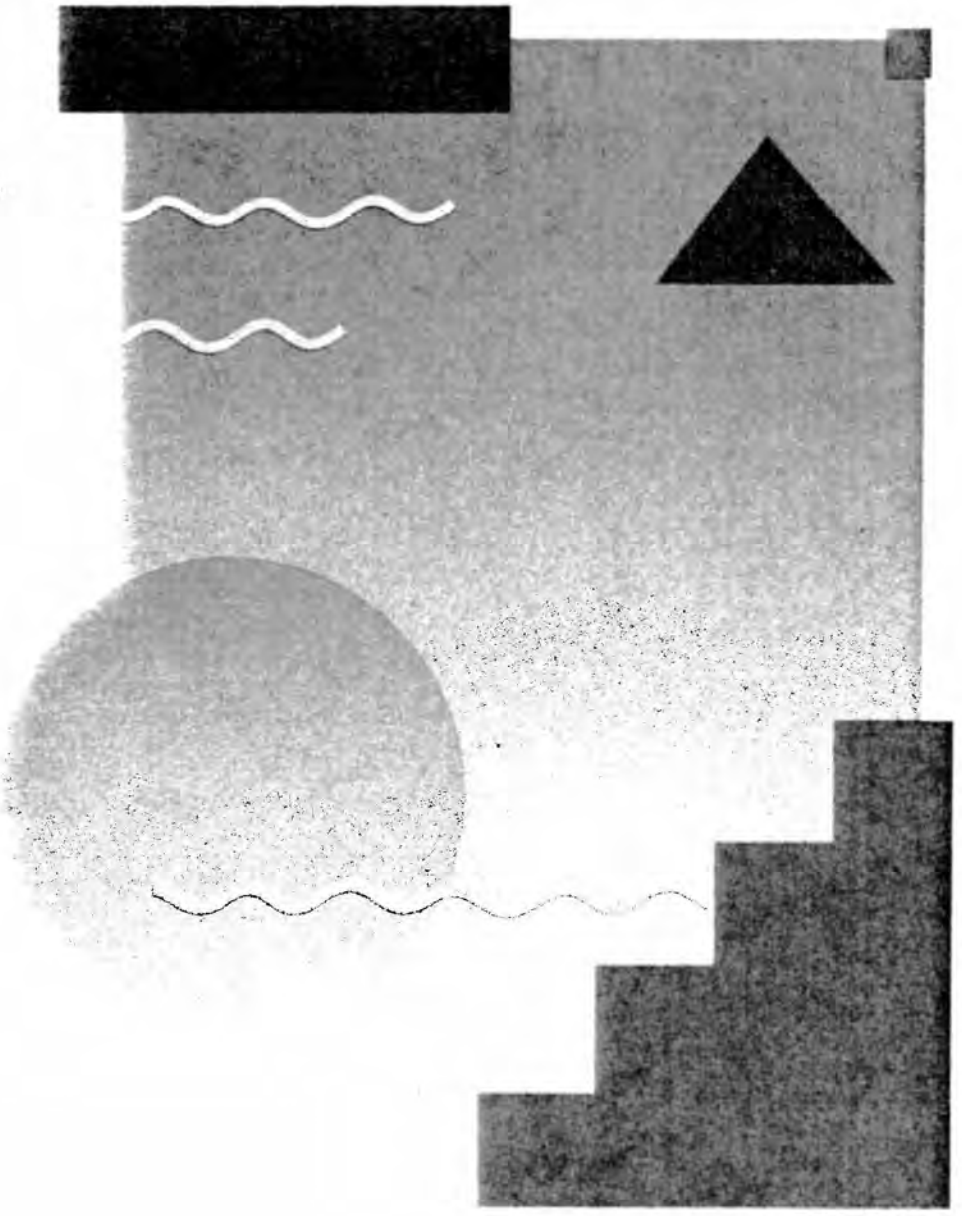
ACT

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

Final Report



Volume I
Pilot Study 1

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

The first pilot study of the 1996 NAEP Science Achievement Levels-Setting (ALS) process was implemented March 21-25, 1996 at the Ritz-Carlton Hotel in St. Louis, MO. This pilot study was the first step in implementing the process of setting achievement levels for the 1996 Science NAEP. Only one pilot study had been planned for the Science ALS process. Because of the inclusion of hands-on tasks in the Science NAEP, however, NAGB suggested that ACT conduct a smaller pilot study to focus more specifically on the hands-on tasks. The second pilot study would be reserved for the more typical role in the ALS process of testing out procedures as planned for implementation in the actual achievement levels-setting process. Thus, the main purpose of this pilot study was to investigate whether procedures that were proposed and used in past ALS processes (geography and U.S. history) could be used successfully in setting achievement levels for the 1996 Science NAEP. This investigation included, but was not limited to, the examination of whether the item-by-item rating process used in setting the cutpoints was applicable to hands-on tasks.

Key Aspects of Pilot Study 1

The design of the pilot study was as similar as possible to the ALS process for the 1994 NAEP in Geography and U.S. History. Please refer to the Agenda in Appendix A. The following discussion is of aspects of the pilot study that differed from the 1994 NAEP ALS processes. Several adjustments were made to the design of the pilot study for the following reasons:

1. The focus on hands-on tasks.
2. The collection of information on the development of borderline descriptions.
3. The availability of materials.
4. The design of the science assessment.
5. The recommendations of TACSS.

Panelists and Item Rating Groups

The sampling design and the nomination and selection procedures implemented were those included in the *Design Document* for the 1996 Science NAEP. The number of panelists included in Pilot Study 1 was different, however. The plan was to have 20, 10, and 30 panelists for grades 4, 8, and 12, respectively. The actual number of panelists in the pilot study was 18, 10, and 29. A list of panelists is included in Appendix B.

For the first time, the specialty of the 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 in science. The nominator was asked to check all that applied for each nominee. Selected nominees were interviewed on the telephone, and they were also asked to identify their

specialties or interests. In addition to the demographic attributes generally considered, science specialty was also used to assign panelists to the two rating groups for each grade. The goal was to have each rating group as equal as possible to the other with respect to panelist type, sex, region, race/ethnicity, and area of science specialty. (Please see Tables 1-3 for data on the demographic characteristics of the panelists nominated, selected, and empaneled. Figures 1-3 show the fields of science for each stage in the panelist selection process.)

Item Rating 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 are paper-and-pencil blocks with both multiple-choice and open-ended items. Concept/Problem Solving blocks include items from the three fields of science (Earth, Physical, and Life) identified in the framework. Theme-Based blocks include items related to one theme identified in the framework, and most of these blocks include items from a single content area. Hands-On blocks involve 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.

Because the hands-on tasks were the focus of the pilot study, adjustments in the usual manner of forming the item rating pool were needed. The goal was to balance the item rating pools with respect to several aspects of the assessment pool: overall average difficulty; test-time for blocks; proportion of multiple choice and constructed response items; number of items in each subscale (field of science); and special types of blocks (e.g., theme blocks).

One criterion was to have all panelists work with all hands-on tasks for their grade level. At the beginning of the pilot study, each panelist took a form of the Science NAEP that included a hands-on block. The remaining three HO blocks were included in each rating pool, two of which were common to the two groups ("common grade blocks"). Two CP and one TB blocks were added to each rating pool so that panelists would have a basis for comparing ratings for different types of item blocks. Item rating pools for each grade level were constructed so that the average p-values were about equal and the distribution of items across content areas and item formats was about the same. (Please see Tables 4-6.) For grade 12, block S3S15, a CP block, was included as an additional common grade block. It was identified by ETS as an "in-depth" block, and ACT felt it might be of special interest to panelists. Notice that not all 15 blocks were included in the item rating pools for each grade level. Moreover, two CP blocks for each grade level were not considered for inclusion in the item rating pools. These CP blocks were field tested in 1995[1] and were not scaled. Those blocks could not be included in the item rating pools for

[1] The rest of the items were field tested in 1993. The field test for the two blocks of CP items was small and did not warrant scaling.

setting cutscores on the achievement levels because there were no item parameters. They were, however, used for practice ratings and exercises, so panelists did become familiar with those items.

Because of changes in the items from the field test to the operational form, many items in other blocks were omitted from computations of the cutscores. Panelists were asked to rate all of the items in their rating pools, despite the fact that some ratings would not be used for computing the cutscores. Items that were significantly changed were dropped from the computations. ACT felt that this would be least confusing to panelists. Further, their ratings could later be used with data from the operational assessment to compare to these using field test data. Table 7 gives the number of items in the item rating pools for which item parameters existed, and the number of items that were used in setting the cutpoints in the pilot study. Data in Table 7 are presented by item type.

The item parameters that were used in computing the cutpoints were not the "official" item parameters. The hands-on blocks were not included in the "official" scaling of field test data, but they were included in a special study by ETS. The latter, more complete, set of item parameters was used in Pilot Study 2.

Several short constructed response items were scored dichotomously. TACSS recommended that they be rated as if they were multiple-choice items; i.e., using the Modified Angoff method to estimate the percentage of correct responses.

Hands-On Sessions

Two sessions were scheduled to address the hands-on tasks. During the first evening, panelists worked with three hands-on blocks for their respective grade levels. (Each panelist had already worked with one hands-on block when taking the Science NAEP.) They were instructed to perform the tasks and answer the questions, just as if they were taking an exam. This activity gave the panelists familiarity with tasks that students performed in the assessment. The goal of this session was to help panelists have a better and more realistic understanding of what was required for students when performing the hands-on tasks. This understanding would be of value when estimating student performance on those tasks.

The second session was included to have the content staff demonstrate a correct way to perform each hands-on task to the panelists. Quite a large amount of time had been spent in consideration of how to handle the hands-on tasks in the rating process. In particular, the technical advisors felt that panelists needed to be aware of the various aspects of the hands-on tasks, some of which are not related to science knowledge and skills. For example, students would have to perform the tasks in classrooms, lunch rooms, libraries, and so forth; not in science labs. Students have varying levels of manual dexterity, and students at grade 4 were assumed to be lacking such skills to some extent. The need for manual dexterity and coordination was expected to be a significant factor in conducting hands-on

tasks. The level of attentiveness required for completing the task under timed conditions was also assumed to have an impact on student performance. And, some students would have never performed such tasks before, while others would have performed a task very nearly the same as that included in their assessment. All of these factors seemed likely to impact student performance. Technical advisors felt that it was important to train panelists in these factors and to make certain that they fully understood how these factors could impact student performance.

ACT Project Staff collected information on filming students taking the NAEP hands-on tasks. It appeared impossible to record actual NAEP sessions without intruding upon the normal testing conditions. Students could be engaged in the hands-on tasks for purposes of filming. The final decision against filming students for purposes of training ALS panelists, however, was based on the amount of time that would be required of panelists to view the videos of students at their grade levels. Several hours would be required for panelists at all grades. Ultimately, agreement was reached on the following steps to train panelists in the hands-on tasks:

- Each panelist would perform each hands-on task.

- Panelists would be instructed in the (non-science) factors that might influence student performance on the hands-on task.

- Content staff would demonstrate the "correct" way to perform each hands-on task and panelists would have the opportunity to discuss the tasks and ask questions for clarification to assure that each panelist understood the purpose of the task and how to perform the task.

The content staff strongly protested the plan to demonstrate the "correct" way to perform the tasks and indicated this would seem demeaning to the panelists. Instead of demonstrating the tasks to the panelists, they facilitated a discussion on the rationale for each hands-on task; i.e., the skills and knowledge that were being tapped by each task. They also agreed to demonstrate all or part of the tasks if there were *any* concern or hesitancy expressed by any panelist regarding the procedure to follow. This agreement was followed to varying extents across the three grade groups. For example, some groups discussed the purposes more than the "how-to," some discussed the purpose in relation to the framework more than the purpose in relation to student performance, and some gave more emphasis to panelists' questions than to initiating discussion and instructions.

The plan had been presented to content staff during the training prior to Pilot Study 1, but the reality of how that would be carried out was less clear during that session. Further, content staff were somewhat "flooded" with information during that training session, and they needed more time to reflect on the implementation of many aspects of the process. This aspect needed more clarification and agreement among staff before implementation in Pilot Study 2.

Performance at the Borderline

In the past, panelists were instructed to use the Achievement Levels Descriptions (ALDs) and their concept of borderline performance at each achievement level to rate the items. Several exercises were performed to help panelists gain confidence in their concept of borderline performance. Since the ratings are provided "at the borderline" of each achievement level, TACSS recommended that more concrete descriptions were in order to assure that all panelists gained a clear understanding of such performance.

Following the recommendation of TACSS, the panelists were directed to produce operational descriptors of borderline performance for each achievement level. Following considerable discussion by TACSS regarding the development, modification, and use of the borderline descriptions, the following instructions regarding the borderline descriptions were implemented.

- Borderline descriptions could be in "bullet" format; sentences were not encouraged.

- The descriptors were for panelists' own use in rating the items and not an outcome of the pilot study.

- Panelists were allowed to modify the borderline descriptions in conjunction with their modifications of the ALDs.

The paper selection exercise did progress more rapidly than usual, perhaps because the panelists had reached agreement on written descriptions of borderline performance.

The development of borderline descriptors did, however, take on a more primary role than intended. Indeed, the development of borderline descriptors seemed to be detrimental to the process of reaching common agreement on the meaning of the achievement levels descriptions. Grade 12, for example, moved into development of borderline descriptors before they had formed a group understanding of the ALDs. All three grade groups struggled with the borderline descriptors and with refining and modifying those instead of the ALDs.

Paper Selection Exercise

ACT had used a paper selection process as the means of rating constructed response items in the 1992 NAEP ALS process for writing, reading, and mathematics. Although the mean estimation method of rating polytomous items was adopted for the 1994 and 1996 ALS processes, TACSS also felt that it was important to maintain some aspects of the paper selection process in the achievement levels-setting process. The paper selection exercise was designed, in accordance with TACSS recommendations, to accomplish the following purposes:

- Provide a reality check about how students respond to open-ended questions.

- Promote a firmer conceptualization of performance at the borderline.

- Give panelists an opportunity to become familiar with the scoring rubrics and scoring for the constructed response items.

In this exercise, panelists were to select responses to represent performance at the borderline for each open-ended item in their rating pool for which papers were available. Most items that were significantly changed from the field test to the operational form were not included in this exercise. ETS was very generous in their support of our goals. They instructed NCS to score hands-on tasks first so that we would have papers from operational test forms for those blocks to use for this pilot study. Scoring of items began only about one week before the pilot study, however, so this made the task of collecting papers with the desired distribution of scores rather difficult. (Tables 4-6 include the approximate number of items that were significantly changed for each block in the rating pools.) For the items that were used in this exercise, the goal was to have three sample responses for each score code. For some items, there were no sample responses at the highest score level. Those items were also excluded from this exercise.

For the hands-on blocks, the plan was to have panelists select a student's set of responses to the whole block to represent borderline performance at each achievement level. This plan was developed because we were informed that the HO blocks would be scored such that one scorer would score all responses for a particular student for a hands-on task. This procedure had been selected because of possible dependencies in responses within the hands-on tasks. Further, students were to be given credit on subsequent answers that followed correct procedures but were based on incorrect answers to earlier questions. Thus, it required that the same scorer evaluate all responses, in tact, for a single student in order to follow the sequence of responses to the whole hands-on task. In fact, only one block per grade level was scored "in tact" for each student. Some items in some of the blocks were scored together (e.g., items 3, 4, and 6 were scored together), and some were scored independently. Panelists were given sample responses to select from in accordance with the way the items were scored.

Feedback

Several pieces of information are typically provided for panelists to use during the rating process. Following each round of ratings, panelists were given information on the overall average ratings at each achievement level (the cutscores) and the variability of the grade-level averages. "P-value" data for each item were provided to inform panelists of the performance of students on each item. They were provided with "Whole Booklet" information that shows the percent of total points required to score at the cutscore of each achievement level set at each round. The booklet on which this information was provided was the same form used to test panelists at the start of the ALS process.

In previous ALS processes, panelists were given information about the variability of individual item ratings relative to their overall cutscore. Because the data from the field test were not complete for the items in the operational test form, not all these feedback procedures could be implemented. Further, some feedback had a lot of "missing data."

The cutpoints and standard deviations were presented to the panelists graphically for the first time. All feedback presented to the panelists are in Appendix C. There were no changes in the way that the interrater consistency feedback was presented. P-value data were only provided to the panelists for those items that were not significantly changed from the field test. For each rating group, there were one or two HO blocks with only one or two items that had not been significantly changed. ACT decided not to give any p-value information for those blocks. At the recommendation of TAT, TACSS decided that intrarater consistency feedback would not be provided to panelists. However, based on comments given in the debriefing session, some feedback of this type should be provided. The Whole Booklet Exercise was not implemented because scores for operational forms of student booklets were not available.[2] No whole booklet feedback[3] was provided because data were too sparse to have confidence in the information that would have been conveyed by the feedback.

Selection of Exemplar Items

One outcome of the ALS process is a set of exemplar items to illustrate the kinds of performance associated with each level of achievement. The purpose of these items is simply to communicate more clearly to users of NAEP data what is meant by Basic, Proficient, and Advanced performance.

Some blocks of items are released for public review after each administration of NAEP. Typically, two or three blocks of items are provided for public review and use. Those blocks of items are those from which exemplar items may be selected to report, along with other information about achievement levels.

ACT has implemented several different procedures for selecting exemplar items. The TACSS recommended that items be statistically selected and presented to panelists for their selection according to the match of the item to the achievement levels descriptions. It is clear that not every item that meets statistical criteria will meet content criteria as a match to the achievement levels descriptions. TACSS recommended that items having a 50% probability, on average, across the range of

[2] Sample student booklets were shown to the panelists for one exercise in which they applied their understanding of the ALDs holistically.

[3] The whole booklet feedback reports the expected score (as the percent of maximum score points) for students whose level of performance was at the cutpoint of each achievement level. The booklet used for this feedback was (would have been) the NAEP form that the panelists took on the first day of the process.

scores included in the achievement level be presented to panelists for their consideration. Following the 1994 geography and U.S. history exemplar item selection process and review by the NAGB Achievement Levels Committee, a second criterion was added to the statistical evaluations of items. The Achievement Levels Committee recommended that a discrimination criterion be added such that items presented at one level have a considerably lower probability of correct response at the next lower level.

Exemplar items were selected for CP and TB blocks based on the two statistical criteria. First, items for consideration as exemplars for an achievement level must have an average conditional p-value (ACP) of at least 50% across that level. Second, for each item, the difference (DACP) between its average conditional p-value at that level and its average conditional p-value at the next lower level was 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 items that satisfied only the first criterion. Panelists were instructed to consider items from the secondary list if they rejected all the items in the primary list.

For HO blocks, panelists were instructed to select the task that would best illustrate what students should know and be able to do across the three achievement levels.

Debriefing Session

Very shortly after the pilot study was adjourned, a debriefing session was held. Present were the process facilitators, and 11 panelists who were selected and invited to participate approximately three weeks prior to coming to St. Louis. ACT selected four persons at each grade level so that the panel of 12 would include panelists of approximately the same composition as represented on the overall panel of raters. Several issues and concerns about the process were discussed, some of which focused specifically on the hands-on tasks in the science assessment. A somewhat annotated transcript is included in Appendix D. Major points are summarized later in this report.

Panelists

The panelist selection process described in the *Design Document* was implemented. Samples were drawn without replacement from each of the three sampling frames (public school districts, private schools, and colleges and universities) for both pilot studies and the ALS. Although the plan was to empanel only 60 persons in this pilot study, the sampling design for recruiting 90 panelists was used. This means that more nominators were contacted than would have been the case for only 60 panelists. Moreover, state officials who were to nominate persons for all three studies were allowed to send their full quota of nominees at one time with the assurance that nominees not selected would be retained for possible selection in future studies.

A total of 462 persons were nominated. (Please see Table 1.) The number of nominees for each grade was not very different. The representativeness of the nominee pool and the subsequent panels was not even. There were only 41 nominees—9% of the nominee pool—for the general public panel positions. The design called for 30% of the panelists to represent the general public. The largest number of nominees were from the central region, and the northeast had the fewest nominees. There were more female nominees than male. At grade 12, however, there were more male nominees. Over 20% of nominees for each grade level were from non-majority racial/ethnic groups. Figure 1 indicates the content area of expertise or interest of the nominees as reported by the nominators. Notice that as the grade level increases, the nominees tend to become more specialized by field of science.

Table 2 gives the distribution of the selected nominees with respect to grade level, type, region, sex and ethnicity. Figure 2 gives the distribution of the selected nominees with respect to (nominator reported) content area specialty or interest.

The required distribution of panelists (55% teachers, 15% nonteachers, and 30% general public) was only missed by very narrow margins. (Please see Table 3.) There was an overrepresentation of the central region, and the southeast was underrepresented. Moreover, minorities were underrepresented in general, and women were hardly represented at all in grade 12. Many grade 4 panelists reported either no specialty (i.e., general science) or specialty in all three content areas. (Please see Figure 3.) Most grade 12 panelists reported Physical science as their area of specialization. The lack of grade 12 panelists in other content areas is probably attributable to our usual recruitment procedures. ACT has specified in the letter to nominators and in the guidelines for nominations that the teachers must "teach 12th grade science." We feel that this generally excluded teachers of subjects other than physics. (The communications will be changed for subsequent recruitment in science.)

Results

Achievement Levels Descriptions

The Process

The plan was to train panelists in the framework and preliminary achievement levels descriptions so that they would understand how the descriptions "fit within" and "come from" the framework. The steps in this process begin with a presentation by content facilitators to the general session at the beginning of Day 2 (the first full day). At that time, all panelists should have learned about the framework, the policy descriptions of the three levels, and the preliminary science achievement levels descriptions. That did not happen. The content facilitators discussed the framework, extensively. They also discussed other such documents, including the National Science Standards. They did not discuss the policy

descriptions or the achievement levels descriptions for the grades. The policy definitions were, however, shown to the panelists before they left the general session. Panelists were instructed to think about those three definitions with respect to the preliminary descriptions when they began work in grade groups.

During the process, panelists were engaged in activities focusing on the achievement levels descriptions. There were three time periods during which they were given the opportunity to modify or make adjustments to the ALDs for their respective grade levels. The largest, single amount of time devoted to the ALDs was Day 2. The entire day was devoted to internalizing the preliminary ALDs and beginning work on the borderline descriptors. The three content facilitators each approached the task differently. In grade 8, the content facilitator had prepared a parsed list of descriptors that showed the alignment of content/themes across the three levels. (Please see Appendix E, pp. E-11 and E-12.) The list was distributed early in the session, and panelists embraced it. In grade 12, the panelists never really addressed the task of reaching an understanding of the ALDs until after the first round of ratings. They did become engaged in developing borderline descriptors, but they went into the first round of ratings without having really worked on the preliminary descriptions and with little internalization of their meaning. In grade 4, panelists spent quite a lot of time working on the ALDs and the borderline descriptions. They spent more time on the borderline descriptions at an earlier period in the process than had been planned, but they did closely follow the planned sequence of events.

The Products

The "chart" of descriptors developed for the grade 8 panel was simply by achievement level. (Please see Appendix E for the ALDs and borderline descriptions developed by each grade group.) The chart that was ultimately developed for the grade 12 panel followed the matrix of the framework with descriptors for each of the different dimensions (ways of knowing and doing science, themes, and so forth). The grade 12 content facilitator changed some parts of this chart *after* the pilot study. The concern shared by one observer of the fourth grade process of modifying ALDs before Round 3 was that the panel was making modifications to accommodate items on the assessment that were not represented in the ALDs. Whether those modifications were consistent with the framework had to be verified more carefully with the content staff.

As the previous discussion perhaps indicates, the focus was on alignment of descriptors across achievement levels. For the most part, the modifications recommended were with respect to the continuity of the descriptions across achievement levels. Grade 12 work can serve as an example. At the Basic level, the preliminary ALD states that students should "be able to apply fundamental facts, concepts, and principles to situations encountered in daily lives." (Please see Appendix E.) Panelists and the content facilitator indicated that statement pertained to practical reasoning. (Please see Appendix E, p. E-24.) However, there was not a statement in the Proficient and Advanced descriptions that pertained to practical reasoning. Thus, in the modified version they added the following

statements: students performing at the Proficient level should "be able to apply facts, concepts, and principles to problems in the global environment"; and students performing at the Advanced level should "be able to use scientific ideas to question common sense ideas, for instance relativity." In addition, they modified the "practical reasoning statement" at the Basic level. (Please see Appendix E, pp. E-28—E-32.)

Borderline Descriptions

During this pilot study, the borderline descriptions seemed to have taken time and received attention at the expense of the ALDs. Grade 12 panelists, for example, did not have time to address modifications to the ALDs on Day 2 because they had used all of their time on borderline descriptions. The paper selection process seemed to go much more smoothly and at a faster pace than in the past. The process facilitator who had worked with the 1994 ALS Process felt that this was largely due to having borderline descriptors developed already.

Evaluations by panelists of their conceptualization of borderline performance did not reveal improvements in clarity of conceptual understanding over previous studies during which the descriptions had not been developed.

Cutscores

The cutscores set in this pilot study are reported on the ACT NAEP-like scale. The cutscores and standard deviations are reported in Table 8. Findings from previous studies (geography and U.S. history) revealed a decreasing standard deviation from the Basic to the Advanced levels. That was **not** the case for the standard deviations of the cutscores in science at any grade or any round except for grade 8, Round 1.

Hands-On Tasks

Because a primary focus of the pilot study was the hands-on tasks, ACT wanted to present rating results for those blocks. There were so many items for which the parameters were not used in estimating the cutscores, however, that this seemed impossible. Cutscores and standard deviations were computed for all items *except* the HO blocks, and the results are reported in Table 9. The differences in the respective cutscores were very small, generally within two points on the scale. This was expected because of the fact that so few items in the HO blocks were included in computing the composite cutscores. Nonetheless, it is interesting to note that the cutpoints set without the HO blocks are consistently higher in all cases except grade 4 Advanced.

ACT was anxious to determine whether that pattern would be found when all item parameters were available for computing these cutscores. The data in Table 10 were computed from the "raw ratings." These data show the ratings with both dichotomous and polytomous items averaged in the percent correct metric. These rating data confirm that panelists at grades 4 and 12 expected student performance to be relatively higher on the hands-on tasks than on the other types of items. That is, these panelists' ratings indicated that they perceived the hands-on tasks to be

relatively easier than the other items in the assessment. Panelists at grade 8 rated both hands-on and other item types very similarly. See Volume 2, Table 19 and Volume 3, Table 18 for comparisons of hands-on ratings versus other item types in Pilot Study 2 and the ALS, respectively.

Dichotomous and Polytomous Items

Because polytomous item ratings have generally been found to result in significantly higher cutscores than those for dichotomous items, a test of significance (at the .05 level) was performed on the differences between cutpoints set for these two types of items. The results are reported in Table 11. Significant differences were found across all rounds for grade 4 at the Proficient and Advanced levels and for grade 12 at the Basic level.

The raw rating data were again used to compute the average percent correct ratings for dichotomous and polytomous items. These results are presented in Tables 12-14 across rounds and grades for each achievement level. Those data show that panelists judged polytomous items to be considerably more difficult than dichotomous items, and that was especially true at the two lower grade levels. It is also instructive to note that in grade 8, the two rating groups rated polytomous items quite differently. Finally, it is clear from the raw ratings that the panelists at grades 4 and 8 either had lower expectations for students or perceived the items to be much harder than was the case for panelists at grade 12. This was especially evident for ratings at the Basic level. This indicates that content experts must examine the achievement levels descriptions to determine whether the statements of what students *should know and be able to do* at grades 4 and 8 are less rigorous than at grade 12. Without complete item parameters, however, it is not possible to determine the impact of the differences in ratings on the cutscores that would have been set on items of the two different types.

Item Content Areas

Analyses were conducted to determine whether there were significant differences in the cutpoints set for different content areas. The results of these analyses are reported in Tables 15-17 for each grade. Significant differences were found only for grade 8 at the Basic level across all rounds. The *pattern* of ratings across the three content areas was not consistent for each of the three grade levels.

The hypothesis tested was that panelists would set higher cutpoints in the content area in which they had expertise or special interest. The analyses to test this hypothesis showed no significant relationship for grade 4. (Please see Table 18.) At grade 8, a significant difference was found for ratings by Earth Science panelists (*versus* all others) and for Physical science panelists (*versus* all others) at the Basic and Proficient levels. (Please see Table 19.) At grade 12, ratings by Physical science panelists (*versus* all others) were significantly higher at the Basic and Proficient levels. (Please see Table 20). Only for grade 12 was the cutscore set for Physical science items by Physical science panelists higher than that set by non-Physical panelists. At the grade 12 Proficient level, however, the cutscore set for

Physical science items was the lowest of the three content areas for both Physical science panelists and all other panelists.

Item Rating Groups

Tests were conducted to determine whether there were significant differences in cutpoints set by panelists in the two item rating groups for each grade level. Recall that rating groups were formed to be as equivalent as possible with respect to demographic attributes and content specialties of panelists. The cutpoints, by item rating group, are reported in Table 21. In order to test the significance of differences, the rating group cutpoints were computed by averaging the cutpoints set by individual raters. The averages and standard deviations are reported in Table 22. There were no significant differences in the cutpoints set by the two rating groups at each grade level.

Panelist Type

Tables 23-25 report the results of the comparison of the cutpoints set by different types of panelists. Previous studies have not revealed a consistent pattern of significant differences by panelist type, despite the expectations of many that such differences would be great. To the extent that significant differences were found, the most frequent pattern (although not the only one) was to find that teachers set cutscores that were significantly lower than those set by general public panelists. For this pilot study, there were no significant differences except for grade 8. Because the number of panelists was so small for grade 8 (10 in all), little importance can be placed on this finding. The *pattern* found for grade 4 was the more frequently found pattern whereby cutscores set by teachers were lowest and those set by general public panelists were highest, but the differences were *not* statistically significant.

Table Groups

Panelists were assigned to table groups (usually five panelists each) according to the same criteria used to form item rating groups. Most of the activities and discussions were within table groups, although those were frequently followed by "cross table" activities and discussions and grade level activities and discussions.

During the debriefing session, a grade 12 panelist commented that panelists at the table adjacent to his were consistently at the high end on the interrater consistency feedback charts. Based on this anecdotal evidence, an analysis of the cutpoints by table groups was performed. The results are reported in Tables 26-28. There were no significant differences *except* for grade 12 at the Basic and Proficient levels across all rounds, where panelists in Table 5 consistently set the highest cutpoint. That table, by the way, *was the table* to which the panelist had referred.

As a result of this information, plans were made to mix table groups for discussion. In previous ALS processes, exercises had been included to engage panelists in discussion with each other across table groups and across rating groups. This was simply an oversight in the planning and implementation of Pilot Study 1.

Exemplar Items

The process for selection of exemplar items was implemented according to the guidelines recommended by TACSS, along with some adjustments/decisions made on-site. The statistical information about the items that passed the statistical criteria is presented in Tables 29-31. The items passing the difficulty criterion (average percent correct across the level ≥ 50%) were rank ordered by their discrimination index. The "DACP" was computed as the difference in the average conditional percent correct across the level in question and that for the next lower level. TACSS had recommended that the DACP corresponding to about the 60th percentile should be used as the discrimination value for selecting exemplar items. Based on Round 2 ratings,[4] a DACP ≥ 30 across all levels for grade 4 would include about 60% of the items. For grade 8 the DACP cutoff was 30 at the Basic level and 32 at the Proficient and Advanced levels. For grade 12, the cutoffs were 24[5], 31, and 32 at the Basic, Proficient, and Advanced levels, respectively. The items passing both the difficulty and discrimination criteria were included in the primary list. All other items that had an average conditional p-value ≥ .50 were included on the secondary list.

Because no blocks had yet been identified for public release, project staff determined the blocks to be used for the exemplar item pool for each grade. Grade 4 panelists selected items from four blocks: one TB and three CP. Information about those items is included under the heading "Presented Items" in Tables 29-31. Grade 8 panelists selected from two blocks: one TB and one CP. Grade 12 panelists selected from three blocks: one Theme-Based and two Concept/Problem Solving. Panelists were instructed to select items from the primary lists. Only if they rejected all items from the primary list at an achievement level were they to consider items from the secondary list.

Because of the nature of the hands-on task blocks, project staff decided to have panelists identify the hands-on task block that they would recommend as representing the knowledge and skills that students should exhibit across the three achievement levels.

The lists of exemplar items selected for each achievement level for each grade are included in Appendix F. More complete statistical information about the exemplar item pool and those recommended is included in Tables 29-31.

[4] Round 2 ratings were used for the pilot study to save time in getting the lists prepared for review by panelists following round 3. Previous experiences indicated that the differences would be minor. That was not, however, the case for this pilot study. As a result, only about 40% of the items were included in the primary list for grade 4 Basic and grade 12 Advanced.

[5] The DACP cutoff for grade 12 at the Basic level was based on round 3 ratings.

Other Results

Consequences Data

Beginning with the 1994 NAEP Geography ALS Process, consequences data have been provided to panelists. NAGB has maintained a policy of having the ALS process be criterion referenced. Consequences data were provided to panelists only after the final round of ratings had been collected. This meant that reactions from panelists to these data could be collected while adhering to NAGB policy. After the achievement levels had been set in Round 3, panelists were given information about how student performance was distributed with respect to the cutpoints; i.e., the percentages of students scoring at or above each achievement level. (Please see Figures 4-6.) These percentages were computed as estimates, based on a normal distribution; they were not based on actual distributions of student performance on the NAEP field test.

Panelists were asked to respond to a questionnaire designed to ascertain their opinions regarding those percentages, and whether they would adjust their ratings in order to increase or decrease the percentages. One grade 12 panelist did not complete the questionnaire.

In response to question 1:

> **Q1**:Given your understanding of borderline student performance at each of the three achievement levels, do these percentages reflect your expectations about the proportions of students at this grade level whose NAEP score would be at or above the cutscore of each of these achievement levels?

Forty-five (80%) panelists said **yes**, and 11 (20%) panelists said **no**. The panelists who said **no** were asked to respond to question 2[6]:

> **Q2**:Having seen the data on the percentages of students at this grade level whose score on the NAEP was at or above the cutscore for each achievement level, would you change one or more of the achievement levels you have set if you could?

Of 11 panelists who responded to this question, 5 said **yes**, and 6 said **no**. Of the 5 who said yes, 4 would **make no changes** at Basic level and one would lower the cutscore; 1 would **make no change** at Proficient and 4 would lower the cutscore; all 5 panelists would lower the cutscore for Advanced.

[6] The data reported in response to the following questions were modified to follow the contingencies of the questionnaire. Complete tables with modified and unmodified data are included in Appendix G.

Fifty-five panelists responded to question 4:

Q4:What recommendations do you wish to make to the National Assessment Governing Board regarding the cutscores set for the achievement levels?

___I would recommend that the achievement levels be reported as set.

___I would recommend changes consistent with my answers above. If you wish, comment on the magnitude of change you would recommend.

Fifty (91%) recommended that the achievement levels be reported as set, and five (9%) recommended changes consistent with their previous answers. Four grade 12 panelists commented on the magnitude of changes that they recommended. One panelist suggested that the percent increas by 10%, but there was no indication of the level(s) to which the increase should be applied. Two panelists indicated that the cutpoints for Proficient and Advanced levels should be lowered. One suggested four points on (the ACT NAEP-like) scale, and the other suggested that the cutpoint should be lowered by one standard deviation. Lastly, one panelist suggested that the Advanced cutpoint should be 192.

We were interested in knowing whether panelists who recommended changes in the cutscores had higher or lower levels of interrater consistency. The hypothesis tested was that panelists who indicated (in response to question 4) that they would change the cutpoints were more distant from the grade-level cutscore, as indicated on their interrater consistency feedback from Round 3. The locations of their respective cutpoints for the levels for which they recommended changes were examined. No pattern was found.

We also investigated whether the panelists who recommended changes were the same panelists who indicated they would not be willing to sign a statement recommending use of the achievement levels resulting from this study; i.e., item 17 in Process Evaluation Questionnaire No. 7. Of the four panelists who recommended changes in achievement levels cutscores on the Consequences questionnaire, one said "No, probably not," two said "Yes, probably," and one said "Yes, Definitely." Apparently those panelists who recommended changes were not so opposed to the levels that had been set that they would not sign a statement to recommend them to others. In response to this question about signing a statement of recommendation regarding the outcomes of the ALS process, two other people who responded "No, probably not" did not recommend changes in the cutpoints. No panelist would "definitely not" sign the statement.

It seemed surprising that no one in either grade 4 or 8 recommended that the cutpoints at the Basic level be raised to decrease the percentage of students scoring at or above the Basic level. For both grades 4 and 8, the estimated percentages of students scoring at or above the Basic level were about 95%. This is not consistent with previous ALS experiences, with the exception of the grade 4 U.S. history pilot study. This point was pursued during the debriefing session. Panelists commented extensively on the low percentages scoring at or above the

Proficient and Advanced levels, but they did not volunteer comments on the high percentages at grades 4 and 8 that would be at or above the Basic level. Please refer to page D-16 of the Debriefing Session Transcript in Appendix D.

Process Evaluation

Seven evaluation questionnaires were completed by panelists, one at the end of each day and one additional one following the presentation of feedback after Round 1 ratings. These questions were included in the geography and U.S. history ALS evaluations, except for additional questions regarding the hands-on blocks and the use of feedback data. Detailed results of the analyses are included as Appendices H and I. Appendix H gives the responses by grade level, and Appendix I gives the responses by panelist type.

Some results of the process evaluation are included in Figures 7-12. Panelists were asked to indicate the clarity of their understanding of the ALDs at each level. As is typically the case, they reported that their understanding of the ALDs was clearer prior to Round 1 than after Round 1. But their level of understanding increased again after Round 1 and was highest after Round 3.

Panelists were also asked to indicate how well formed their concept of borderline student performance was during each round of rating. Figure 8 shows that their conceptualization of borderline student performance became more well formed across rounds.

A cursory look at responses to these items by U.S. history ALS panelists indicated that the mean level of understanding the ALDs reported by science Pilot Study 1 panelists was slightly higher by Round 3 than that for U.S. history ALS panelists. The mean level of clarity regarding their conceptualization of borderline performance was slightly higher at each round than that for the history panelists. The differences in mean levels of clarity of conceptualization of borderline performance at Round 3 were .12 at Basic, .05 at Proficient, and .08 at Advanced. Given the fact that the borderline descriptions were given more emphasis in Pilot Study 1 than intended, the gains in clarity over the process for which no descriptions were developed for borderline performance seem very low.

Responses regarding the rating methods used in setting the cutpoints show that the clarity and ease of applications of the methods increased across rounds. (Please see Figures 9 and 10.) The mean estimation method, however, was consistently judged to be less clear and less easily applied than the modified Angoff method. This finding was consistent with those for geography and U.S. history.

When asked about the clarity and ease of applications of the rating methods with respect to items in HO blocks, panelists indicated that the method was less clear and less easily applied to the HO blocks. (Please see Figures 11 and 12.)

Debriefing Session

Several interesting recommendations were put forth by panelists during the debriefing session. The following questions/topics were covered during the session.

1.What were the strongest/most positive aspects of the process and what were the weakest or most negative aspects? Any suggestions for changes to the process?

Panelists felt that the review of student booklets and papers was particularly helpful, and that having three rounds of ratings was a positive aspect. The long days and frustration experienced in the beginning were negative aspects.

A very interesting suggestion for change was offered whereby panelists would participate in the first round of ratings much earlier in the process. For that round of ratings, they would be given student performance data (p-value feedback). They would discuss their ratings and feedback from that round. Later, ratings for Round 2 and Round 3 would be done without reference to the Round 1 ratings. Panelists would never see the Round 1 ratings again. Those Round 1 ratings would truly be a practice round.

Intrarater feedback data were described. Panelists were asked whether that would seem helpful. They felt that it would have given them something to focus on in their final round of ratings.

Comments by panelists revealed a very strong reliance on the interrater consistency feedback information. Panelists seemed to have determined whether the location of their individual cutscores were "acceptable" with respect to the grade level cutscores. If they felt comfortable with that, they were reluctant to change their ratings. Grade 12 panelists indicated that even though they had no clear agreement on the ALDs when they provided their Round 1 ratings, they were very reluctant to change their ratings in subsequent rounds. This discussion led to the discussion for having Round 1 serve as a training round.

General public panelists voiced a real frustration with the process and a real sense of impatience with the pace of the process.

2.**If**, for some reason, the hands-on tasks were omitted from the reporting scale so that they were not used in computing the final composite performances measures, what impact would this have? Would the achievement levels descriptions still be valid? Would the assessment still reflect the content of the framework adequately?

The discussion of this was lengthy, and took several turns. The final consensus seemed to be that the assessment would not reflect the framework completely

and adequately without the hands-on tasks nor would the achievement levels descriptions be valid if the hands-on tasks were omitted. In part, the discussion was lengthy because panelists were so committed to the hands-on tasks. They all felt that the hands-on tasks were outstanding and that the assessment would not be nearly as good without it. In general, they finally agreed that it would be possible to measure **most** of the knowledge and skills that students *should* have at each level of achievement, but the wording of the achievement levels descriptions would have to be modified if hands-on tasks were omitted. They also discussed the fact that the descriptions of what students *should* know and be able to do would hold without hands-on tasks, since they believed that hands-on tasks should be included. If the assessment excluded hands-on tasks, however, it would not reflect the framework. Thus, without the hands-on task items, the framework would not be tied to the achievement levels descriptions.

3.What were panelists reactions to the consequences data? Were they surprised? Did the percentages of students performing at or above each level seem reasonable and in line with their expectations? Would they have preferred having the data earlier in the process or more frequently in the process?

Panelists were not surprised by the consequences data, generally. One grade 12 panelist *did* believe that her experiences and observations suggest that more students achieve at the Advanced level than the consequences data indicated.

One panelist indicated that he felt that he would surely be considered at the grade 12 Advanced level, yet he was not certain that he would fit into that percentage distribution indicated by the consequences feedback.

The discussion also focused on the fact that most of the grade 12 teacher panelists were involved with Advanced Placement courses in science. There was agreement that panelists at grade 12 were generally likely to have contact *exclusively* with Advanced students while panelists at lower grade levels would have contact with a wider range of students on a more regular basis.

References

American College Testing (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.