Developing Achievement Levels for the 1998 NAEP in Writing Interim Report: Pilot Study

Susan Cooper Loomis, Patricia L. Hanick, & Wen-Ling Yang
ACT, Inc.

December 2000

The work for this report was conducted by ACT, Inc. under contract ZA97001001 with the National Assessment Governing Board.

Table of Contents

EXECUTIVE SUMMARY

Susan Cooper Loomis

The achievement levels-setting (ALS) process for the 1998 National Assessment of Educational Progress (NAEP) was implemented as a pilot study prior to the operational ALS. The writing pilot study was an opportunity to continue studying and refining procedures that were introduced in earlier field trials for writing and in the 1998 Civics NAEP ALS Pilot Study. The pilot study enabled ACT to identify further modifications to training, instructing, timing, and other key elements of the process to be made before the operational Writing ALS process was implemented.

ACT was particularly interested in evaluating panelists' reactions to incorporating the new Reckase Charts as feedback in the ALS process. Reckase Charts had not been implemented in that manner prior to the writing pilot study. Some significant changes in the process were introduced as a result of observations and findings from the civics pilot study. In addition to a modification to the role of Reckase Charts in the process, consequences data were provided to writing pilot study panelists three times: after the second round of ratings, after the third round of ratings, and at the end of the entire process. Panelists had an opportunity to change cutscores based on both the Round 2 and Round 3 feedback. The writing pilot study was a final check on procedures to assure a successful operational Writing NAEP ALS.

OBSERVATIONS AND EVALUATIONS OF THE WRITING PILOT STUDY PROCESS

AGENDA ADJUSTMENTS NEEDED

- ✓ The early days of the process were still too busy.
- ✓ Panelists appreciated having a social time, and they enjoyed the social mixer. Try to arrange the agenda so that Day 1 activities are completed with the dinner meal.
- ✓ With few items to rate, the rating process required little time after the first round. Adjust the agenda to include a longer break between Round 3 ratings and Round 3 feedback.

UNDERSTANDING COMPENSATORY SCORING

- ✓ Concern has been expressed about the need for panelists to understand that a compensatory model is used to scale NAEP data.
- ✓ Panelists have no direct experience with a compensatory model when they use an item-by-item rating method.
  - When engaged in exercises requiring holistic judgments, they were resistant to accepting the reality of uneven student performance and the concept of compensatory scoring.
  - Some panelists were very surprised by the fact that a student might give an "off task" response to one writing task and write a good response to the second.

RATING PROMPTS

- ✓ Panelists were quite disturbed about estimating an average score.
  - Many were fixed on the whole numbers associated with the scoring rubric and could not reconcile themselves to an average score including a decimal value.
  - When instructing future panelists in the rating method, the meeting facilitator will include an example of computing an average for which the result requires a decimal value.

WRITING BORDERLINE DESCRIPTIONS

- ✓ The ALDs for Basic performance for all grades were very brief.
  - It was difficult for panelists to describe borderline Basic performance that was more minimal than that described by the Basic achievement level description.
  - Writing borderline descriptions did not have a significant impact on the panelists' perception of the clarity with which they conceptualized borderline performance. Panelists in the writing pilot study who wrote borderline descriptions reported no clearer concept of borderline performance than panelists in other ALS studies who only discussed borderline performance.

PRESENTING FEEDBACK IN ORDER

- ✓ Beginning with the writing pilot study, the order of presenting feedback was from the most holistic level to the most item-specific level.
  - The exception to the ordering was consequences data which were distributed for the first time after Round 2.
  - The ordered feedback seemed to help panelists identify the different pieces of feedback more readily, but several panelists seemed to want to move to rater location feedback and student performance data rather than examine student performances in the whole booklet exercise. This ordering was used:
    1. Cutscores and their standard deviations (all rounds)
    2. Whole booklet feedback (all rounds)
    3. Whole booklet exercise (Round 1, only)
    4. Rater location data (all rounds)
    5. Student performance data (available for all rounds after Round 1)
    6. Reckase Charts (Rounds 1 and 2)
    7. Consequences data (Rounds 2 and 3 and Final)

USING AN ELECTRONIC SCANNER FOR DATA ENTRY

- ✓ The experience with the scanner was not positive, overall.
  - Because there were few prompts to rate in the Writing NAEP, the problems were not cumbersome.
  - The scanner did not eliminate the necessity for key entry of a significant number of rating forms.

CONCLUSIONS DRAWN FROM WRITING PILOT STUDY RESEARCH

THE ISSUE OF IMPROVING AND REFINING THE STANDARD SETTING PROCESS

Years of refinements have led to the current process, which has been considerably enhanced by the most recent addition of the Reckase Charts. The charts were created specifically for use in setting NAEP standards, although they could be used easily in other standard-setting contexts. Incorporating the charts into the ALS process helped to overcome difficult technical challenges to setting achievement levels for NAEP.

- ✓ The Reckase Charts proved to be a powerful tool that enabled laypersons to work with item measurement data that otherwise would have been too technical to comprehend.

- ✓ When asked to identify the single feedback information that they would choose to use during the rating process if they could choose only one, 64% of all panelists selected the Reckase Charts.

A concern associated with incorporating the Reckase Charts into the ALS process was that panelists would rely on the chart data to the exclusion of other sources of relevant feedback, possibly deferring their judgment to the statistical data shown on the chart.

- ✓ Although panelists were greatly impressed by the usefulness of the charts and the ease of using them, they indicated that they considered other forms of feedback as well when forming their judgments.
  - There was no evidence of too much reliance on the Reckase Charts.
  - The Reckase Charts did not overly influence panelists when modifying their ratings, to the exclusion of other types of feedback.

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

One persistent challenge to improving the ALS process has been to provide panelists with information about the consistency of their item ratings. This sounds relatively simple, but the issue is *how* to inform them in a way that they can understand and in a way that does not lead to incorrect interpretations.

- ✓ After panelists studied the Reckase Charts, they generally adjusted their ratings to be more similar to the IRT-based estimates of student performance at the cutscores—either their own cutscores or the grade-level cutscores. This finding was consistent for all three achievement levels at all three grades.

- ✓ Panelists could evaluate ratings for different types of writing to determine whether they had judged one type as being more or less difficult than others, relative to their overall cutscores and student performances at that point.

- ✓ None of the judges adjusted his/her ratings to be identical to IRT-based performance estimates.
  - This suggested that panelists considered the achievement levels descriptions and other forms of feedback in addition to the charts when forming their judgment of student performance.
  - Responses to the process evaluation questionnaires supported this interpretation.

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

It was reasoned that if panelists understood the item rating method and the feedback produced by the method, they would adjust their ratings from round to round. If panelists did not adjust their ratings at all, it indicated that they probably did not understand the rating method or the feedback.

- ✓ The ALS process designed by ACT provided panelists with extensive feedback and instructions for interpreting information when forming their judgments of student performance.

- ✓ Panelists were encouraged to reconsider their ratings and adjust them according to their interpretation of the many sources of feedback and information available to them.

- ✓ The Writing Pilot Study panelists exhibited "reasonable" intrajudge consistency across rounds based on the percentage of item ratings changed and the magnitude of change in item ratings.

THE ISSUE OF COGNITIVE COMPLEXITY

ACT has collected considerable data during the Writing Pilot Study and previous research where panelists have reported their capacity to perform the tasks associated with estimating student performance on an item-by-item basis.

- ✓ Judges perceived that they performed the required estimation and judgmental tasks with relative ease.

- ✓ They reported that they were confident in their judgments and satisfied with the results.

- ✓ There is no evidence to indicate that panelists felt unable to make the item-by-item judgments or that they were incapable of estimating probabilities with reasonable accuracy.

THE ISSUE OF PROVIDING CONSEQUENCES DATA WITHIN THE RATING PROCESS

Panelists were given grade-level consequences data after the second round of ratings and individual-level consequences data after the third round of ratings. This was a significant change, and NAGB allowed it for the pilot study with the understanding that it would be implemented in the ALS unless problems were identified.

- ✓ Panelists appeared to be pleased to have this information.

- ✓ They were concerned about the fact that so few students scored at or above the Advanced level.

- ✓ They were not inclined to make major adjustments in their cutscores as a result of the consequences data.

Developing Achievement Levels for the 1998 NAEP in Writing Interim Report: Pilot Study [1]

INTRODUCTION

The achievement levels-setting (ALS) process for the 1998 National Assessment of Educational Progress (NAEP) was implemented as a pilot study prior to the operational ALS. The writing pilot study was an opportunity to continue studying and refining procedures that were introduced in earlier field trials. The field trials were designed to explore new methods for collecting and summarizing judgments used in setting achievement levels for NAEP. The pilot study enabled ACT to determine whether modifications were needed to training, instructing, timing, and other key elements of the ALS process. ACT was particularly interested in evaluating panelists' reactions to incorporating the new Reckase Charts as feedback in the ALS process. Reckase Charts had not been implemented in that manner prior to the pilot study. The writing pilot study was a final check on procedures to assure a successful operational Writing NAEP ALS.

RESEARCH CONDUCTED PRIOR TO THE WRITING PILOT STUDY

ACT carried out two field trials each for the 1998 Writing and Civics NAEP (Loomis, Bay, Yang, & Hanick, 1999; Loomis, Hanick, Bay & Crouse, 2000a and 2000b). All four of those studies were completed and reviewed by ACT's technical advisory committees prior to convening the panels for the 1998 Writing pilot study.[2] It was expected that the data collected during the field trials, based on panelists' reactions to the different methods, would indicate a preferred method for setting NAEP standards that would be further refined through the pilot study. Taken together, the field trials research provided important information about various elements that constitute the standard-setting process designed by ACT.

FIELD TRIAL #1

The purpose of the first writing field trial was to evaluate the Item Score String Estimation (ISSE) rating method relative to the Mean Estimation (ME) rating method used by ACT in the 1994 and 1996 NAEP ALS procedures. To set cutscores on NAEP, ACT has always used an item-by-item rating method requiring judges to estimate the performance of students at the borderline of each achievement level. ACT proposed to study the ISSE method as potentially a new method for collecting item-by-item ratings in the NAEP ALS Process. ACT selected the ISSE method because it appeared to be easy for panelists to understand and use (Impara & Plake, 1997). Further, ACT devised a method for producing item rating consistency feedback data that was analogous to the rating method, so the feedback also appeared to be easy for panelists to understand and use. ACT had conducted computer simulations (Chen, 1998) with the ISSE method with encouraging results. The next step in the research was to evaluate panelists' reactions to the method.

---

[1] This report and the studies in which the report is based were conducted under contract ZA97001001 with the National Assessment Governing Board.

[2] The members of the Technical Advisory Team, ACT's internal advisory group, and the Technical Advisory Committee on Standard Setting (TACSS), the "official" advisory committee, are listed in Appendix A.

Results of the first field trial in writing indicated that panelists were able to use the ISSE method without difficulty. The panelists expressed satisfaction with and confidence in the ISSE method and the outcomes of the process. The procedures were implemented with ease. The ISSE cutpoints and their standard deviations appeared to be reasonable when compared with those produced by ratings of the same prompts with the ME method. The ISSE method resulted in lower cutscores for the Basic level and higher cutscores for the Advanced level than those from the ME method. The standard deviation of the cutpoints for the ISSE method was lower than for the Mean Estimation method. Further research showed the ISSE method to be biased in such a way that cutscores would be higher for the Advanced level and lower for the Basic level when compared with the "true" scores or "true" judgments of the panelist (Reckase & Bay, 1999). Because of this bias, further research using the ISSE method was discontinued, and it was eliminated as an alternative for implementation in the Writing pilot study and subsequent ALS.

FIELD TRIAL #2

The purpose of the second field trial was to identify the procedures that would be used for the 1998 ALS process. ACT's goal was to complete the research phase prior to the pilot study, and the second field trial was the final opportunity to conduct research with panelists before the pilot study. The ISSE method had been eliminated from further consideration, and no final decision had been made regarding the rating method to use in the 1998 ALS process.

In field trial #2, ACT implemented the Booklet Classification method and the new Reckase method (Reckase, 1998) as alternatives to the Mean Estimation method for setting cutscores. ACT compared the two methods. In addition, ACT examined the effect of providing consequences data to judges throughout the ALS process.

Results of the second field trial indicated that panelists had little difficulty with either the Booklet Classification method or the Reckase method. There was no reliable method for computing cutscores for the Booklet Classification method, and there was not enough time to conduct sufficient research to identify a method. Previous research conducted by ACT had consistently pointed to higher cutscores with the Booklet Classification method than with item-by-item methods although the Advanced cutscore for the Booklet Classification method in the field trial was lower than with the Mean Estimation method. Logistic requirements for the Booklet Classification method were known to be great. Finally, ACT research indicated that panelists were likely to use a noncompensatory model for judging student performances with the Booklet Classification method. The decision was made to use the Mean Estimation method with Reckase Charts incorporated as a step in the process for the writing pilot study[3].

There was no statistically significant difference between cutscores set by groups of panelists who were informed of the consequences of their ratings or classifications after the first round of ratings and cutscores set by panelists who were not informed until after the last round of item ratings or classifications. The cutscores of panelists using the Reckase method were higher for panelists who received consequences data throughout the process than for those who received it at the end of the process. The decision was made to provide consequences data after the final round of item-by-item ratings and to allow panelists to make adjustments to the cutscores. The

[3] The decision had also been made to use the Mean Estimation method with Reckase Charts for the civics pilot study. Results of the civics pilot study were evaluated carefully before implementation of the writing pilot study, and adjustments were made as appropriate.

recommendations would be used to produce the final cutscores that would be used for selecting exemplar items and as recommendations to NAGB.

PURPOSE OF THE PILOT STUDY

After reviewing the results of the field trials, it was agreed that research would continue on the use of Reckase Charts for setting NAEP achievement levels. The Reckase Charts were judged to be a promising addition to the ALS process designed by ACT. They appeared to have added substantially to panelists' understanding of the process without a significant increase in the cognitive demand. It was agreed that the charts would be used in the writing pilot study, with the expectation that they would also be used for the writing ALS.

The purpose of the Writing Pilot Study (PS) was to continue studying and refining the procedures for the ALS. Evidence collected in the Civics Pilot Study (Loomis, Hanick & Yang, 2000) led to an adjustment in the design of the Writing Pilot Study regarding the role of the Reckase Charts. In the Civics Pilot Study, Reckase Charts were separated from the other feedback and conceptualized as a distinct step in the rating process. The decision was made to present Reckase Charts as one of several pieces of feedback to consider in preparation for each subsequent round of ratings. Further, ACT developed a program to electronically transfer each panelist's ratings to the Reckase Charts. The method had to be efficient because there were approximately 60 panelists in the pilot study and 90 were expected for the ALS. Thus, Reckase Charts were produced for each panelist to show item ratings for each item at each achievement level for Round 1 and for Round 2. Reckase Charts were not produced for Round 3 ratings.

Because Reckase Charts were now to be incorporated as another form of feedback, TACSS recommended that ACT establish an order for presenting the feedback to maximize the informational value of each piece of feedback and to help panelists identify the feedback by name. Observations by facilitators and statements by panelists suggested that panelists had some difficulty in associating the feedback forms with the names ACT used to identify them. Thus, more emphasis was to be placed on identification, and maintaining a specific ordering of feedback was regarded as an aid in this effort. The feedback was ordered from most holistic to most item specific so that Reckase Charts were the last feedback distributed and evaluated.

A change in consequences data feedback was implemented in the Writing Pilot Study as a result of findings from the Civics Pilot Study. Cutscores in the Civics Pilot Study were found to increase slightly at each level for each grade for each round of ratings. This finding led TACSS to suggest a change in the timing of consequences feedback data. Consequences feedback data would be provided after Round 2 and after Round 3. The feedback after Round 2 would be consequences data related to the grade-level cutscores for each achievement level. The feedback after Round 3 would be consequences data related to each individual cutscore at each achievement level.

Finally, ACT implemented scannable forms for item ratings so that rating data could be entered electronically. This addition of technology was implemented for the first time in the writing pilot study, and ACT needed to study its effectiveness with respect to the logistics of data entry and with respect to item ratings.

The pilot study examined:

- how panelists reacted to the ALS process that included Reckase Charts as feedback;

- whether having item ratings electronically marked on the Reckase Charts appeared to diminish panelists' understanding of the relationship between Reckase Chart data and their item ratings;
- how panelists reacted to using a specified ordering of feedback;
- how panelists reacted to the grade-level consequences data provided after Round 2 and the individual consequences data provided after Round 3;
- how panelists reacted to using scannable forms and the effectiveness of the scanner in the production of feedback data;
- whether further modifications could be identified for the ALS process that included Reckase Charts to make it more successful when implemented for the actual achievement levels-setting study.

The criteria for evaluating the ALS process were measures of the reasonableness of the cutpoints, standard deviations, interjudge consistency, and intrajudge consistency. Also included in the criteria were panelists' comments and responses to process evaluation questionnaires, which reflected their perception of the entire ALS process. Of particular interest were panelists' reactions to the Reckase Charts and logistical issues related to using the charts.

PANELISTS SELECTION PROCESS

The following summary highlights the main features of each step in the process of selecting panelists for the Writing Pilot Study. Please see Appendix B for additional details of the process.

SELECTION OF SCHOOL DISTRICTS

School districts served as the basic sampling unit for the panelist selection process. Principles of sampling were used for drawing stratified random samples of school districts from a national database. ACT drew samples that were proportional to the regional share of districts. The regional proportions were as follows:

- Northeast 20%
- Southeast 20%
- Central 33%
- West 27%

The samples of districts were drawn to include at least 15% with enrollments of 25,000 or more students, and 15% with at least 25% of the population below the poverty level. A total of 258 public districts and 40 private schools were sampled. Please see Table 1 for the distribution of districts and schools sampled. In addition, 15 colleges and universities were sampled from the Higher Education Directory (Rodenhouse & Torregrosa, 1998). Persons in specific positions were identified as nominators in those two- and four-year institutions, both public and private. The total number of districts selected and the proportion in each nominator type were based on previous experience with response rates from nominators in other subjects. Details of the process and the projected number of nominators in each category are provided in the *Design Document* (ACT, 1997).

Table 1
Writing Pilot Study: Distribution of Districts and Private Schools Sampled

Nominator Type	Public Districts	Private Schools	Total
Teacher	15	34	49
Nonteacher Educator	116	6	122
General Public	127	-	127
Total	258 (87%)	40 (13%)	298

NOMINATORS

ALS nominators were identified by drawing three separate samples of districts without replacement.[4] One sample of public school districts was drawn from which nominators of teacher panelists were identified, a second for nominators of nonteacher educators, and a third sample for nominators of general public representatives. Nominators of private school teachers and nonteacher educators were identified from a sample of private schools drawn separately. A total of 748 nominators were contacted. Please see Table 2 for the distribution of nominators. Nominators were persons holding a specific title or position, such as the following.

Nominators of teachers were:
- district superintendents
- leaders of teacher organizations
- state curriculum directors (nominated teachers from throughout the state)
- principals or heads of private schools

Nominators of nonteacher educators were:
- non-classroom educators (e.g., principals, district social studies curriculum coordinators)
- state assessment directors (nominated nonteacher educators from throughout the state)
- deans of colleges and universities (two-year and four-year; public and private)

Nominators of members of the general public were:
- education committee chairpersons of the local Chambers of Commerce
- mayors
- school board presidents
- employers of persons in a writing-related position or with a writing-related background

Table 2
Writing Pilot Study: Distribution of Nominators Contacted

Nominator Type	Public Districts	Private Schools	State	College/ Universities	Employers	Total
Teacher	243	30	42	-	-	315
Nonteacher	13	6	11	29	-	59
General Public	295	-	-	-	79	374
Total	551 (74%)	36 (5%)	53 (7%)	29 (4%)	79 (11%)	748

[4] The districts were sampled from a data file produced by Market Data Retrieval for 1997.

POOL OF NOMINEES

Nominees represented a specific grade perspective (4th, 8th, or 12th) and filled a specific role (teacher, nonteacher educator, or member of the general public). Guidelines were sent to nominators detailing the requirements and criteria. Nominators could submit up to four candidates for each grade whom they judged to be "outstanding" in their writing-related field. From the 748 persons who were contacted to serve as nominators, a total of 88 persons were identified as nominators. They nominated a total of 419 candidates. Please see Appendix B for the distribution of the nominee pool.

CHOOSING PANELISTS

A computerized algorithm was developed to select panelists from the pool of nominees. Nominees were rated according to their qualifications based on information provided on the nomination form (e.g., years of experience, professional honors and awards, degrees earned). Nominees with the highest ratings had the highest probability of being selected, other factors being equal. The selection program was designed to yield panels with:

- 55% of the members representing grade-level classroom teachers
- 15% of the members representing nonteacher educators
- 30% of the members representing the general public
- 20% of the members from diverse minority racial/ethnic groups
- up to 50% of the members male
- 25% of the members representing each of the four NAEP regions

Sixty panelists were required for the panels, 20 for each of the three grade groups. Approximately 30 persons were selected from the nominee pool for each grade and contacted about serving as panelist. Some of the persons who were selected were unable to serve at the scheduled time. ACT was unable to recruit the planned number of panelists, and 57 panelists participated in the pilot study representing 28 states. ACT did not limit the number of panelists who were recommended by the same nominator, which had been done in past studies. ACT attempts to select only one person from a district or school to serve on the grade-level panels. In order to assure the highest quality panels with representation of other important characteristics, this selection criteria was waived, and as many as two nominees were used from the same nominator. The overall representation by region, gender, and race/ethnicity approached the targeted percentages across the three grade groups. The demographic profiles for the nominee pool and the panels have been included in Appendix B.

THE ACHIEVEMENT LEVELS-SETTING PROCESS

The writing pilot study lasted five days, October 1 – 5, 1998 (Thursday-Monday). It was conducted at the St. Louis Ritz-Carlton Hotel. Sessions generally started at 8:30 AM and lasted until 5:00 PM or later. The study employed three grade panels and included all grades assessed by NAEP (4th, 8th, and 12th). The NAEP ALS Project Director served as the primary trainer and general session facilitator for the five-day study. Three content facilitators and three grade group facilitators (one for each grade) assisted the Project Director during the meeting. All facilitators

had been trained before the writing pilot study panels were convened and were experienced in the procedures used for the study.[5]

SESSION FORMATS AND FACILITATION

The Project Director presented all training and instructions in general sessions so that every panelist had the same instructions and the same information regarding tasks, purposes, and procedures. Following each general session, panelists broke into grade-level sessions where they were trained using group discussions, exercises, practice ratings, and so forth. All procedures, except producing final cutscore recommendations, were implemented in grade-level sessions. The Project Director presented a general overview of the process that included graphics and flow charts to illustrate the process, as well as a step-by-step summary of the procedure to be followed. Information regarding the tasks to be accomplished and the methods by which they would be accomplished was provided to panelists at the start of each day during general sessions.

A grade-level process facilitator and a content facilitator led each grade-level panel. Process facilitators took the lead in implementing training exercises and answering "process" questions. Process facilitators received approximately 40 hours of training prior to the pilot study. Content facilitators led the discussions of the *1998 Writing NAEP Framework* and achievement levels descriptions, and answered "content" questions. All content facilitators had participated in developing the Writing NAEP. They participated in a full-day, joint training session with the content facilitators led by the Project Director before the pilot study.

Each morning before the session started, the facilitators met to review activities for the day and to coordinate plans for implementing tasks. Any problems or issues were discussed and resolved. Facilitators generally reviewed all process evaluation questionnaires to determine whether any panelists were having problems or needing additional help with specific aspects of the process.

To ensure that grade-level facilitators provided uniform instructions, they followed a highly detailed outline of the achievement levels-setting process. The outline provided instructions for each activity in each grade-level session. In addition, instructions were displayed on overhead transparencies for panelists to follow during each part of the procedure. The meeting agenda and the facilitators' outlines have been included in Appendix C.

ITEM RATING GROUPS AND TABLE DISCUSSION GROUPS

Within each grade group, panelists were divided into two different item rating groups of 9 or 10 persons: group A and group B. These groups provided a means of monitoring the ALS process by evaluating the similarity of ratings of both groups at different stages of the process. Each rating group was further divided into 2 discussion groups of 4 or 5 persons per table for each grade group. The demographic attributes of panelists were considered when assigning members to the item rating groups and to table groups; otherwise, the assignments were random. The goal was to have groups as equal as possible with respect to panelist type, gender, region, and race/ethnicity. The demographic profiles for the item rating groups and the table discussion groups have been included in Appendix B.

[5] A list of ALS staff and observers has been presented in Appendix A. Note that for the writing pilot study, two content experts participated in facilitation of grade 4 sessions. Dr. Catherine Welch served as the content facilitator through the first few days when panelists internalize the achievement levels descriptions and form a clear concept of borderline performance. Dr. Patricia Porter was on site to answer content questions of panelists during the rating process and to gain experience in the overall process.

ITEM RATING POOLS

The 1998 NAEP Writing data were used for the pilot study. Two item rating pools for each grade were constructed so that they were as nearly equal as possible with respect to difficulty and prompt type. Detailed information about the item pools has been presented in Appendix D. The design included two item rating groups and two item rating pools which provided the opportunity to examine ratings from each item rating group as a replication of the other item rating group for each grade.

The grade 4 Writing NAEP consisted of 20 prompts divided into 20 blocks. Each block contained 1 prompt. Each rating group rated 12 blocks of items (12 prompts): 5 narrative prompts, 4 informative, and 3 persuasive. Eight prompts in each rating pool were unique to each rating group, and four prompts were common to both rating groups for grade 4. Group A rated prompts from 12 of the 20 blocks; group B rated items from 12 of the 20 blocks; and both groups rated the same prompts for 4 of the 12 blocks.

The item rating pools for grades 8 and 12 were divided in the same way as those for grade 4, with one exception. Although all groups rated 12 blocks of items, grade 8 rated 4 narrative prompts, 4 informative, and 4 persuasive, and grade 12 rated 3 narrative prompts, 4 informative, and 5 persuasive.

STEP 1: BRIEFING MATERIALS

Before panelists arrived in St. Louis, they were mailed materials that contained important background information on setting achievement levels. (See Appendix E.) The first advance packet was mailed August 10, 1998, and contained materials that panelists were required to study. The second mailing was September 24, 1998 and contained detailed instructions related to travel arrangements and accommodations. The briefing materials and information included:

- 1998 NAEP *Writing Framework*;
- 1998 NAEP Writing Achievement Levels Descriptions;
- *Briefing Booklet* for 1998 Writing NAEP;
- *Multiple Challenges*, a booklet about the 1998 NAEP;
- NAGB brochure;
- *The NAEP Guide*;
- Cover letters with instructions for preparing for the study;
- Assessment Item-Use and Nondisclosure Agreement;
- Check Request Form;
- Request for Taxpayer I.D. Number and Certification;
- Information about St. Louis;
- Map and directions to the meeting.

STEP 2: GENERAL ORIENTATION AND TRAINING EXERCISES

In the opening session, panelists were given an orientation to the achievement levels-setting process and a complete overview of the procedures planned for the pilot study. The overview was presented with the aid of a computer presentation program that provided animated graphics as examples or demonstrations of key aspects of the process. During the orientation session, a member of the NAGB staff presented a history of NAEP, a general overview of the NAEP

program, a description of the method used to develop the *1998 Writing NAEP Framework*, and other such general information about NAEP and NAGB.

At the start of the second full day, a different computer presentation was used as an aid in the session designed to reinforce the information presented in the initial orientation to the process. The second presentation focused on outcomes of the process and how they are produced. The second computer show was presented for the first time at the Writing Pilot Study, and it seemed to be a positive addition to the process. An animated computer presentation of the *Top Ten Misconceptions about the NAEP ALS Process* was also shared with panelists at the end of the first general session on day 2. That presentation had been successful in previous ALS processes as a means of addressing panelists' concerns and anxiety about the complexity of the process. It seemed to be a hit with the writing panelists too.[6]

Panelists were urged to use their *Briefing Booklet* as an instructional tool and as a review guide for each session. The *Briefing Booklet* included a sketch of each activity in each session, in the order that it occurred in the agenda. It described the purpose of the activity and how it was to be accomplished. A copy of the meeting agenda is in Appendix C.

The process includes several opportunities for panelists to receive instructions in each element of the procedure. By design, all instructions and training are first provided in a general session so that each person hears the same information. Grade group facilitators then implement the training exercises and ALS procedures using the same instructions. Each day, a list of tasks that panelists must accomplish for the day is presented in the general session, along with information about the purpose(s) of the activities and instructions on how the tasks will be accomplished. These lists are again presented in the grade group sessions to help panelists stay focused and to help identify the activities for the panelists.

Facilitators are given an outline to follow. The outline is shared with panel members in each grade so that they can refer to the steps in the outline while performing exercises and tasks. These procedures, along with the *Briefing Booklet* for panelists, make it relatively easy for panelists to identify each element in the process and to understand how each one fits into the overall ALS process.

TAKING A FORM OF THE NAEP

Following the general orientation session on the first day, panelists went to their assigned grade groups where they took a form of the NAEP developed for their grade group.[7] After completing the assessments, they reviewed their own responses relative to the scoring guides. Two forms of the assessment were administered to the panelists for each grade. Item blocks in the form administered to rating group A were excluded from their item rating pool, and the same was true for the blocks in the form administered to rating group B.

---

[6] Panelists engaged in a social mixer on the first evening. A grid with clues was distributed to each panelist and they were to identify a person who matched each clue. A book was presented to the winner, drawn from random among those who had correctly completed the form by Day 2. The mixer was distributed to panelists as a part of a social hour just before dinner on the evening of the first day. This activity was implemented as a result of suggestions from civics pilot study panelists, and it seemed to help panelists in the general orientation process.

[7] The NAEP forms administered to panelists were later used as Whole Booklet Feedback and The Whole Booklet Exercise, which are described in "Step 3: The Item Rating Process and Feedback."

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS

During a general session the morning of Day 2, the three content facilitators presented an overview of the *NAEP Writing Framework* and the ALDs as a general session.[8] Panelists had been instructed to read the Framework and to study the achievement levels descriptions prior to the meeting. To reinforce this learning, the general session presentation provided a clear, comprehensive account of the content and organization of the *NAEP Writing Framework* and a clear explanation of how the ALDs were related to both the Framework and to the NAGB policy definitions.

In grade-level sessions, content facilitators guided the panelists through an extensive training session focused specifically on the achievement levels descriptions for their grade. Panelists were led in an evaluation of the ALDs to compare performance across levels in their grade and to compare performance across each grade within each level. Panelists discussed the ALDs and participated in several training exercises to help them understand the descriptions. They were lead through an explanation of the scoring guide and the description for each score point in the generic rubric for each type of prompt: narrative, informative, and persuasive. This exercise was designed to help panelists become familiar with prompts of different types and to understand how the ALDs relate to all types of prompts.

Next, panelists had the opportunity to apply their understanding of the ALDs with samples of student papers. A sample of ten student papers was given to judges to review and discuss with respect to their understanding of the ALDs. The particular writing exercise was the first one in the NAEP booklet form that was administered to the panelists on the first day of the ALS session. Panelists were asked to determine if the performance exhibited in the paper should be classified as Basic, Proficient or Advanced. After classifying each paper independently, panelists discussed their classifications with each other.

Panelists then looked at whole booklets produced by the same students who had written the first paper used in the previous exercise. Panelists confronted potential differences in performance by the same student when reviewing the two writing tasks presented in the NAEP booklet. Panelists classified the booklet performance into the four[9] categories based on the ALDs. Their paper classifications from the previous exercise had been collected and were not available to them when they classified the complete student booklets. Following this task, panelists received their earlier classifications of single papers to compare with their classifications of performance on both writing exercises. They discussed the differences between performance on a single writing task and performance on a whole booklet that required writing for two prompts. This exercise helped panelists gain a better understanding of the ALDs and become familiar with additional NAEP items and scoring rubrics. In addition, it helped panelists understand that students do not perform at the same level on two different writing exercises. Some panelists were quite disturbed by the unevenness of student performance on the two writing tasks. They were also disturbed to learn that students who performed poorly on one task and well on the other could score relatively well on the booklet as a whole. Please see Appendix F for a sample classification form.

[8] The achievement levels descriptions were evaluated thoroughly and modified extensively through a national review process involving focus groups and expert review panels. The final version of the writing ALDs that resulted from this process were approved by NAGB for use in this pilot study (Loomis & Hanick, 2000).

[9] The four categories were below Basic, Basic, Proficient, and Advanced.

UNDERSTANDING BORDERLINE PERFORMANCE

After working with the ALDs throughout the morning and early afternoon, panelists were trained in the concept of borderline performance and instructed in rating prompts to reflect performance at the borderline. Borderline performance refers to the level of performance that is minimally acceptable for each achievement level. Panelists are trained to understand that performance that minimally qualifies for a level is within that level. Thus, the borderline concept used in rating items refers to the lower boundary of each achievement level. As part of the training in borderline performance, the general session included a demonstration of how items would be rated. This training exercise was designed to help panelists understand the importance both of forming a clear understanding of borderline performance and of having the same understanding shared among panelists. In the pilot study for civics, panelists participated in a "Preview Rating Session." Observations by staff and feedback from panelists in the debriefing session indicated that the Preview Rating Session should be dropped. The decision was made to replace the Preview Rating Session with a training demonstration in rating items. The goal was to give panelists an understanding of why it was important to have a clear, well-formed concept of borderline performance, and it seemed possible to accomplish this purpose through a demonstration of rating procedures.

Following the demonstration in the general session, panelists returned to their grade groups to begin working on borderline descriptions. Panelists wrote descriptions of student performance at the borderline of each achievement level. Developing descriptors of borderline performance assisted panelists in forming a common understanding of the ALDs as well as a common understanding of borderline performance. Each grade group had drafted a set of borderline descriptions by the close of Day 2. Content facilitators evaluated those sets across all grades to make certain that they represented "reasonable" descriptions of borderline performance—not too low and not too high.

The borderline descriptions were distributed to panelists at the start of Day 3 for review and modification. They were aware that the first round of item ratings would begin later that day, and the goal was to make certain that they had a useful set of descriptions to use in the rating process. A review of borderline descriptions was scheduled just prior to the training session for Round 1 ratings. As a means of keeping panelists focused on the ALDs, they evaluated borderline descriptions throughout the process. Borderline descriptions were evaluated and modified, as needed, until panelists were ready to begin the final round of item-by-item ratings on the last day. Please see Appendix C for the final versions of the borderline descriptions for each achievement level and each grade.

PAPER CLASSIFICATION EXERCISE

Instructions in the Paper Classification Exercise followed the review and discussion of borderline performance on the morning of Day 3. The purpose of the exercise was to have panelists evaluate student papers and determine whether any could be found to represent borderline performance. The Paper Classification Exercise required panelists to examine three student papers scored at each of the six rubric score points for a total of 18 papers for each prompt. The papers represented each of the three writing prompts that were common to all panelists in the grade group.

Panelists were instructed to first sort papers into four categories: below Basic, Basic, Proficient, and Advanced. They then were to look at the papers within each category and determine whether or not any of those represented performance at the borderline of the level. If no paper represented borderline performance, then no paper was classified as borderline performance. For each of the

three prompts, panelists classified each paper, discussed their classifications as a group, and modified their classifications if they chose to do so. Panelists classified and discussed 54 student papers and 7 achievement categories[10]. After selecting a paper to represent borderline performance at each achievement level, panelists could refer to a sheet where scores for each paper were recorded. The basis for selection, however, was to be their understanding of the ALDs and borderline performance, not paper scores. The discussion of paper classifications provided some feedback for panelists to use in the process of selecting exemplar performances on Day 5. Further, their classifications were recorded on a form so that the numbers of persons classifying each paper at each of the levels and borderlines could be used later as feedback in the exemplar performance selection process.

This training activity was designed to accomplish the following purposes:

- to provide a reality check on how students responded to the prompts;
- to promote a clear conceptualization of performance at the borderline;
- to familiarize panelists with the scoring rubrics of prompts.

STEP 3: THE ITEM RATING PROCESS AND FEEDBACK

The general procedure followed for the item rating process included instruction in a general session involving all panelists to assure that they were given the same information. Process facilitators reviewed the instructions and answered questions from panelists in the grade-level sessions. Three papers scored at each rubric score point were included in packets for all prompts in each panelist's rating pool. Papers written for prompts that were not in the paper classification exercise had the score written on the first page. The papers were to serve as a reference for panelists during the rating process so they could see examples of student performances at each rubric point. The rating tasks were performed by panelists in grade-level sessions. Similarly, feedback information was first presented in a general session where panelists learned what it was and how to use it. All feedback for the first two rounds was distributed to panelists for review and discussion in their grade groups.

ROUND 1 RATINGS

Following the Paper Selection training exercise on Day 3, all panelists participated in a general session that involved instruction in the item-by-item rating process. The Mean Estimation method (ME), which is a form of the modified-Angoff rating method, was used. The rating method had been described in the orientation sessions of the first two days, and a demonstration had been given on Day 2. The procedure was reviewed in detail, and panelists were instructed in marking their rating forms. For rating the prompts, judges estimated the mean or average score (e.g., 2.4 on a scale of 1-6) of students performing at the borderline of each level. They were told to think of a class with 100 borderline students for each achievement level and estimate the average score for those students on each prompt. Once trained, the panelists were ready for Round 1 rating. During the rating process, panelists could refer to student papers that served as examples of student performance at each rubric point. The references included three papers scored at each rubric score point for all prompts in each panelist's rating pool. Scores appeared on the papers for prompts that had been omitted from the paper classification exercise.

Panelists were told to read each item carefully, compose a mental response to the item, and refer to the scoring rubric. This procedure would help panelists form a clear concept of what was

[10] The seven categories were: Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient, Borderline Advanced and Advanced.

required of students. For each item in their item rating pool, panelists marked their estimate of borderline performance at each of the three achievement levels. Panelists were not allowed to discuss item ratings with each other. They were encouraged to refer to the achievement levels descriptions and descriptions of borderline performance. Some panelists completed the Round 1 ratings in less than one hour, and no panelist needed more than about two hours to complete the task. A copy of a rating form has been included in Appendix C. Before leaving, panelists completed a process evaluation questionnaire. They were also asked to be available until their item ratings had been entered into a data file and checked. This concluded Day 3.

FEEDBACK AFTER ROUND 1

Staff scanned the rating data into electronic files on site and verified the accuracy of the data. Feedback data were produced and ready for distribution to panelists at the start of Day 2. In a general group session, panelists were given instructions in the use of feedback data resulting from their first round of ratings. The cutpoints for each grade-level were presented in the general session for all panelists to see. Instructions in feedback included an explanation of feedback forms and information about the source of the feedback data, how to interpret the data, and how to use the data to modify ratings to raise or lower cutscores. These forms of feedback have been described in the *Briefing Booklet*, and defined for the NAEP ALS context only. Copies of the feedback based on Round 1 ratings have been included as Appendix G.

Cutpoints

The cutpoints are computed from the combined ratings of all raters and all prompts for each achievement level for each grade. Cutpoints are computed for each grade-level across ratings by panelists in the two rating groups, Group A and Group B. The cutpoints are presented on the ACT NAEP-like scale which is a linear transformation of the NAEP score scale. This transformation decreases the potential for achievement level data from other NAEP subjects to influence panelists in the writing pilot study. Item parameters produced by an IRT model are used in computing the cutscores. (See Chen & Loomis, 2000 for a description of computational procedures.)

Standard Deviation

The standard deviation is the indicator of the level of variability around each cutscore for a grade. The cutscores are computed as the mean score over all items and raters within a grade. The standard deviations reported to panelists on graphs with their cutscores are computed as the variability of the individual raters' cutscores with respect to the grade level cutscore. (See rater location data below.)

Whole Booklet Feedback

Whole booklet feedback is produced for the set of prompts in the NAEP exam booklet that were administered to panelists as part of the orientation process on Day 1. Each rating group (A and B) had a different assessment form. The whole booklet feedback reports the percent of total possible points that a student needs to earn in an assessment booklet in order to meet the minimal requirements for performance at the cutscore of each achievement level. For example, the whole booklet feedback report might state: "Based on the cutscore for your grade, students performing at the borderline Advanced level are expected to get 92% of the total possible score points for this booklet." A similar statement is given for each achievement level. This feedback is based on the cutpoints the grade group had set during the first round of ratings, and is updated after subsequent

rounds of ratings. Panelists are informed of the reasons that would cause the percentages to differ for the two booklet forms used by Group A and B, i.e., different prompt combinations resulting in different student performance and total points possible.

Whole Booklet Exercise

As part of Round 1 feedback, the panelists participated in a whole booklet exercise, which was an extension of providing whole booklet feedback. They were shown actual student booklets with scores near the cutpoints that had been set by Round 1 ratings. The booklets were the same form used for the training exercise "Taking a Form of the NAEP." Booklets scored within 4% above or below the total possible points associated with each cutpoint were evaluated by panelists. Panelists might be shown a booklet representing borderline Basic performance that earned 49% of the total possible points, if that corresponded to performance at the Basic cutscore ±4% points. Furthermore, the combination of prompt scores was considered in selecting the booklets. For example, three booklets were selected to represent borderline Proficient level. All three booklets were scored as 8 points. The score combinations varied, however. One booklet was scored 5 on the first prompt and 3 on the second, another booklet was scored 4 on both prompts, and the final booklet was scored 3 on the first prompt and 5 on the second.

Panelists were asked to examine the responses of the student to both prompts in the booklet as a whole and determine if the responses represented student performance expected at the lower borderline of Basic, for example. If they perceived a discrepancy between the expected performance and the observed performance in the booklets scored at the cutpoint, they discussed the achievement levels descriptions and borderline performances again with other panelists to try to understand the cause for this discrepancy. Performance higher than expected would signal that they had set their cutpoints too high. Performance lower than expected would signal that they had set their cutpoints too low.

Panelists are given up to 4 booklets to review as representative of borderline performance at each achievement level. One hundred booklets for each of the six NAEP forms (one for each rating group) are randomly selected for this exercise. Panelists review photocopies of student responses, but no student background data are shared. Because the booklets are randomly selected, there is a fairly high probability that none will be available to represent performance at some cutscore(s). In cases for which no booklets are available that have a score within 4% of the total possible points associated with the cutscore, no booklets are presented to panelists for that achievement level. Panelists are given a complete explanation of the source of booklets and the reason for which no booklet are available.

Rater Location Feedback Charts

The rater location feedback charts are histograms representing the distributions of panelists' cutscores. The horizontal axis represents scores on the ACT NAEP-like scale, and the vertical axis represents the number of raters. Letter codes that identify individual raters are positioned along the ACT NAEP-like scale at the point where each panelist sets his/her cutscores based on his/her individual ratings. Letter codes are used so the cutscores for each panelist may remain confidential. (In fact, most panelists openly and freely discussed their rater location data.) The graphs indicate the cutscores that result from the item ratings by each panelist for Basic, Proficient, and Advanced levels, and the relationship of the panelists' ratings to each other (interjudge consistency). One chart is produced to display rater locations for each of the three achievement levels within each grade.

If the cutscores for panelists in a grade are scattered across a relatively wide score range, this indicates a low level of interrater consistency which probably resulted from a lack of agreement on the meaning of borderline performance. The rater location charts show in detail the information reported as the standard deviation of each cutscore. Panelists are informed of this relationship between rater location charts and standard deviations, and they are informed that low interrater consistency/a high standard deviation is an indication of the need to discuss their understanding of borderline performance for the achievement level(s).

Facilitators examine the patterns of cutscores on the charts to identify panelists who are "outliers" or panelists who could be experiencing problems with the item rating process. For example, facilitators check for panelists who tend to set very high or very low cutscores for all levels relative to other panelists in the grade group. They also check for panelists who set cutscores that are very close together or very far apart. Facilitators make a specific point of discussing these findings with the panelists to make certain they understand the implications of their cutscore patterns and how to change them through subsequent ratings, if the panelist so desires.

Student Performance Data

Panelists receive information about overall student performance on each prompt. The mean (average) score is reported for each prompt, along with the percentage of student responses scored at each rubric score point. The data also report various categories of "no response" for each item. Student performance data serve as a reality check because they show how students actually perform on each item. The data indicate how easy or difficult the prompts are for all students who took the 1998 Writing NAEP. They do not indicate how easy or difficult the prompts are for students at different achievement levels.

Reckase Charts

The meeting director introduced the Reckase Chart to panelists by using a special computer presentation. The computerized show displayed only one column of the chart followed by only one row of the chart. This instructional design was selected to avoid "number shock" for the panelists. The entire chart was displayed after the data in columns and rows had been explained. The computer show allowed the presenter to slide up and down the chart to reveal different areas. It also provided the capability to "zoom in" during instruction to highlight specific features, such as markings to distinguish ratings for the three achievement levels. The demonstration used colored markings to indicate individual cutscores, grade-level cutscores, and item ratings on the chart. The presentation aided panelists to evaluate their own charts and interpret the information displayed on the charts. Each panelist received a printed copy of the Reckase Chart that presented information for each prompt in the rating pool and was marked with his or her own Round 1 ratings.

Panelists are given a Reckase Chart that indicates expected performance for students scoring at each score point on the ACT NAEP-like scale for each prompt in the item rating pool. Each column represents the range of IRT-based performance estimates for one assessment item. Each row represents IRT-based performance estimates for the prompts for students scoring a specific point on the ACT NAEP-like scale. The ACT NAEP-like scale scores range from the score associated with the lowest asymptote value for any prompt in the grade-level item pool to the value associated with the highest asymptote. The expected performance across scale score points can be observed for each prompt, as can the expected performance across prompts for students scoring at a particular scale score. A sample Reckase Chart and instructions have been included in

Appendix H. Please note that only data for odd-numbered scale scores are reported on the charts in order to save space and fit the necessary data on the 11"x17" charts.

Panelists mark their charts with both the grade-level cutscore and their own cutscore for each achievement level. Panelists' individual item ratings are marked by computer on the Reckase Charts. Panelists draw a line to connect one item rating to the next for all ratings at each achievement level. They use three colored markers to distinguish the three achievement levels. All of the prompts for one grade are printed on one chart page.

By examining the charts, the panelists are able to consider the relationship between their estimates of student performance for each item and the IRT-based expected student performance at the cutscores. Further, panelists can consider any observable patterns in their ratings, such as differences in ratings for prompts requiring different types of writing (narrative, informative, or persuasive). They can also look for indicators of rater fatigue, such as less consistent ratings for prompts toward the end of the rating pool. Panelists are informed that if their judgments of students performing at the borderline of each achievement level exactly fit the estimates generated by a statistical model based on actual student performance, all of their ratings would fall along a single row. In other words, if panelists' ratings are on a single row, their ratings perfectly match IRT-based estimates of student performance.

ROUND 2 RATINGS

Panelists studied and discussed the feedback information from Round 1. To prepare for Round 2, they reviewed the ALDs and modified the borderline descriptions as needed. Panelists rated the same items a second time using the same rating method. They could change all, some or none of their ratings for any or all achievement levels. As is typically the case, Round 2 ratings on Day 4 took less time than Round 1 ratings. Panelists again completed a process evaluation questionnaire. Item ratings were again scanned into data files for computations and analyses, and staff verified data entry on site. Feedback data, based on Round 2 ratings, were produced for distribution on the following day.

FEEDBACK AFTER ROUND 2

Day 5, the last day, was a busy day. Feedback information was presented in a general session where cutscores and standard deviations for each grade were shared. Feedback information was reviewed before panelists returned to their grade level panels. The same types of feedback were provided to panelists after Round 2 as were distributed after Round 1. Consequences data were added and the whole booklet exercise was omitted, however.

Panelists were trained in the consequences data in the general session. All panelists were informed about the consequences of each cutscore at each achievement level at each grade. Data were reported as the percentages of students scoring at or above each achievement level. The data were reported as bar graphs clearly showing the percentages at or above Advanced and Proficient that are included in the percentage at or above Basic, for example. In addition to the bar graphs, pie charts represented the percentages of students scoring within each achievement level and the percentage scoring below the Basic level.

Feedback was again distributed to panelists in grade groups where they could ask questions and discuss the results. Panelists again marked their cutscores on the Reckase Charts and connected their ratings for each item at each achievement level to examine the consistency of their ratings. Panelists had time to review the feedback data, ask questions, and discuss concerns before

beginning the third round of ratings. They also had the opportunity to review the ALDs and modify and borderline descriptions prior to the Round 3 ratings. (Please see Appendix G for feedback information based on the second round of ratings.)

ROUND 3 RATINGS

Panelists rated the same items a third time using the same methodology. They could change all, some or none of their ratings for items at any or all achievement levels. For this final round of item ratings, panelists were allowed to discuss ratings for specific items with other panelist in their table group. Round 3 ratings were completed before noon on Day 5.

FEEDBACK AFTER ROUND 3

Round 3 item ratings were again scanned into data files for computations and analyses. Feedback data were produced for panelists, based on Round 3 ratings. Reckase Charts were not marked for Round 3 ratings.

The graphical representations of consequences data presented as feedback after Round 2 were updated and presented again after Round 3. In addition, consequences data were presented in two new formats. Rater location charts were modified to include the percentages of students who scored at or above score points, reported in increments of 5 points on the ACT NAEP-like scale. This provided panelists with fairly detailed information about the distribution of student performances relative to the reporting score scale. The other format was individual consequences data that were listed for each panelist in each grade group. The list contained panelists' secret ID codes, their cutscores on the ACT NAEP-like scale for each achievement level, and the percentages of students performing at or above the individual panelists' cutscores. Together, these different ways of presenting consequences data provided panelists with a large amount of rather specific information they could use to make recommendations for their final cutpoints.

Cutscores, standard deviations, and consequences data for each grade were presented in general session. Paper copies were distributed to each panelist including individual level consequences data. Panelists also received updated whole booklet feedback and rater location charts based on Round 3 ratings. Round 3 feedback data were distributed in the general session. Panelists were seated by grade group and arranged by their identification number. This seating plan was announced to panelists in advance. The plan was designed to facilitate distribution of feedback data and collection of recommendations for final cutpoints.

STEP 4: MAKE RECOMMENDATIONS FOR FINAL CUTPOINTS

Panelists were given a few minutes to review the consequences data before they received a consequences data questionnaire. A sample questionnaire has been included in Appendix I. The questionnaire items asked whether panelists would want to make changes to any of the cutscores after learning the consequences of their cutscores. The relationship between cutscores and consequences data was made clear, i.e., raising cutscores lowered percentages of students performing at or above the cutscores. Panelists could recommend a different cutscore to represent each achievement level for any or all three cutscores. The individual Round 3 cutscores were used to compute the final grade-level cutscores for panel members who recommended no changes to their cutpoints. Panelists were fully informed that these would be the final cutpoints, and they would be used as the standard for selecting exemplar items.

STEP 5: SELECTION OF EXEMPLAR PERFORMANCES

After the panelists recommended their final cutpoints, they were trained in the selection of exemplar performances for each achievement level. The final cutpoints were computed and, based on these new cutpoints, lists of exemplar papers were prepared for review and selection by panelists. Panelists also received feedback from the Paper Classification Exercise on Day 3 when they classified 54 student papers into 7 achievement categories. They reviewed the frequencies of panelists' classifications for three prompts, one of each type of writing.

Panelists in each grade group selected student performances that they considered appropriate to illustrate knowledge and skills associated with the description of each achievement level. The exemplar performances were selected by panelists to use in reporting the NAEP results and were a primary outcome of the ALS process. The exemplar performances lists were drawn from prompts that had been marked for release to the public when the results of the 1998 Writing NAEP were reported. The goal of the exemplar selection process was to provide papers to illustrate student performance for each type of writing at each NAEP achievement level.

The statistical selection of exemplars for review by panelists was based on item difficulty. The average conditional probability of correct response served as the indicator of item difficulty. To be on the list of exemplars, the score for a prompt had at least a 50% average probability of correct response across the score range of an achievement level. Each rubric score point was evaluated as if it were an item, so prompts could appear on the list six times (once for each credited response). Papers were "assigned" to the list of items at the lowest achievement level for which this criterion was met. If the criterion were met at a score below the Basic cutpoint, the item was not listed because the performance could not represent Basic achievement.

Panelists determined whether or not each paper that qualified as an exemplar would serve as a good illustration of performance required at the specific achievement level, based on the achievement levels descriptions. They identified papers that matched the descriptions of student performance at each achievement level and satisfied the statistical criterion that qualified the paper as an exemplar. They "approved" or "vetoed" each paper. The lists of exemplar performances for each grade have been included in Appendix J. Also displayed with the lists are the average conditional probabilities.

STEP 6: EVALUATIONS THROUGHOUT THE PROCESS

Panelists completed seven process evaluation questionnaires throughout the five-day meeting. The questionnaires were distributed at the conclusion of each stage of the process, usually at the end of each day.

FEEDBACK AFTER RECOMMENDATIONS FOR FINAL CUTPOINTS

The final grade group cutscores, based on panelists' recommendations, were used to compute the final consequences data. These final consequences data were presented to panelists in a general session after all grade groups had completed the process of selecting exemplar items. Panelists were given a few minutes to consider the final consequences data.

After reviewing the final cutscores and grade-level consequences data, each panelist was again asked to respond to a questionnaire regarding the consequences data and the final cutscores he/she would recommend to NAGB. Panelists were aware that their responses were only recommendations and that no changes would be made in cutscores on the basis of those

recommendations. The stated purpose of collecting their recommendations was to inform NAGB of panelists' opinions regarding the final cutpoints and the consequences associated with them.[11] When the panelists completed the final questionnaire, they were thanked for their work and the meeting was adjourned.

OUTCOMES OF THE WRITING PILOT STUDY

The writing pilot study for the 1998 ALS process was planned as a "dry run" for the operational ALS to determine whether modifications to the process were needed. The writing pilot study was an opportunity to continue studying and refining the incorporation of Reckase Charts into the NAEP ALS process. Throughout the pilot study, ACT collected information about the reactions of panelists to the ALS process and the Reckase Charts. Their suggestions were considered when adjusting the process to assure smooth implementation of the methodology when used for the operational ALS meeting. In addition to pilot study panelists' comments, the criteria for evaluating the ALS process were measures of the reasonableness of the cutpoints, standard deviations, and intrajudge consistency resulting from implementing the method.

EVALUATION OF THE CUTPOINTS AND THEIR STANDARD DEVIATIONS

The cutscores and their standard deviations have been included in Table 4. For all grades and levels, the cutscores *increased* from round to round, while the standard deviations *decreased* for each round of ratings. It has been common for the standard deviation to decrease from round to round, so this positive outcome was expected. The cutscores, however, showed an uncommon pattern of increasing across each round at each level for each grade. This pattern was also observed for the civics pilot study, for which the same process was implemented. These results, however, have not been observed for data generated from previous NEAP ALS studies.

Grade 8 judges had rather large variability in their Basic ratings, particularly for Rounds 1 and 2 and in general were more variable than panelists for grades 4 and 12. The range of cutscores as displayed on the rater location charts provides additional evidence of variability for grade 8 panelists. By Round 3, the range of cutscores for Basic was smaller for grade 4 and grade 12 panels (4^{th} grade = 139 – 151; 12^{th} grade = 133 – 142) than for grade 8 (135 – 165).

No patterns of statistically significant differences appeared when comparing cutscores by panelist type, grade, round of ratings, gender, region, or race/ethnicity. Occasionally, a statistically significant difference was found. For example, the Round 1 Proficient cutscore for grade 12 teacher panelists was significantly lower than that for general public panelists. When comparing cutscores within grade by rating groups (groups A and B) and table groups, no major differences were noted. Given the very large number of comparisons, the surprising finding was that so very few were statistically different. For a detailed report of the test results for group differences, please refer to Appendix K.

[11] ACT presented these recommendations to TACSS for review and evaluation, as well as to NAGB.

Table 4
1998 Writing NAEP Pilot Study Outcomes:
ACT NAEP-Like Scale Score Cutpoints, Standard Deviations, and
Percentages of Students Who Scored At or Above Each Achievement Level

Grade	Achievement Level	Data	Round 1	Round 2	Round 3	Final
4	Basic	Cutpoint	141.6	144.5	145.0	145.3
		SD	5.2	4.0	3.3	2.6
		%≥	**83.0**	**77.2**	76.3	75.6
	Proficient	Cutpoint	165.1	167.5	168.0	167.1
		SD	9.2	4.0	3.9	3.0
		%≥	**23.7**	**18.5**	17.7	19.4
	Advanced	Cutpoint	186.7	189.3	189.1	186.6
		SD	8.0	3.5	4.2	3.1
		%≥	**1.2**	**0.7**	0.7	1.2
8	Basic	Cutpoint	140.2	145.6	149.8	151.2
		SD	10.1	9.9	7.5	5.0
		%≥	**85.4**	**75.0**	64.6	60.4
	Proficient	Cutpoint	165.0	171.0	172.3	170.9
		SD	8.4	5.9	5.2	4.3
		%≥	**23.7**	**12.6**	10.8	12.8
	Advanced	Cutpoint	186.1	189.5	190.7	188.6
		SD	4.7	6.5	5.2	4.9
		%≥	**1.3**	**0.7**	0.5	0.8
12	Basic	Cutpoint	135.9	137.0	137.5	138.3
		SD	4.6	3.5	2.7	2.1
		%≥	**91.4**	**90.1**	89.3	88.4
	Proficient	Cutpoint	156.1	159.1	157.9	158.9
		SD	5.5	3.5	5.2	2.3
		%≥	**47.0**	**38.4**	41.9	38.8
	Advanced	Cutpoint	179.9	182.2	182.8	181.7
		SD	8.4	4.3	4.0	3.7
		%≥	**3.8**	**2.4**	2.4	2.8

Bold font represents data that were not presented to panelists.

A SPECIAL CASE

A general public, grade 4 panelist had many errors on his Round 1 rating form. All three of his cutscores were nearly identical. His Round 1 ratings were excluded from computations of the grade-level cutscores, and the rater location charts for feedback after Round 1. The meeting facilitator worked intensely, one-on-one with this panelist while others were evaluating their feedback data. The panelist discussed Round 2 ratings with the facilitator during the rating process. The two discussed the panelist's Round 2 results, which showed his ratings to be near the grade-level cutpoints for the three achievement levels. The panelist seemed satisfied with the rating procedures and felt capable of making logical and reasonable judgments. His Round 2 ratings were included in the computations for feedback after Round 2 and thereafter for the remainder of the meeting.

EVALUATION OF INTRAJUDGE CONSISTENCY

Intrajudge consistency, both within rounds and across rounds, is generally regarded to be a reasonable criterion by which to judge a standard setting process. Indicators of intrajudge

consistency include both the magnitude of change in item ratings from round to round, and the number of item ratings changed from round to round. ACT examines these indicators as part of the data analyses *after* an ALS process has been completed. These comparisons of rating changes are "across rounds" measures of intrajudge consistency.

ACT has examined within rounds forms of intrajudge consistency data as well. ACT has provided intrajudge consistency feedback to panelists during the ALS process to inform them about the consistency of their ratings for specific items, relative to their overall item ratings. The difference between panelists' individual item ratings and the overall estimate of student performance at the borderline or cutscore provides a within rounds indicator of intrajudge consistency. Previous efforts to provide this intrajudge consistency data as feedback were not considered successful. Reckase Charts provided a means of providing this type of consistency information to panelists, along with several other consistency indicators.

INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The consistency of a judge's ratings across rounds can be examined by evaluating the percentages of items for which the ratings were changed from round to round and the magnitude of change in ratings from round to round. These data can be found in Appendix L. After reviewing the feedback presented following Round 1, Writing Pilot Study panelists were given the opportunity to change their ratings for Round 2. These changes have been reported as percentages of item rating changes in Table 5. The same procedure was followed after Round 2 when panelists could change their ratings for Round 3. These percentages of item rating changes have been displayed in Table 6. Tables 7 and 8 display the magnitude of average rating changes following the same procedures. Sets of bar graphs show the percentages of items for which ratings were raised, lowered, and unchanged from one round to the next by grade level (Figure1) and the magnitude of average rating changes from one round to the next for different types of items by grade level (Figure 2).

Table 5
Average Percentages of Item Rating Changes, by Rating Group
Round 1 to Round 2

		Raise			Lower		
Grade	Group	Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	38	39	39	17	29	24
	B	42	43	53	17	21	17
8	A	73	76	65	2	3	4
	B	30	18	12	24	20	20
12	A	30	41	44	29	24	22
	B	48	41	34	15	26	29

Table 6
Average Percentages of Item Rating Changes, by Rating Group
Round 2 to Round 3

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	27	17	11	19	15	19
	B	25	27	23	7	7	24
8	A	34	23	28	3	7	11
	B	45	34	39	3	3	3
12	A	16	16	19	17	23	24
	B	32	19	20	16	19	23

Table 7
Magnitude of Average Rating Changes (in Percentages), by Rating Group
Round 1 to Round 2

Grade	Group	Achievement Level		
		Basic	Proficient	Advanced
4	A	5.2	6.5	5.5
	B	5.4	6.6	7.7
8	A	11.8	12.0	6.4
	B	4.1	3.3	1.5
12	A	4.2	6.1	6.5
	B	4.9	5.5	4.5

Table 8
Magnitude of Average Rating Changes (in Percentages), by Rating Group
Round 2 to Round 3

Grade	Group	Achievement Level		
		Basic	Proficient	Advanced
4	A	2.4	1.5	1.2
	B	1.1	1.6	1.4
8	A	5.1	1.6	1.2
	B	5.0	1.8	1.5
12	A	1.7	1.5	1.6
	B	2.2	3.5	1.4

Figure 1

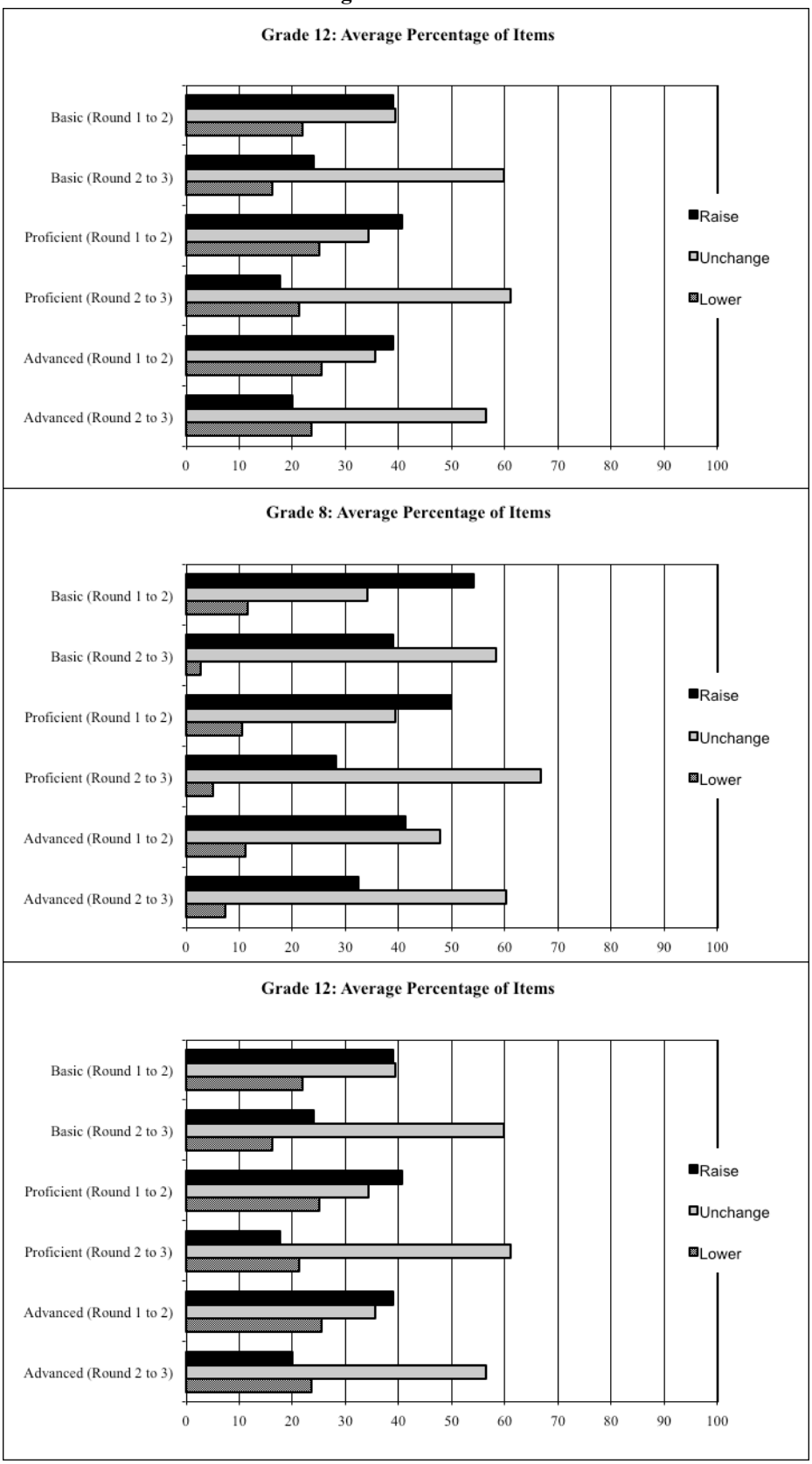
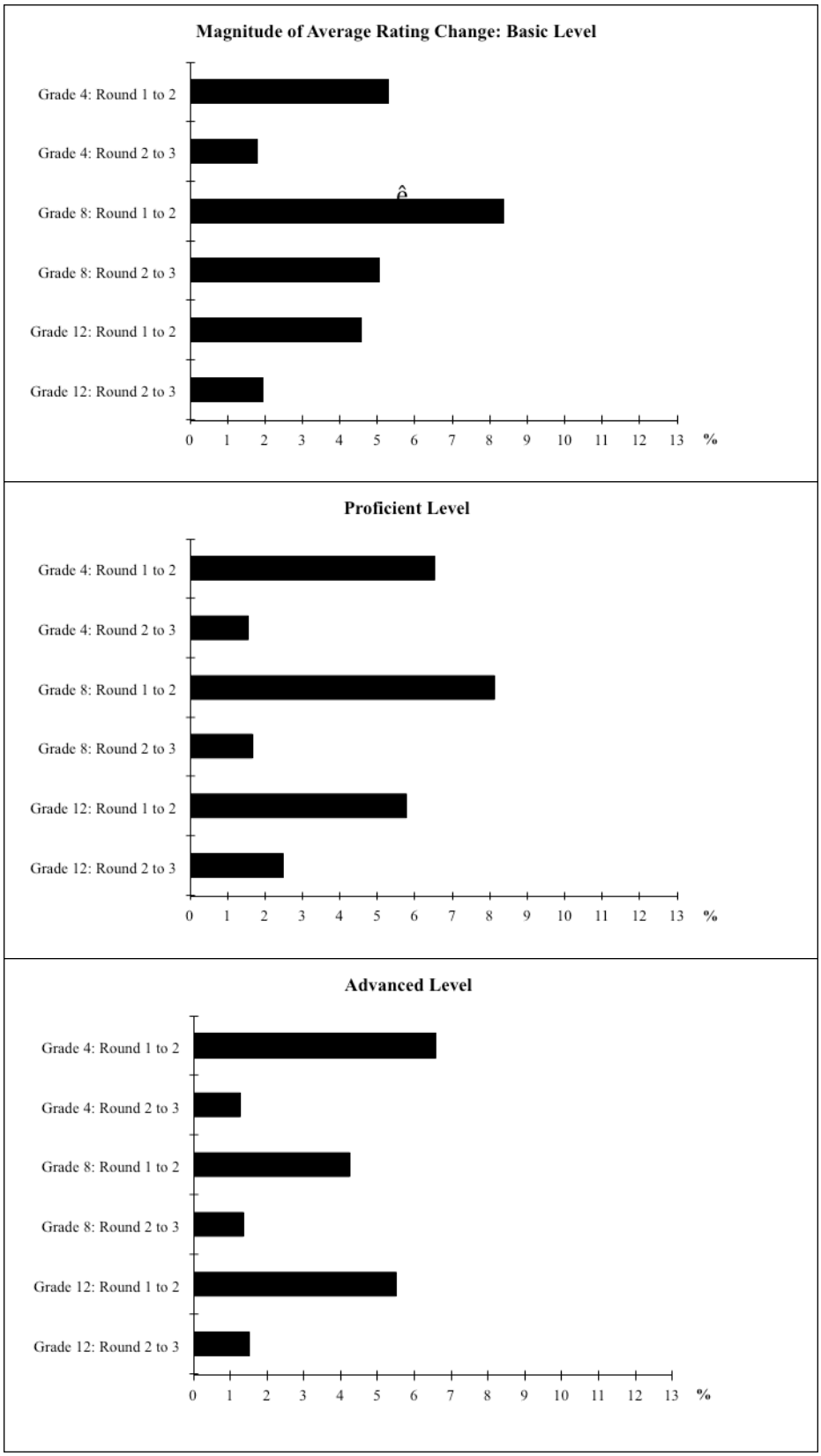

Grade 12: Average Percentage of Items
Basic (Round 1 to 2)
Basic (Round 2 to 3)
Proficient (Round 1 to 2)
Proficient (Round 2 to 3)
Advanced (Round 1 to 2)
Advanced (Round 2 to 3)
0 10 20 30 40 50 60 70 80 90 100
Raise
Unchange
Lower
Grade 8: Average Percentage of Items
Basic (Round 1 to 2)
Basic (Round 2 to 3)
Proficient (Round 1 to 2)
Proficient (Round 2 to 3)
Advanced (Round 1 to 2)
Advanced (Round 2 to 3)
0 10 20 30 40 50 60 70 80 90 100
Raise
Unchange
Lower
Grade 12: Average Percentage of Items
Basic (Round 1 to 2)
Basic (Round 2 to 3)
Proficient (Round 1 to 2)
Proficient (Round 2 to 3)
Advanced (Round 1 to 2)
Advanced (Round 2 to 3)
0 10 20 30 40 50 60 70 80 90 100
Raise
Unchange
Lower

Figure 2


Magnitude of Average Rating Change: Basic Level
Grade 4: Round 1 to 2
Grade 4: Round 2 to 3
Grade 8: Round 1 to 2
Grade 8: Round 2 to 3
Grade 12: Round 1 to 2
Grade 12: Round 2 to 3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 %
Proficient Level
Grade 4: Round 1 to 2
Grade 4: Round 2 to 3
Grade 8: Round 1 to 2
Grade 8: Round 2 to 3
Grade 12: Round 1 to 2
Grade 12: Round 2 to 3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 %
Advanced Level
Grade 4: Round 1 to 2
Grade 4: Round 2 to 3
Grade 8: Round 1 to 2
Grade 8: Round 2 to 3
Grade 12: Round 1 to 2
Grade 12: Round 2 to 3
0 1 2 3 4 5 6 7 8 9 10 11 12 13 %

Findings from previous studies show that across all grades and all achievement levels, panelists usually change their ratings on fewer items from Round 2 to Round 3 than from Round 1 to Round 2. The rating changes for the writing panelists followed this pattern. In general across grades and achievement levels, panelists raised their ratings more frequently than they lowered them. Several panelists raised every item rating at all achievement levels from Round 1 to Round 2. Very few panelists changed every item rating from Round 2 to Round 3. The two grade 8 groups changed their item ratings in different patterns. Across achievement levels, group A raised ratings on a large proportion of items, but only lowered ratings on a few items. Group B, however, raised and lowered ratings on a smaller proportion of items.

These findings suggest that panelists understood the feedback data and adjusted their item ratings in light of the information provided to them. Had panelists made large adjustments to item ratings between rounds 2 and 3, it would have indicated that panelists were perhaps confused by the feedback data or item rating methods.

Intrajudge Consistency Within Rounds (Reckase Charts Analyses)

The Reckase Charts were introduced to the Writing Pilot Study panelists as a step in the ALS process. Panelists marked their individual cutpoints and grade-level cutpoints directly on the Reckase Charts. They analyzed their ratings to discern patterns of consistency and inconsistency with respect to item ratings of any particular prompt type or category. Panelists repeated this analysis after Round 2.

Evaluation of Consequences Data

For the Writing Pilot Study, consequences data were introduced as feedback during the item-by-item rating process. Previous research by ACT had indicated that cutscores would not be significantly impacted by this feedback. At the recommendation of TACSS, ACT requested that NAGB allow consequences data to be introduced after the second round of ratings. NAGB approved the plan. This change was prompted by the findings from the civics pilot study that cutscores for each achievement level increased at each round across the three grades. The increases were small but consistent. The Reckase Charts provided panelists with a great amount of item-level information. It seemed appropriate to provide panelists with consequences information as well.

Panelists were told the percentage of students at each grade performing at or above each achievement level as feedback from Round 2 ratings. They had the opportunity to adjust their cutpoints in response to those and other Round 2 feedback data. Panelists discussed the consequences data in grade groups, and they were generally concerned about the small percentages of students scoring at or above the Advanced levels.

Individual Consequences Data

Following Round 3 ratings, panelists were given both grade-level and individual-level consequences data. In general, the effects of giving panelists consequences data appeared to be consistent with previous ACT research. That is, the data had little impact on the cutscores (Loomis, Hanick, Bay & Crouse, 2000a and 2000b). Most panelists made no changes in their cutscores after receiving consequences data, even though they had the opportunity. When asked if these percentages of students scoring at or above his/her cutscores reflected the panelist's expectations, 22 of the 57 panelists (39%) answered "yes" and 35 panelists (61%) said "no." All

35 panelists recommended changes to the cutscores. Table 9 displays these data. The changes were a mix of raising and lowering the cutscores for different achievement levels. The net effect of changes in cutscores was to slightly raise the Basic cutscores for all grades, and to slightly lower the Advanced cutscores for all grades. Panels for grades 4 and 8 lowered their Proficient cutscores and grade 12 raised their Proficient cutscore. Individual consequences data for each panelists are reported in Appendix G.

Table 9
Writing Pilot Study: Number of Changes Made to Individual Cutscores in Response to the Consequences Data Questionnaire by Grade Groups

	Grade 4 (n=21)	Grade 8 (n=18)	Grade 12 (n=18)
Data Reflects Your Expectations?			
Yes	5 (24%)	9 (50%)	8 (44%)
No	16 (76%)	9 (50%)	10 (56%)
No Response	-	-	-
(If no) Change one or more cutscore(s)?			
Yes	17 (81%)	10 (56%)	10 (56%)
No	-	-	-
No Response	-	-	-
Recommend Changes to Cutscores			
Basic			
Raise	8	3	6
Lower	3	3	0
Proficient			
Raise	4	0	6
Lower	9	5	3
Advanced			
Raise	0	0	0
Lower	12	8	5

GRADE-LEVEL CONSEQUENCES DATA

Panelists received grade-level consequences data after Round 2, after Round 3, and during the final wrap-up session. Final cutscores were computed, based on the cutscores recommended by each panelist in response to the consequences data presented as feedback after Round 3. Consequences data were computed again, and the consequences of the final cutscores were presented to panelists during the final wrap-up session.

In the wrap-up, when panelists were asked if the final percentages reflected their expectations for the proportions of students scoring at or above the grade-level cutpoints, 40 of the 57 panelists (70%) answered "yes" and 17 (30%) answered "no." Half of the grade 12 panelists (9 of 18) said that the percentages did not reflect their expectations. One judge did not complete the questionnaire. All of the panelists who responded negatively recommended changing the grade level cutpoints for one or more of the achievement levels. Results of the recommendations have been presented in Table 10. Thirty-eight panelists (68%) recommended that NAGB report the achievement levels as set, while 16 panelists (29%) recommended changes consistent with their expectations about the proportions of students scoring at or above the cutscores. One panelist from each of the grades did not respond. These responses were collected to document panelists' evaluations of the final cutscores. No adjustments were actually made to the cutscores.

Table 10
Writing Pilot Study: Number of Changes Made to Grade Group Cutscores in Response to the Consequences Data Questionnaire #2 by Grade Groups

	Grade 4 (n=21)	Grade 8 (n=18)	Grade 12 (n=18)
Data Reflects Your Expectations?			
Yes	0 (0%)	0 (0%)	0 (0%)
No	3 (14%)	4 (22%)	9 (50%)
No Response	-	-	-
(If no) Change one or more cutscore?			
Yes	3	3	9
No	0	1	0
No Response	-	-	-
Recommend Changes to Cutscores			
Basic			
Raise	0	0	6
Lower	0	0	1
Proficient			
Raise	0	1	1
Lower	1	0	2
Advanced			
Raise	1	0	2
Lower	3	3	4
Recommend to NAGB			
Grade Cutscores as Set	16 (76%)	14 (82%)	8 (47%)
Grade Cutscores Changed	5 (24%)	3 (18%)	9 (53%)
Uninterpretable/No Response	-	1	1

EVALUATION OF PANELISTS' COMMENTS AND RESPONSES TO PROCESS EVALUATION QUESTIONNAIRES

Panelists were asked their opinions about the 1998 ALS process using seven process evaluation questionnaires. Most responses were collected on a Likert-type scale, but several responses were narratives that addressed specific aspects of the process. Some questions dated back to the 1992 ALS process. Others have been added in the interim, and still others have been included to ascertain opinions about and reactions to features of the ALS process implemented for the Writing Pilot Study.

Some of the responses of panelists in the 1998 Civics Pilot Study and the 1992 Writing ALS have been presented for comparison with those of panelists in the Writing Pilot Study. In general, Writing Pilot Study panelists' responses were quite positive with respect to their standard setting experiences. The process implemented in the 1992 Writing ALS was considerably different from the one used in the 1998 Writing Pilot Study. The responses from the 1992 ALS have been included as a point of interest, rather than a reliable reference for comparison. Even though civics and writing are different content areas, the Civics Pilot Study responses have been included for comparison purposes because of the similarity of the processes used in the two studies.

Summaries of responses to all questions on the process evaluation questionnaires have been included in Appendix M. Responses highlighted here are those for evaluation items that have been consistently reported for several subjects and those for items related to special features of the 1998 ALS process implemented in the Writing Pilot Study.

UNDERSTANDING THE RATING PROCESS AND CONFIDENCE IN RATINGS

Data reported in Table 11 show the average responses (5=most positive and 1=most negative) to questions about the rating sessions round by round. As expected, panelists' responses generally reflected an increase in understanding and confidence as the rounds of ratings progressed. By Round 3, the responses were very high to questions about the *clarity of instructions* and the *level of understanding* of the tasks (range 4.76 – 4.88). The *level of confidence* increased substantially from Round 1 to Round 3 as reflected by the considerable increase the degree of positive response for each grade (grade 4 increased 1.62 points, grade 8 increased 1.72 points, and grade 12 increased .91 points). Many judges commented that providing the first round of rating was a difficult task for them, which is reflected in the lower responses after Round 1.

Table 11
Writing Pilot Study Evaluation Questionnaires
Summary of Responses to Questions Related to Ratings

Questions	Round	Writing Pilot Study		
		Grade 4 (n=21)	Grade 8 (n=18)	Grade 12 (n=18)
1. The instructions on what I was to do during the 1st/2nd/3rd rating session were: *(5=Absolutely Clear; 1=Not at all Clear)*	1	3.04	3.33	3.71
	2	4.57	4.44	4.55
	3	4.81	4.88	4.83
2. My level of understanding of the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Totally Adequate; 1=Totally Inadequate)*	1	3.00	3.38	3.71
	2	4.52	4.44	4.61
	3	4.76	4.77	4.77
3. The amount of time I had to complete the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Far too Long; 3=About Right;1=Far too Short)*	1	3.09	3.55	3.18
	2	3.71	3.77	3.66
	3	4.00	4.11	3.55
4. The most accurate description of my level of confidence in the ratings I provided to represent the three achievement levels during the 1st/2nd/3rd rating session is that I was: *(5=Totally Confident; 1=Not at all Confident)*	1	2.57	2.83	3.47
	2	3.71	4.05	3.88
	3	4.19	4.55	4.38

Another point of interest was the response to the question related to the amount of time panelists had to complete the rating tasks. After Round 1, most panelists indicated that the amount of time was about right to complete the task (5= far too long, 3 = about right, and 1= far too short). For each successive round, panels responded that they had more time than they actually needed to do their work. It would seem that participants had plenty of time to complete their rating tasks and were not rushed through the rating sessions.

UNDERSTANDING OF THE ACHIEVEMENT LEVELS DESCRIPTIONS AND BORDERLINE PERFORMANCE

A typical response pattern that has emerged from past ALS meetings is that panelists generally understand performance across the levels associated with achievement levels descriptions better than borderline performance. The questions about achievement levels descriptions ask about the level of understanding, and the questions about borderline performance ask about the extent to which the concept is well formed. Panelists' understanding of both categories of performance usually increases over rounds so that the difference between the two diminishes by Round 3. Results from the Writing Pilot Study varied slightly from this expected pattern. For the Writing Pilot Study, understanding the definition of borderline performance was noticeably lower than

understanding the definition of achievement level performance for grade 4 and grade 8 panels after Round 1, but not for grade 12. Responses were reversed for grade12; that is, their understanding of borderline performance after Round 1 was slightly *higher* than their understanding of performance for overall achievement levels. All three grade-level panels indicated highly positive responses when asked about their understanding of the definitions of achievement level performance and borderline performance after Round 3. Panelists' understanding of student performance across the achievement levels approached *absolutely clear* by Round 3. The mean of their responses ranged from 4.33 to 4.55 by Round 3. Their conception of borderline performance approached *very well formed* for all achievement levels by Round 3, with the means ranging from 4.28 to 4.47. Table 12 shows that panelists' overall understanding of student performance at the borderline and at the three achievement levels increased with each round of ratings for each grade.

Table 12
Writing Pilot Study Evaluation Questionnaires
Summary of Responses to Questions Related to Achievement Levels Descriptions

Questions	Round	Writing Pilot Study		
		Grade 4 (n=20)	Grade 8 (n=18)	Grade 12 (n=18)
1. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Basic level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.81 4.19 4.33	3.76 4.22 4.55	3.66 4.50 4.55
2. At the time I provided the 1st/2nd/3rd set of ratings my conception of Borderline Basic performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.38 4.04 4.28	3.24 4.05 4.44	3.88 4.44 4.38
3. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Proficient level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.66 4.09 4.33	3.71 4.22 4.44	3.66 4.05 4.44
4. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Proficient performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.38 3.71 4.47	3.24 4.00 4.38	3.77 4.05 4.33
5. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Advanced level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.95 4.04 4.42	3.72 4.33 4.44	3.65 4.22 4.55
6. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Advanced performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.52 3.85 4.47	3.16 4.00 4.38	3.82 4.00 4.33

Although grade 12 panels appeared to understand borderline performance after Round 1 better than performance across achievement levels, having panelists write borderline descriptions did not seem to have an obvious, significant impact on their ability to form a clear concept of borderline performance. Panelists for the Writing Pilot Study and the Civics Pilot Study wrote borderline descriptions and modified them throughout the process. In contrast, other ALS panelists discussed their concept of borderline performance with other panelists and used it in training exercises prior to rating items. It was anticipated that the Writing Pilot Study panelists and the Civics Pilot Study panelists would respond more positively about their understanding of borderline performance than their understanding of ALDs, given that they wrote borderline descriptors but did not modify ALDs. When compared with responses from other ALS panelists,

the pilot study panelists' understanding of student performance at the borderline *and* at the three achievement levels was noticeably higher after Round 3 in most cases. Understanding of borderline performance did not surpass that for performance across achievement levels, however.

EVALUATIONS OF FEEDBACK

Many different types of feedback information were given to the panelists during the ALS process. When asked if they were planning to use *all* the feedback information to adjust their ratings during Round 2, most panelists agreed (grade 4 = 20; grade 8 = 17; grade 12 = 17; when 5 = *totally agree* and 1 = *totally disagree*). These data suggest that when panelists were modifying their ratings, they were not overly influenced by one type of feedback to the exclusion of all others. Most panelists agreed that the Reckase Chart was the most useful type of feedback information (please see Table 13). With regard to the amount of feedback given to panelists during the rating process, most panelists remarked that they were able to manage the amount of information without confusion, but acknowledged that they were reaching their limit.

Table 13
Writing Pilot Study: Response Frequencies for Choosing the Most Useful Type of Feedback

Types of Information	Grade 4	Grade 8	Grade 12
If you had to choose one, and only one, of the following types of information to use during the rating process, what would it be? (Reported after Round 3 ratings)			
Cutscores and their standard deviations	1 (5%)	2 (11%)	1 (6%)
Student performance data on each item	5 (24%)	1 (6%)	2 (11%)
Rater location feedback	0 (0%)	2 (11%)	1 (6%)
Whole booklet feedback	2 (11%)	1 (6%)	2 (11%)
Reckase Charts	12 (57%)	12 (67%)	12 (67%)
Blank	1 (5%)	0 (0%)	0 (0%)

Panelists were asked about the decision model they used when determining performance at the Advanced achievement level. In particular, judges were asked questions that would indicate whether they used a compensatory model. Panelists' responses are displayed in Table 14. Only 5 panelists responded that students performing at the Advanced level should demonstrate superior performance on *all* prompts in the assessment (noncompensatory model). Fifty-two of the 57 panelists indicated that students performing at the Advanced level should demonstrate superior performance on *most* or *some* prompts in the assessment (compensatory model). Only 6 panelists responded that students performing at the Advanced level should demonstrate superior performance on *all* types of prompts in the assessment (noncompensatory model). Fifty-one panelists indicated that students performing at the Advanced level should demonstrate superior performance on *one* or *two* prompts in the assessment (compensatory model).

Table 14
Writing Pilot Study: Response Frequencies for Compensatory/Conjunctive Judgments

Question	Response	Grade 4	Grade 8	Grade 12
Students performing at the Advanced level should demonstrate superior performance on All/Most/Some prompts in the assessment.	All	2 (10%)	1 (6%)	2 (11%)
	Most	19 (90%)	13 (72%)	13 (72%)
	Some	0 (0%)	4 (22%)	3 (17%)
Students performing at the Advanced level should demonstrate superior performance on All/Two/One type(s) of prompt(s) in the assessment.	All	2 (10%)	1 (6%)	3 (17%)
	Two	16 (76%)	12 (67%)	8 (44%)
	One	3 (14%)	5 (28%)	7 (39%)

Comments about the Reckase Charts

Panelists' comments about the Reckase Charts were overwhelming positive, although participants offered some suggestions regarding the chart. Several participants remarked that the visual display helped them conceptualize the information. A summary of panelists' diverse comments about the charts has been included in Appendix M.

THE OVERALL ALS PROCESS

Data reported in Table 15 show the average responses to questions from the final questionnaire about the overall ALS process used for the Writing Pilot Study. When asked if the achievement levels were "defensible" and "reasonable," Writing Pilot Study panelists' responses were consistently lower compared with those for the Civics Pilot Study and 1992 Writing ALS. This was particularly true for grade 8 and grade 12 panels. Responses were less positive when Writing Pilot Study panelists were asked about their willingness "to sign a statement" recommending the achievement levels than responses from the other two sessions. Forty-nine of the 57 judges were willing to sign a statement recommending the use of the achievement levels, 6 panelists responded "no, probably not," and 1 said "no, definitely not." These results indicated that the Writing Pilot Study panelists had greater reservation about endorsing results of the ALS process than panelists from either the Civics Pilot Study or the 1992 Writing ALS.

Table 15
Final Writing Pilot Study Evaluation Questionnaire
Summary of Mean Responses to Questions About the ALS Process Taken as a Whole

Questions	Writing Pilot Study			1992 Writing ALS			Civics Pilot Study		
	Grade 4 (n=20)	Grade 8 (n=18)	Grade 12 (n=18)	Grade 4 (n=22)	Grade 8 (n=21)	Grade 12 (n=23)	Grade 4 (n=16)	Grade 8 (n=18)	Grade 12 (n=16)
1. The most accurate description of my level of confidence in the achievement levels ratings I provided was: *(5=Totally Confident; 1=Not at all Confident)*	4.10	4.11	4.11	4.06	4.11	3.82	3.81	4.06	4.19
2. I would describe the effectiveness of this achievement levels-setting process as: *(5=Highly Effective; 1=Not at all Effective)*	3.95	3.89	3.67	4.18	4.44	3.86	4.00	3.67	3.88
3. I feel that this NAEP ALS process provided me an opportunity to use my best judgment in rating items to set achievement levels for the NAEP Writing Assessment: *(5=To a Great Extent; 1=Not at All)*	4.10	4.11	4.00	4.36	4.18	4.09	4.19	4.17	4.00
4. I feel that this NAEP ALS process produced achievement levels that are defensible: *(5=To a Great Extent; 1=Not at All)*	4.10	3.83	3.94	4.27	4.57	4.13	4.31	4.17	4.25
5. I feel that this NAEP ALS process produced achievement levels that will generally be considered reasonable: *(5=To a Great Extent; 1=Not at All)*	4.10	3.27	4.00	4.43	4.76	4.43	4.31	4.00	4.25
6. I would be willing to sign a statement (after reading it, of course) recommending the use of the achievement levels resulting from this achievement levels-setting procedure: *(4=Definitely, 3=Probably, 2=Probably not, 1=Definitely not)*	40.0% 45.0% 15.0% 0.0%	44.4% 44.4% 11.1% 0.0%	44.4% 44.4% 5.6% 5.6%	50.0% 50.0% 0.0% 0.0%	85.7% 9.52% 0.0% 0.0%	52.2% 43.5% 4.35% 0.0%	75.0% 25.0% 0.0% 0.0%	50.0% 50.0% 0.0% 0.0%	50.0% 50.0% 0.0% 0.0%

EVALUATION OF THE SELECTION OF EXEMPLAR PERFORMANCES

One of the primary outcomes of the NAEP ALS meeting is the identification of assessment performances that illustrated the knowledge and skills associated with each achievement lev each type of writing (narrative, informative, persuasive) to use in reporting NAEP results. Appendix J includes panelists' classifications of all papers that qualified for consideration a exemplars, and those papers selected by the Writing Pilot Study panelists to serve as exemp performances for reporting the NAEP achievement levels.

Panelists reviewed and discussed the student writing that qualified statistically for considera as exemplars. (Please see Table 16.) In general, there was a mix of prompt types for panelis consideration in the exemplar selection process. The exceptions were at the grade 4 Advan level; grade 8 Proficient level; and grade 12 Proficient and Advanced levels. For those level there was no score level that met the statistical criterion to qualify a prompt in a particular t writing. For example, there was no score that met the criteria for the Persuasive prompt at t grade 4 Advanced level.

Panelists were instructed to "veto" performances that met the statistical criterion but that di meet the criteria described in the ALDs. Judges were trained in the statistical requirements t had been used for selecting the performances (described earlier in this report). In addition, panelists were instructed to use their knowledge of the achievement levels descriptions to evaluate each paper in terms of its quality as an illustrative or exemplar performance. Panel were given tables with feedback on their paper classifications for these prompts. That information allowed them to review the number of panelists who had judged each paper sco the qualifying level to be in each achievement level or at the borderline of each.

Table 16
Writing Pilot Study: Number of Student Papers that Qualified and Were Selected as Exemplars

	Number Primary List	Number Selected
Grade 4		
Basic	3	1
Proficient	9	5
Advanced	3	3
Grade 8		
Basic	9	6
Proficient	6	4
Advanced	12	6
Grade 12		
Basic	15	9
Proficient	12	6
Advanced	12	4

The process was based on general agreement among panelists regarding whether each paper should be used as an exemplar of performance at a specific achievement level. Facilitators allowed the group, as a whole, to determine whether papers would be accepted or vetoed. In instances, papers were approved although a few people vetoed the paper. An intense minor however, succeeded in their veto. The general will of the group was the guide.

REMARKS FROM DEBRIEFING SESSION

Shortly after the pilot study was adjourned, the Project Director held a debriefing session for invited panelists and facilitators. Panelists were selected from each grade group and panelist type. Four facilitators and eleven panelists of those invited were able to stay after the meeting was adjourned. The composition of the debriefing panel approximated the overall pilot study panel. A general discussion was held to evaluate key elements of the standards-setting process. A complete list of the discussion topics included in the debriefing session has been included in Appendix N. When the session was concluded, the last of the panelists were thanked for their work and the meeting was officially over. The following is a summary of the remarks made during the debriefing session.

ABOUT THE ACHIEVEMENT LEVELS DESCRIPTIONS

Panelists at the debriefing expressed concern about the precision of the language used in the ALDs. In particular, the descriptions for grade 12 stated that Basic required an "effective" response. Grade 12 panelists felt that an "effective" response actually described Advanced writing, not Basic. Grade 12 panelists had been distracted throughout the process by the wording of some of the ALDs. They also took exception to the use of the term *genre* and recommended that staff refer to narrative, informative, and persuasive *types* of writing.

ABOUT FEEDBACK

Panelists were asked their reactions to the feedback in general and to the whole booklet feedback, in particular. Of interest was to determine if panelists truly understood these data. For example, the whole booklet feedback after Round 2 showed that grade 4 students needed to score 11 of the 12 possible points to perform at the Borderline Advanced level. The grade 4 panel felt that this outcome was consistent with their expectations of Advanced performance. Participants in the debriefing session verbalized an understanding of these data and a general agreement that the results were "reasonable."

ABOUT IMPROVING THE ORGANIZATION OF THE PROCESS

Participants voiced few complaints about their experience on the panels. They thought that the agenda was too busy at the beginning of the meeting, and somewhat uneven after the third round of ratings on the last day. They understood that the pilot study was practice for the ALS meeting and offered suggestions to improve the timing of the process.

ABOUT ESTIMATING THE AVERAGE SCORE

Panelists were somewhat shocked when they were instructed to estimate the average score as the way they were to rate student performance for each prompt. The meeting director questioned them about how this could be improved. The panelists suggested modifying the first training exercise when they learned about the relationship between the generic scoring rubrics and the ALDs. They suggested that panelists be instructed to think about the range of scores for which the rubric seemed to match the ALDs. They would then conceptualize an average score that would best capture that range. This alteration not only would cause judges to start thinking about average performance, but also would break any association between specific rubric scores and the ALDs.

EVALUATION OF THE OVERALL ALS PROCESS

One of the primary purposes of the writing pilot study was to test the procedures planned for implementation in the Writing ALS. ACT conducted the Civics Pilot Study before the Writing Pilot Study. Consequently, ACT made many adjustments to the process implemented for the Writing Pilot Study as a result of earlier observations and experiences. These modifications seemed to improve the process so that the Writing Pilot Study panelists had few complaints and problems. Nonetheless, ACT staff still recommended several improvements for the Writing ALS meeting as a result of the Writing Pilot Study. The following section summarizes the procedural issues that needed modifications and the adjustments that were planned for the 1998 Writing NAEP ALS.

ADJUSTING THE AGENDA

Check-in on Day 1 started at 2:00 PM and the opening session began at 3:00 PM. Some panelists complained about being instructed to arrive an hour early for the meeting. Since this was a travel day, many panelists had not eaten lunch and were hungry when they arrived. No refreshments had been arranged. This will be corrected in the future.

ACT had added a social "mixer" as a pre-dinner activity at the end of Day 1. Panelists appeared to enjoy the event. They also enjoyed being free to socialize after dinner as long as they wished. The agenda for several previous ALS studies has required panelists to return to grade groups and take the NAEP or participate in some other training exercise after dinner. ACT will continue to study the agenda and the physical, social, and logistical considerations for panelists on Day 1. It is difficult to have Day 1 serve as a travel day for most panelists and to also be the key process orientation day and social "get acquainted" day.

Day 2 was too busy. The ALD discussion lasted for the entire morning even without explaining the scoring procedures in detail. A demonstration of how to rate an item replaced the Preview Rating Session, which was an adjustment made after implementing the Civics Pilot Study. The demonstration required less time than the preview session, but panelists still did not have adequate time to develop their initial written versions of borderline descriptions.

Day 5 was very busy and too uneven in pace. With few prompts to rate, panelists had more than an hour of unstructured time between Round 3 ratings and lunch. To streamline the distribution of consequences data after Round 3, panelists were assigned seats for the general session. This arrangement eliminated the possibility that panelists could discuss the consequences data openly among grade-level panelists before they made recommendations for their final cutpoints.

UNDERSTANDING ACHIEVEMENT LEVELS DESCRIPTIONS RELATIVE TO SCORING RUBRICS

ACT staff agreed with the panelist's recommendation made during the debriefing session to modify the ALD training exercise. Panelists should discuss the details of the scoring rubrics in relation to the ALDs and "estimate the average" score they would associate with the ALDs. This modification would give panelists early exposure to the concept of average scores and diminish the tendency to form a one-to-one association between the scoring rubric and the ALDs.

UNDERSTANDING COMPENSATORY SCORING

Concern has been expressed about the need for panelists to understand that a compensatory model is used to scale NAEP. Panelists would have no direct experience with a compensatory model

when they use an item-by-item rating method. The instructions for the Reckase Charts directly addressed the issue of compensatory scoring when they explained average performance (please see Appendix H for the exact instructions). During the paper/booklet classification exercise on Day 2, some panelists were very surprised by the fact that a student might give an "off task" response to the first writing task and write a reasonably good response to the second. Although panelists were made aware of this issue, some seemed resistant to accepting the reality of uneven student performance and the concept of compensatory scoring.

RATING PROMPTS

Panelists were quite disturbed about estimating an average score. Many were fixed on the whole numbers associated with the scoring rubric and could not reconcile themselves to an average score including a decimal value. During training the meeting director inadvertently omitted a demonstration of computing an average score using whole numbers that resulted in a decimal. This will be added to the demonstration for rating a prompt.

WRITING BORDERLINE DESCRIPTIONS

The ALDs for Basic performance for all grades were very brief. It was difficult for panelists to describe Borderline Basic performance that was more minimal than that described by the Basic achievement level description. The meeting director suggested that panelists consider writing borderline descriptions by starting at the Advanced level and working to Basic. It was reasoned that because the ALDs included many attributes that described Advanced performance, working from the Advanced descriptions might add information to the borderline Basic ALD.

PRESENTING FEEDBACK IN A LOGICAL ORDER

The order of presenting feedback was rearranged. Panelists received feedback from the most holistic level to the most item-specific level. This seemed to work well and to help panelists identify the different pieces of feedback more readily.

1. Cutscores and their standard deviations
2. Whole booklet feedback
3. Whole booklet exercise
4. Rater location data
5. Student performance data
6. Reckase charts

SELECTING BOOKLETS FOR THE WHOLE BOOKLET EXERCISE

For the Whole Booklet Exercise, ACT expected to select booklets to represent performance at the cutscores by first estimating the theta for each cutpoint and then selecting booklets scored near the theta value. The plan was conceived to help alleviate the problems associated with writing scores that ranged only from 2-12 points on the raw score scale. The score distribution for writing was unusual in that the scores tended to cluster in clumps rather than to spread across a normal distribution curve. Unfortunately, there were few booklets in the sets of 100 randomly-selected booklets for each form that were scored near that theta value. With no booklets on site that matched the cutscore on the theta scale, ACT resorted to the typical method of estimating the percent of total possible points using the test characteristic curve for the particular booklet forms used in the whole booklet feedback.

RECOMMENDING FINAL CUTPOINTS

For the Consequences Data Questionnaire, some panelists remarked that they would have preferred to recommend a percentage of students performing at or above the cutpoint rather than a cutpoint. Because they had given careful consideration to their item-by-item ratings, many panelists thought that changing the cutpoints was arbitrary.

USING AN ELECTRONIC SCANNER FOR DATA ENTRY

ACT used an electronic scanner for the first time to enter item ratings in the Writing Pilot Study. Several errors in ratings were detected during the meeting, and those were corrected on site. Because many problems were encountered, key entry was necessary for numerous forms. The data were rechecked after the meeting and further corrections were made.

CONCLUSIONS DRAWN FROM WRITING PILOT STUDY RESEARCH

The Writing Pilot Study was the final opportunity for ACT to evaluate procedures to be used for the operational Writing NAEP ALS. The purpose of the study was to identify needed adjustments in the ALS process for training, instructing, timing, and other key activities to assure a successful ALS. Another important objective of the study was to evaluate panelists' reactions to incorporating the new Reckase Charts into the ALS process. The following summaries are the conclusions drawn from the Writing Pilot Study.

THE ISSUE OF IMPROVING AND REFINING THE STANDARD SETTING PROCESS

Years of refinements have led to the current process, which has been considerably enhanced by the most recent addition of the Reckase Charts. The charts were created specifically for use in setting NAEP standards, although they could be used easily in other standard-setting contexts. Incorporating the charts into the ALS process helped to overcome difficult technical challenges to setting achievement levels for NAEP. The Reckase Charts proved to be a powerful tool that enabled laypersons to work with item measurement data that otherwise would have been too technical to comprehend.

A concern associated with incorporating the Reckase Charts into the ALS process was that panelists would rely on the chart data to the exclusion of other sources of relevant feedback, possibly deferring their judgment to the statistical data shown on the chart. In particular, ACT, TACSS, and NAGB's COTR were all concerned that panelists would loose their standards-based focus—their focus on ALDs as **the** criteria by which to judge student performance—and rely solely upon the model-based estimates of student performance. Although panelists were greatly impressed by the usefulness of the charts and the ease of using them, they indicated that they considered other forms of feedback as well when forming their judgments. The Reckase Charts did not overly influence panelists when modifying their ratings, to the exclusion of other types of feedback. There was no evidence of undue influence based on observations of panelists working with the charts and panelists responses to questionnaire items.

As a result of recommendations by Civics Pilot Study panelists, ACT had developed a computer procedure for transferring ratings onto the Reckase Charts being presented as feedback for Round 1 and Round 2 ratings. The markings seemed relatively easy for panelists see and identify with ratings for each achievement level. Panelists seemed particularly impressed with the understanding they gained by connecting their ratings for each prompt at each achievement level. In addition, ACT had developed a special computer presentation for instructing panelists in the

use of Reckase Charts, and that seemed quite successful. ACT staff and observers had no further recommendations for improving the Reckase Charts.

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

One persistent challenge to improving the ALS process has been to find a way to provide panelists with information about the relationship between their individual item ratings and student performance. This sounds relatively simple, but the issue is *how* to identify the relevant level of student performance. Individual item ratings can be used to compute a cutscore for each panelist. That cutscore then becomes the representation of a panelist's concept of borderline performance for a level of achievement. The panelist's ratings for each prompt are associated with an overall performance score (cutscore). If all of the item ratings are for the same performance score, then the panelist has managed to estimate student performance for each prompt to be perfectly consistent with the IRT model used to estimate student performance on NAEP. Certainly, that was not the case in the 1998 process. Most panelists judged some prompts to be much harder or much easier than others, relative to their overall cutscore. Intrarater consistency is a measure of the extent to which individual item ratings are consistent with the overall cutscore estimated from the individual item ratings, given student performance on the prompts. Although this information has been given to panelists in previous ALS meetings, there was little indication that panelists either understood the information, or found it useful when forming their judgments about student performance. Reckase Charts made that information easy to assess.

After panelists studied the Reckase Charts, they generally adjusted their ratings to be more similar to the IRT-based performance estimates of students at the cutscores—either their own cutscores or the grade-level cutscores. This finding was consistent for all three achievement levels at all three grades.

It is important to note, however, that none of the judges adjusted his/her ratings to be identical to IRT-based performance estimates. Such an adjustment would be indicated by judges rating all prompts at a single scale score or a single row on the chart. The fact that this did not happen suggested that panelists considered the achievement levels descriptions and other forms of feedback in addition to the charts when forming their judgment of student performance. After considering all of this information, panelists formed judgments that were not exactly the same as the IRT-based estimates of student performance. Responses to the process evaluation questionnaires supported this interpretation.

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The ALS process designed by ACT provided panelists with extensive feedback and instructions for interpreting information when forming their judgments of student performance. Panelists were encouraged to reconsider their ratings and adjust them according to their interpretation of the many sources of information available to them. It was reasoned that if panelists understood the item rating method and the feedback produced by the method, they would adjust their ratings from round to round. If panelists did not adjust their ratings at all, it indicated that they probably did not understand the rating method or the feedback. On the other hand, if they changed all—or most—of their ratings after two rounds, it indicated that they probably did not understand the rating method or the feedback. The Writing Pilot Study panelists exhibited "reasonable" intrajudge consistency across rounds based on the percentage of item ratings changed and the magnitude of change in item ratings.

THE ISSUE OF COGNITIVE COMPLEXITY

The charge has been made that item-by-item rating methods cannot produce valid cutpoints because panelists are incapable of performing the cognitively complex task of estimating probabilities or mean scores with reasonable accuracy (NAE, 1993: Shepard, 1995: Impara and Plake, 1998). ACT has collected considerable data during the Writing Pilot Study and previous research where panelists have reported their capacity to perform the tasks associated with estimating student performance. Judges perceived that they performed the required estimation and judgmental tasks with relative ease. They reported that they were confident in their judgments and satisfied with the results. There is no evidence to indicate that panelists felt unable to make the item-by-item judgments or that they were incapable of estimating probabilities with reasonable accuracy.

PLANNING FOR THE WRITING ALS

The details of implementing the standard setting process were closely scrutinized and evaluated during the Writing Pilot Study. Panelists offered a wealth of information about adjusting the ALS process, from their difficulties in rating prompts the first time, to managing the amount of feedback information presented for their consideration. As a result of extensive research conducted not only for the 1998 Writing NAEP, but also for all the assessments administered since 1990, ACT anticipated conducting the most refined and technically precise NAEP ALS meeting to date.

REFERENCES

ACT (1997). *Developing achievement levels on the 1998 NAEP in civics and writing: Design document.* Iowa City, IA: Author.

Chen, Wen-Hung (1998, April). *Setting achievement level standards for NAEP using response pattern estimation: A simulation study*. Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

Chen, Wen-Hung & Loomis, S.C. (2000). "Computational procedures used in field trials, pilot studies, and the operational achievement levels-setting studies for the 1998 NAEP in civics and writing" in Chen, Wen-Hung, Loomis, S.C. & Fisher, T., *Developing achievement levels on the 1998 NAEP in civics and writing: Technical report.* Iowa City, IA: ACT.

Impara, J.C. & Plake, B.S. (1997). *Standard setting: An alternative approach.* Paper presented at the annual meeting of the American Educational Research Association, 1997, Chicago.

Impara, J.C. & Plake, B.S. (1998). Teachers' ability to estimate item difficulty: a test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement*, 35(1), 67-81.

Loomis, S.C., Bay, L., Yang, W.L., & Hanick, P.L. (1999). *Field trials to determine which rating method(s) to use in the 1998 NAEP achievement levels-setting process*. Paper presented at the meeting of the NCME, Montreal.

Loomis, S.C. & Hanick, P.L. (2000). *Setting standards for the 1998 NAEP in civics and writing: Finalizing the achievement levels descriptions*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., Bay, L. & Crouse, J.D. (2000a). *Developing achievement levels on the 1998 National Assessment of Educational Progress in civics: Field trials final report*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L., Bay, L. & Crouse, J.D. (2000b). *Developing achievement levels on the 1998 National Assessment of Educational Progress in writing: Field trials final report*. Iowa City, IA: ACT.

Loomis, S.C., Hanick, P.L. & Yang, W.L. (2000). *Developing achievement levels for the 1998 NAEP in civics interim report: Pilot study*. Iowa City, IA: Author.

MDR's School Directory (20th Edition) [Electronic data]. (1997). Shelton, CT: Market Data Retrieval [Producer and Distributor].

National Academy of Education (1993). *Setting Performance Standards for Student Achievement*, Robert Glaser, Robert Linn, and George Bohrnstedt, eds. Panel on the Evaluation of the NAEP Trial State Assessment. Stanford, CA: Author.

Reckase, M.D. (1998). *Setting standards to be consistent with an IRT item calibration*. Iowa City, IA: ACT.

Reckase, M.D. & Bay, L. (1999). *Comparing two methods for collecting test-based judgments*. Paper presented at the annual meeting of the National Council on Measurement in Education, 1999, Montreal.

Rodenhouse, M.P. & Torregrosa, C.H. (1998). *1998 Higher Education Directory*. Falls Church, Virginia: Higher Education Publications.

Shepard, L.A. (1995). *Implications for Standard Setting of the NAE Evaluation of NAEP Achievement Levels.* Proceeding of the Joint Conference on Standard Setting for Large Scale Assessments. Washington, DC: National Assessment Governing Board and National Center for Educational Statistics.