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

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

December 2000

Developing Achievement Levels for the 1998 NAEP in Civics 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 Civics Pilot Study was an opportunity to continue studying and refining procedures that were introduced in earlier field trials for both civics and writing. 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 Civics ALS process was implemented.

ACT was particularly interested in evaluating panelists' reactions to incorporating the new Reckase Charts as a step in the ALS process. Reckase Charts were taken from the Reckase Method, as implemented in the field trials, and incorporated into the ALS process as a step in preparation for the rating process. This was the first opportunity for ACT to study the process with Reckase Charts used in this manner.

The decision had been made to provide consequences data after the third, final round of item-by-item ratings. Panelists would be given consequences data associated with their own cutscores, and they could change any or all of their cutscores as a result of this information. The cutscores resulting from these recommendations would be "final." They would be used for selecting exemplar items and as recommendations to the National Assessment Governing Board (NAGB) unless some reason could be found to choose the Round 3 cutscores instead.

OBSERVATIONS AND EVALUATION OF THE OVERALL PILOT STUDY PROCESS

AGENDA

Issue: Panelists wanted to lengthen the five-day meeting to allow more time to assimilate information, get to know other participants, and relax after a full day of intense work.

Adjustment: ACT will start the meeting earlier on Day 1 and add a social "mixer" to the activities planned at the end of the day. ACT did not extend the ALS meeting to six days because of concerns about the loss of participation in the process by outstanding panelists.

ACHIEVEMENT LEVELS DESCRIPTIONS

Issue: ACT anticipated that some panelists would be troubled by the fact that they would be working with the finalized versions of the ALDs without the opportunity to modify the descriptions to reflect their own judgment of student performance.

Adjustment: ACT's concern was unfounded. There was no evidence that panelists were troubled by not having the opportunity to modify the ALDs. None of the judges expressed a desire to revise the descriptions.

Issue: Participants commented that working with the ALDs would be easier if they were in bullet format.

Adjustment: Because the ALDs were written in a narrative format, the meaning of the ALDs would change if they were reformatted to bulleted statements. NAGB's policy is to report ALDs in narrative format. No recommendation was made to NAGB to change the format.

WRITING DESCRIPTIONS OF BORDERLINE PERFORMANCE

Issue: As a training exercise, panelists spent considerable time and effort writing descriptions of borderline student performance. There was no evidence to suggest that writing descriptors for borderline student performance enhanced panelists' understanding of borderline performance more so than discussions of borderline performance.

Adjustment: ACT decided to retain this training exercise even though panelists did not respond more positively to the evaluation questions related to their level of understanding borderline performance.

A primary reason for including the exercise of writing borderline descriptions was to sharpen the focus of panelists on the ALDs. Since no modifications were allowed to the ALDs, this was the major motivation for panelists to study the ALDs carefully.

PREVIEW RATING SESSION

Issues: A Preview Rating Session had been added to the agenda to give panelists an understanding of how important the information about and training in borderline performance would be later in the process. Judges struggled greatly with rating NAEP items during the Preview Rating Session. ACT staff concluded that panelists had difficulty making the transition from applying the ALDs to evaluate assessment items, to applying the ALDs to estimate borderline student performance. Further, panelists were disturbed by the fact that they did not receive any type of feedback after the Preview Rating Session. Since training for the rating process had not been completed, ACT did not want to provide feedback to inform panelists about their ratings.

Adjustments: ACT will change the agenda so that panelists will be introduced to the rating process early in the meeting through a demonstration The introductory activity will illustrate the rating process by giving an example of each step in the procedure. Panelists will not participate in the rating of items at this early stage of the process so there will no longer be a question of providing feedback. This activity will be referred to as a training exercise, rather than the "Preview Rating Session."

PAPER SELECTION EXERCISE

Issue: The Paper Selection Exercise required judges to examine three student papers scored at each score point for all of the constructed response items in each panelist's item rating pool. The grade 4 group reviewed 138-141 student papers, the grade 8 group reviewed 165-168 papers, and the grade 12 group reviewed 171-174 papers. Although most of the student responses were very short answers, panelists complained that this exercise was tiring and caused them to feel fatigued before undertaking the first round of item-by-item ratings. They wanted more time to discuss their selections with other panelists in their group, and they wanted feedback from the exercise so they could judge how "accurately" they had selected papers.

Adjustments: ACT will reduce the number of papers that ALS panelists will examine during this exercise. Panelists will review the student papers only for the common blocks of items that all panelists rated in their grade group. This will provide time to discuss the selections and enhance the training process. As a result, 72 student papers will be the maximum number reviewed by grade 4 panelists, and about 65 papers will be the maximum number reviewed by grade 8 and 12 panelists. Scored papers for the remaining constructed response items in the rating pool will be available for judges to review prior to round 1 ratings. Further, ACT will give panelists feedback from the Paper Selection Exercise. Frequencies will be reported so that panelists will know how many panelists selected specific papers to represent performance at the borderline of Basic, Proficient, and Advanced achievement levels.

RECKASE CHARTS

Issues: Many panelists complained that they could not see the numbers on the Reckase Charts easily. They found transferring their ratings to the charts by hand to be tedious work. Some suggested that their ratings should be marked electronically. A few panelists suggested that the Reckase Charts should be distributed earlier in the process rather than prior to the second round of ratings. Some panelists thought that the charts emphasized the importance of item ratings over panelists' judgments. They suggested that the significance of the charts be moderated by describing them as only one of many sources of information panelists could consider when forming their judgments.

Adjustments: Although the charts themselves cannot be enlarged further, ACT will develop a computer presentation that will augment and display the Reckase Charts clearly during instructions. In addition, a computer program will be developed to electronically mark panelists' item ratings on the charts. Panelists will continue to connect their marked ratings by hand to assist them in identifying individual rating patterns. The charts will not be distributed until after the first round of ratings has been completed. The Reckase Charts will be presented as a form of feedback for the ALS meeting, rather than as a distinct step in the rating process.

CONCLUSIONS DRAWN FROM CIVICS 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 about the most useful feedback information provided, the highest average level of agreement was with the statement that Reckase Charts were the most useful.

- ✓ Pilot study panelists advised that the significance of the charts be moderated by describing them as only one of many sources of information panelists might consider when forming their judgments.

- ✓ Panelists urged that item ratings be marked on the charts electronically so that transferring item ratings by hand would be eliminated.

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 question 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 items to determine whether they had judged multiple choice items as being more or less difficult than constructed response items, relative to their overall cutscores and student performances at that level.

- ✓ Panelists could evaluate ratings for items in different content areas to determine whether they had judged content as being more or less difficult than another, relative to their overall cutscores and student performances at that level.

- ✓ 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 Civics 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 Differences Between Multiple Choice and Constructed Response Items

In general, the NAEP cutscores set for polytomous items are higher than those set for dichotomous items. While the Civics Pilot Study cutscores for the two item types were much closer after the initial round of ratings, polytomous item ratings still resulted in higher cutscores. Differences by item type were very small, however.

- ✓ Panelists in the Civics Pilot Study adjusted their ratings of polytomous and dichotomous items so that the differences in cutscores between the two types of items were reduced as the rounds of ratings progressed.
  - ▪ Grade 4 panelists tended to lower their polytomous ratings and raise their dichotomous ratings across the three rounds of ratings.
  - ▪ Grade 8 panelists generally raised their dichotomous ratings while keeping their polytomous ratings relatively constant.
  - ▪ Grade 12 generally raised their dichotomous cutscores from Round 1 to Round 2, and they also adjusted polytomous ratings to a slightly lower overall level.

THE ISSUE OF COGNITIVE COMPLEXITY

ACT has collected considerable data during the Civics 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 BEFORE FINAL CUTSCORES ARE DETERMINED

Panelists were given individual-level consequences data after the third round of ratings. They were provided with information about the consequences associated with each of their own cutscores and those of other panelists in their grade group. Further, rater location charts were modified to include additional data about the percentages of students scoring at or above points on the score scale, reported in increments of five points. Panelists recommended different cutscores, if they wished, to be used in computing the final cutscores. The final cutscores would be recommended to NAGB, unless some reason was found to do otherwise. This was a significant change.

- ✓ Two sets of cutscores would be available for NAGB's consideration.
  - ▪ Cutscores based on Round 3 ratings were not influenced by consequences data.
  - ▪ The final cutscores, based on panelists' recommendations were informed by consequences data.

- ✓ 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 CIVICS INTERIM REPORT: PILOT STUDY [1]

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

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 civics 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 civics pilot study was a final check on procedures to assure a successful operational Civics NAEP ALS.

RESEARCH CONDUCTED PRIOR TO THE CIVICS 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 Civics 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 civics 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 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 a method for collecting item-by-item ratings in the NAEP ALS Process. The ISSE method 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 it seemed likely that the feedback would also 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 civics 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 items with the ME method.[3] The ISSE method resulted in higher cutscores and lower percentages of students performing at or above the Proficient and Advanced levels than those from the ME method. The cutscores for the Basic level were approximately the same for the two methods. 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 inherent bias, further research using the ISSE method was discontinued, and it was eliminated as an alternative for implementation in the Civics 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. ACT planned to study whether item maps could be incorporated into the rating process as a means for panelists to adjust their cutscores without a third round of item-by-item ratings. In addition, ACT wanted to study the effect of providing consequences data to panelists during the rating process. ACT studied both the effect of item maps and the timing of providing consequences data on the outcomes of the ALS process.

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 new Reckase method (Reckase, 1998) as an alternative to the Mean Estimation method for setting cutscores. The extent to which item maps interfaced with an item-by-item rating method was evaluated with the Mean Estimation method as the item-by-item rating method coupled with the item mapping procedure.

Results of the second field trial indicated that panelists had little difficulty with either the item maps or the Reckase method. Both methods groups used the Mean Estimation method for the first two rounds of item-by-item ratings, and both groups marked their third round of cutscores directly on a chart or map. The interface between an item-by-item rating method and item maps seemed smooth, and the overall process had coherence. The critical role of the response probability in the item mapping procedure together with the fact that there was no clear choice of response probability to use led TACSS to recommend that ACT not consider item mapping as a procedure in the NAEP ALS process.

Panelists appeared to understand the Reckase Charts, and they were able to use the charts to select a cutscore for their final round of ratings with little apparent difficulty. Reckase Charts provided a means of evaluating the consistency of ratings for items along several important dimensions. The finding that panelists understood how to use the Charts and interpret their rating consistency represented an important and significant improvement in the NAEP ALS Process. TACSS recommended that ACT continue evaluating the use of Reckase Charts by incorporating them into the ALS process to be implemented for the pilot study.

[3] The NAEP Geography data were used to test the methods in the field trials for civics since civics data were not available. Item ratings collected during the 1994 NAEP Geography ALS were compared to those collected in the field trial using the ISSE method.

Findings regarding the effect of consequences data on the cutscores were consistent with evidence collected in earlier studies by ACT. There was no statistically significant difference between cutscores set by groups of panelists who were informed of the consequences of their ratings throughout the process (starting with feedback for the first round of ratings) and cutscores set by panelists who were not informed until after the last round of item ratings. Panelists appeared to value the consequences information, but there was not a significant, measurable impact on their cutscores.

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 in the process 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 civics pilot study, with the expectation that they would also be used for the civics ALS.

The purpose of the civics pilot study (PS) was to continue studying and refining the procedures for the ALS with particular attention focused on incorporating Reckase Charts into the process. The Reckase Charts were used as part of the Reckase method implemented in the second field trials for both civics and writing. (Please see Loomis, Hanick, Bay, & Crouse, 2000a & 2000b for a description of the field trials in which the Reckase method was implemented.) The decision was made to use Reckase charts in the pilot study as a step in the ALS procedure. Instructions in the use of Reckase Charts would be given in a general session convened after the standard feedback had been evaluated and discussed. Following instructions, Reckase Charts were distributed in grade groups where panelists marked them and evaluated their ratings relative to the data on the charts. The Reckase Charts were conceptualized as a step in preparation for the next round of ratings. The pilot study examined:

- how panelists reacted to the ALS process that included Reckase Charts;
- how cutscores that resulted from the final round of item ratings compared to cutscores computed after panelists received consequences feedback data; and
- 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 included various indicators and measures of the reasonableness of the cutpoints, standard deviations, interjudge consistency, and intrajudge consistency. Also included in the criteria were observations and panelists' evaluations of the process. Were instructions clear and concise? Did panelists feel confident in their ability to perform tasks? Did panelists feel satisfied with the outcomes of the process? Was there enough time (too much or too little) for each aspect of the process? Panelists' comments and responses to process evaluation questionnaires, which reflected their perception of each aspect and of the entire ALS process, were evaluated. A de-briefing session was conducted at the close of the pilot study with a representative sample of panelists, and facilitators and observers were de-briefed as well. Because the Reckase Charts were the most significant change in the procedure, there was particular interest in learning about 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 Civics Pilot Study. Please see Appendix B for additional details.

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. Three samples were drawn without replacement: one to identify nominators of teachers; one to identify nominators of nonteacher educators; and one to identify nominators of general public representatives[4]. A total of 260 public districts and 58 private schools were sampled. Please see Table 1 for the distribution of districts and schools sampled by nominator type. 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, 1997b).

Table 1
Distribution of Districts and Schools Sampled

Nominator Type	Public Districts	Private Schools	Total
Teacher	127	52	179
Nonteacher Educator	19	6	25
General Public	114	0	114
Total	260 82%	58 18%	318

NOMINATORS

Three separate samples of school districts were drawn to identify nominators. A separate sample of private schools was drawn to identify nominators of private school teachers and nonteacher educators. A total of 782 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.

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

Nominators of teachers were:

- district superintendents
- leaders of teacher organizations
- state curriculum directors (nominees throughout 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 (nominees throughout 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
- civic leaders, elected officials, employers of persons in a civics-related position or with a civics-related background

Table 2
Distribution of Nominators Contacted

Nominator Type	Public Districts	Private Schools	State	College/ Universities	Employers	Total
Teacher	224	52	43	0	0	319
Nonteacher	18	6	15	29	0	68
General Public	289	0	0	0	106	395
Total	531 68%	58 7%	58 7%	29 4%	106 14%	782

POOL OF NOMINEES

Nominees were to represent a specific grade perspective (4th, 8th, or 12th) and fill 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 whom they judged to be "outstanding" in their civics-related field for each grade. From the 782 persons who were contacted to serve as nominators, a total of 70 persons submitted nominees. They nominated a total of 329 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 (4th, 8th, or 12th) 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 initially 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 53 panelists participated in the pilot study representing 39 states and 1 U.S. territory. ACT did not strictly limit the number of panelists who were nominated by the same person, 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 four candidates were selected from the same nominator. For the grade 12 panel, three panelists were selected from the same nominator. For the grade 4 panel, there was a shortage of both male nominees (a typical problem for grade 4 panels) and nonteacher educators. Although the overall representation by region, gender, and race/ethnicity approached the targeted percentages across the three grade groups, balanced representation was lacking in specific grade groups. The demographic profiles for the nominee pool and the panels have been included in Appendix B.

The Achievement Levels-Setting Process

The civics pilot study lasted five days, August 13-17, 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 civics 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 Civics NAEP Framework* and achievement levels descriptions, and answered “content” questions. All content facilitators had participated in developing the Civics NAEP, and were trained for the ALS process. One content facilitator had served in the same capacity for the 1994 U.S. History NAEP ALS process They participated in a

[5] A list of ALS staff and observers has been presented in Appendix 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. A copy of 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 about 8 to 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.

ITEM RATING POOLS

The 1998 NAEP Civics data were used for the Civics Pilot Study. Two item rating pools for each grade were constructed so that they were as nearly equal as possible with respect to item difficulty, item content area, and item 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. Table 3 presents a summary of information describing items in the rating pool for each rating group.

Table 3
Description of Items in Each Item Rating Pool for Each Item Rating Group

Grade Group	Total Items	# Items in Content Area					Item Type*			Student Performance (P-values)			
		1	2	3	4	5	MC	SCR	XCR	Mean	SD	Min	Max
4A	60	13	9	10	7	21	46	9	9	53.4	19.6	18.0	90.8
4B	60	14	13	10	4	19	46	19	4	52.1	19.9	12.7	93.0
8A	94	13	21	28	13	19	77	13	4	49.1	17.2	16.0	89.4
8B	94	10	24	28	14	18	77	12	5	48.6	21.2	12.6	86.8
12A	95	11	16	29	17	22	77	15	3	52.1	16.0	21.3	90.0
12B	95	11	18	28	16	22	77	14	4	52.6	17.0	19.1	89.6

*MC = Multiple choice; SCR = short constructed response; XCR = extended constructed response

The grade 4 Civics NAEP consisted of 90 items divided into 6 blocks. Each block contained 15 items. Each rating group rated four blocks of items (60 items). Two item blocks in each rating pool were unique to each rating group and two blocks were in common with the rating pool of the other rating group for grade 4.

The grade 8 assessment consisted of 151 items divided into 8 blocks. Seven blocks contained 19 items and one block contained 18. Five blocks (94 items) were assigned to the item rating pool for each grade 8 rating group. Group A rated items from 3 of the 8 blocks, and group B rated items from a different set of 3 blocks. Both groups rated the same items from two common blocks.

The grade 12 assessment consisted of 152 items divided into 8 blocks with 19 items in each block. Five blocks (95 items) were assigned to the item rating pool for each grade 12 rating group. Group A rated items from 3 of the 8 blocks, and group B rated items from a different set of 3 blocks. Both groups rated the same items from two common blocks.

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 July 1, 1998, and contained materials that panelists were required to study. The second mailing was August 4, 1998 and contained detailed instructions related to travel arrangements and accommodations. The briefing materials and information included:

- 1998 NAEP *Civics Framework*;
- 1998 NAEP Civics Achievement Levels Descriptions;
- *Briefing Booklet* for 1998 Civics 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 Civics NAEP Framework*, and other such general information about NAEP and NAGB.

At the start of the second full day, a reduced version of the general orientation show was presented to review the outcomes of the process with panelists. 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 civics panelists too.

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. Panelists are urged to use their *Briefing Booklet* as an instructional tool and as a review guide for each session. The *Briefing Booklet* includes a sketch of each activity in each session, in the order that it occurs in the agenda. It describes the purpose of the activity and how it will be accomplished. The *Briefing Booklet* is included in Appendix E.

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.[6] 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.

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS

During a general session the morning of Day 2, the three content facilitators presented an overview of the *NAEP Civics Framework* and the ALDs as a general session.[7] 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 Civics 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. In one exercise, panelists used their understanding of the ALDs to determine the level of achievement that would be required of students to answer correctly and completely all items in one item block. This exercise was designed to help panelists become familiar with items of different types (i.e.,

[6] 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.”

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

multiple choice and constructed response) and to understand how the ALDs relate to all types of items. The exercise also helped panelists to become familiar with items that would not be included in their rating pools. Finally, it helped them to become familiar with the structure of NAEP item blocks and with scoring rubrics.

In another exercise, panelists applied their understanding of the ALDs more holistically. Rather than matching an individual item to an achievement level description, they were now being asked to match performance in an entire booklet to the description of one level of achievement. A sample of ten student assessment booklets was given to judges to review and discuss with respect to their understanding of the ALDs. Panelists were asked to determine if the performance exhibited in each booklet should be classified as Basic, Proficient or Advanced. After classifying each booklet independently, panelists discussed their classifications with each other. This exercise helped panelists to gain a better understanding of the ALDs and to become familiar with additional NAEP items and scoring rubrics. Discussions of the performances in booklets, relative to the ALDs, helped panelists to become more conversant with the ALDs and to internalize their meaning more completely.

UNDERSTANDING BORDERLINE PERFORMANCE AND PREVIEW RATING SESSION

After working with the ALDs throughout the morning and early afternoon, panelists were trained in the concept of borderline performance and instructed in rating items to reflect performance at the borderline. Panelists were trained to understand that borderline performance is the minimum performance qualifying for each level of achievement. That is, borderline performance is just at the lower cutscore for each level of achievement.

The general session included instructions for rating and marking item-rating forms in preparation for the Preview Rating Session. Panelists rated items in the two blocks that had been administered to them in the initial training exercise when they took the NAEP. They were already familiar with the items and with the scoring rubrics for the items. The Preview Rating Session was an addition to the process, and it was intended 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 their grade group. Panelists were informed that because this was a preview session, neither results nor feedback would be provided. For multiple-choice items, panelists estimated the probability that a student performing at the borderline of each achievement level would respond correctly. Panelists were instructed to think of this task as one of estimating the number of students who would give a correct response. They were told to think of a class of 100 students whose performance just met that of the ALD for a particular level, and estimate the number of correct responses from 100 borderline students. For constructed response items, judges estimated the mean or average score (e.g., 2.4 on a scale of 1-3) of students performing at the borderline of each level. They could again think of a classroom with 100 borderline students for each achievement level and estimate the average score for those students on each constructed response item.

Panelists struggled with the Preview Rating Session. They appeared to have difficulty with applying their concepts of borderline performance to performance. The achievement levels descriptions are statements of what students should know and be able to do. Panelists were just beginning to form a concept of the minimal level of performance that students should have. This exercise asked them to apply that concept to actual performance. Panelists had problems with the difference. They were annoyed by the absence of feedback.[8]

[8] Providing feedback for ratings collected at this early stage did not seem advisable. The purpose was not served by reporting rating feedback to panelists before they had gained a clear concept of borderline

Once the Preview Rating Session was concluded, panelists returned to the task of developing their concept of borderline performance. In grade groups, 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 (Day 3), 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.

PAPER SELECTION EXERCISE

Instructions in the Paper Selection Exercise followed the review and discussion of borderline performance on the morning of Day 3. The paper selection exercise required panelists to examine three student papers scored at each score point for all of the constructed response items in their item rating pool. Some of these items required short written answers (score range 1-3 points), while others required extended answers (score range 1-4 points). Although many student papers were included in this exercise (between 138 and 174, depending upon grade) most of the written responses were brief and could be read quickly.

Panelists were instructed to choose one student paper that represented performance at the borderline of each achievement level for each constructed response item in their rating pool. As a group, panelists reviewed and discussed paper selections for the constructed response items in one block before they independently reviewed the constructed response items in the remaining blocks. The papers selected to represent borderline Basic performance were marked with a yellow “flag;” those selected to represent borderline Proficient performance were marked with a blue flag; and those selected to represent borderline Advanced performance were marked with a red flag. After selecting a paper to represent borderline performance at each achievement level for all the constructed response items in one block, 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. If no paper was found to represent borderline performance, then no paper should be selected for that particular item at that particular level. The purpose of the exercise was to have panelists evaluate papers and determine whether any could be found to represent borderline performance. Training in borderline performance was accomplished whether or not a paper could be identified to represent borderline performance at each achievement level.

This training activity was designed to accomplish the following purposes:
- to provide a reality check on how students responded to open-ended questions;
- to promote a clear conceptualization of performance at the borderline;

performance. This exercise had been intended only to raise awareness. After discussion with TACSS, ACT decided to eliminate the Preview Rating Session for the Writing Pilot Study and to replace it with a demonstration of rating.

- to familiarize panelists with the scoring rubrics for constructed response items.

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. Panelists performed the rating tasks 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 item-by-item rating, was used. The rating method had been described in the orientation sessions of the first two days. Panelists had used the method for rating two blocks of items in the Preview Rating Session on Day 2. The procedure was reviewed in detail, and panelists were again instructed in marking their rating forms. Panelists were instructed to estimate the percentage of students at the borderline of each achievement level who would correctly answer each multiple choice item and the average score on each constructed response item of students performing at the borderline of each achievement level. They marked their rating forms accordingly. Rating forms included notations to help panelists in rating items of the different formats. A copy of a rating form has been included in Appendix C. Once trained, the panelists were ready for Round 1 rating.

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 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. No panelist needed more than about 2.5 hours to complete the rating process for Round 1. After completing their Round 1 ratings, panelists responded to a process evaluation questionnaire. They were asked to stay in the general vicinity until their item ratings were entered and verified. This completed Day 1.

FEEDBACK AFTER ROUND 1

Staff entered 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. 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 F.

Cutpoints

The cutpoints are computed from the combined ratings of all raters and all items 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 civics 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.)

The cutpoints for each grade-level are presented in the general session for all panelists to see. Cutpoints and standard deviations are presented on graphs for each grade and shared with everyone in the general session. Grade level graphs are distributed to each panelist in the grade groups, along with the other feedback data.

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 items 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 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 Basic level are expected to get 49% 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 item combinations resulting in different student performance and total points possible.

Whole Booklet Exercise

As part of Round 1 feedback, the panelists participate in a whole booklet exercise, which is an extension of providing whole booklet feedback. They are shown actual student booklets with scores near the cutpoints that had been set by Round 1 ratings. The booklets are the same form used for the training exercise "Taking a Form of the NAEP." Panelists evaluate booklets scored within 2% above or below the total possible points associated with each cutpoint. They are asked to examine the responses of the student to all items in the booklet as a whole and determine if the responses represent student performance expected at the lower borderline of Basic, for example. If they perceive a discrepancy between the expected performance and the observed performance in the booklets scored at the cutpoint, they discuss 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 2% 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 set 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 relative to other panelists in the grade group. They also check for panelists who set cutscores that are very close together for two levels 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 item. The proportion of students who gave the correct answer is listed as the actual "p-value" for each dichotomously scored item. The mean (average) score is reported for each polytomous item, 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 items are for all students who took the 1998 Civics NAEP. They do not indicate how easy or difficult the items are for students at different achievement levels.

Reckase Charts

After reviewing and discussing the feedback from Round 1, panelists met again in general session for instructions in the use of Reckase Charts. For the pilot study, the Reckase Charts were conceptualized as a step in the rating process. Their role was something more than an additional source of feedback. Training in the use of Reckase Charts and implementing that training were separated from the other types of feedback.

For each block of items in the item rating pool, panelists are given a Reckase Chart that indicates expected performance for students scoring at each score point on the ACT NAEP-like scale. Each column represents the range of IRT-based performance estimates for one assessment item. Each row represents IRT-based performance estimates for the items in one block for students scoring at 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 item in the grade-level item pool to the value associated with the highest asymptote. The score range for grade 4 is 39 to 273; for grade 8, the range is 13 to 301; and for grade 12, the range is 27 to 303. For dichotomous items, the probability of correct response (p-values) at each scale score point is reported for each item. For polytomous items, the expected score (mean) is reported for each item at each scale score. The expected performance across scale score points can be observed for each item, as can the expected performance across items for students scoring at a particular scale score. A sample Reckase Chart and instructions have been included in Appendix G. 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 also mark their individual item ratings from Round 1 onto 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 pens to distinguish the three achievement levels. Each block of items is printed on one chart page. Grade 4 panelists mark four charts and grade 8 and 12 panelists each mark five charts

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 the of ratings for multiple choice and constructed response items relative to a cutscore, or varying levels of consistency in item ratings with respect to a specific achievement level. They can also look for indicators of rater fatigue, such as less consistent ratings for items in the last block 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. They are reminded that their guide to item ratings is the description of what students should know and be able to do—not model-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 were informed that 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. Item ratings were again entered 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. The whole booklet exercise was omitted, however. Feedback was again distributed to panelists in grade groups where they could ask questions and discuss the results. Panelists transferred their ratings from round 2 onto the Reckase Charts and marked the charts a second time. (Please see Appendix F for feedback information based on the second round of 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.

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 by noon on Day 5.

FEEDBACK AFTER ROUND 3

Round 3 item ratings were again entered 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. Individual level consequences data were presented to inform the panelists about the percentages of students scoring at or above each cutscore for each grade. Lists were distributed with cutscores and consequences data associated with each. Data for each panelist in a grade group were listed and identified by secret identification codes. Panelists could look at the consequences of their own cutscores and of all other panelists in the grade group. In addition, rater location charts were modified for Round 3 feedback to include information about the distribution of student performances at or above score points on the ACT NAEP-like score scale. Panelists could review the percentages of students who would score at or above different score levels, and those data were reported for scores in increments of five points on the ACT NAEP-like scale. Together, these two pieces of feedback data provided panelists with considerable information for modifying their own cutscores.

Cutscores, standard deviations, and consequences data for each panelist in each grade were presented in general session. Paper copies were distributed to each panelist as well. 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 in order 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 M. The questionnaire items asked whether panelists would want to make changes to any of their 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 to be used as the standard for selecting exemplar items.

Consequences questionnaires were collected as rapidly as possible, and the recommended cutscores were entered to compute new grade level cutscores.

STEP 5: SELECTION OF EXEMPLAR ITEMS

After the panelists recommended their final cutpoints, they were trained in the process of selecting exemplar items for each achievement level. The final cutpoints were computed during this time, and used to prepare lists of exemplar items for review and selection by panelists when they returned to the grade groups.

Panelists in each grade group selected assessment items that they considered appropriate to illustrate student knowledge and skills associated with the description of each achievement level. The exemplar items are for use in reporting the NAEP results and are a primary outcome of the ALS process. The exemplar item lists were drawn from two item blocks at each grade. Exemplar items were selected from blocks of items selected for release to the public when the results of the 1998 Civics NAEP were reported. The goal of the exemplar selection process was to provide the maximum number of items to illustrate student performance at each NAEP achievement level.

Two statistical criteria guided the statistical selection of exemplars for review by panelists: item difficulty and item discrimination. The average conditional probability of correct response served as the indicator of item difficulty. To qualify as an exemplar, a multiple-choice item and constructed-response score needed at least a 50% average probability of correct response across the score interval of an achievement level. Each rubric score point for constructed response items was evaluated as if it were an item. Short constructed response items could appear on the list two times (once for each of two credited responses) and extended constructed response items could appear on the list three times. Items were “assigned” at the lowest achievement level for which this criterion was met. If the criterion were met for scores below the Basic score range, the items were eliminated from further consideration.

The items and response scores that met the first criterion were screened further for discrimination. The indicator of discrimination was the difference between the average probability of correct response across one achievement level, compared with that of the next lower level. To meet the discrimination criterion, the difference between the two levels must be at or above the discrimination level of 60% of the items in the entire grade-level item pool. The discrimination value was applied to the items in the blocks marked for release to determine which items qualified under that criterion.

Items that met the statistical criteria for difficulty *and* discrimination constituted the primary list of exemplars, whereas items that met the difficulty criterion *only* constituted the secondary list. Items were listed on a primary and secondary list for each achievement level for consideration as exemplars at the three grades.

Panelists determined whether each item on the lists would serve as an appropriate illustration of performance required at the specific achievement level, based on the achievement levels descriptions. From the list of items that satisfied the statistical criteria, panelists identified items and student response scores that matched the descriptions of student performance at each achievement level. They approved or *vetoed* each item and student response. The number of exemplars selected for each achievement level ranged from 1 to 9 items. The primary and secondary lists of exemplar items for each grade have been included in Appendix H. 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.

FINAL CIVICS PILOT STUDY WRAP-UP

Panelists gathered for the wrap-up session to complete the seventh process evaluation questionnaire and finish the last of the tasks related to consequences data.

Feedback After Recommendations for Final Cutpoints

The final grade group cutscores, based on panelists' recommendations, were used to compute the final consequences data. Data were represented graphically. Bar graphs were used to report the percentages of students performing at or above the cutscore for each achievement level, and pie charts were used to report the percentages of students performing within each achievement level score range. 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.[9] When the panelists completed the final questionnaire, they were thanked for their work and the meeting was adjourned.

OUTCOMES OF THE CIVICS PILOT STUDY

The civics 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 civics 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.

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

EVALUATION OF THE CUTPOINTS AND THEIR STANDARD DEVIATIONS

The cutscores[10] and their standard deviations have been included in Table 4. For all grades and levels, the cutscores *increased* from round to round during the item-by-item rating process, 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 of item-by-item ratings at each level for each grade. No patterns of statistically significant differences appeared when comparing cutscores by panelist type, grade, round of ratings, gender, region, or race/ethnicity. When comparing cutscores within grade by rating groups (groups A and B) and table groups, no major differences were noted. For a detailed report of the test results for group differences, please refer to Appendix I.

Table 4
1998 Civics 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	144.7	146.0	149.3	148.9
		SD	15.9	8.3	5.2	3.6
		%≥	**76.9%**	**73.7%**	65.7%	66.8%
	Proficient	Cutpoint	161.5	162.9	165.0	164.1
		SD	6.9	5.1	3.9	3.7
		%≥	**32.1%**	**28.6%**	23.7%	25.9%
	Advanced	Cutpoint	174.2	176.0	178.8	176.2
		SD	7.6	6.5	5.5	5.1
		%≥	**8.6%**	**6.7%**	4.5%	6.5%
8	Basic	Cutpoint	152.2	153.3	154.2	154.1
		SD	9.5	6.8	5.7	5.5
		%≥	**58.9%**	**55.0%**	52.2%	52.6%
	Proficient	Cutpoint	165.5	166.1	167.3	167.1
		SD	5.2	5.1	4.2	4.1
		%≥	**22.8%**	**21.3%**	19.1%	19.4%
	Advanced	Cutpoint	176.9	177.6	179.2	179.6
		SD	5.9	5.8	4.7	4.6
		%≥	**5.9%**	**5.3%**	4.2%	4.0%
12	Basic	Cutpoint	147.6	148.4	149.0	149.3
		SD	6.0	3.7	3.3	3.2
		%≥	**70.0%**	**68.3%**	66.5%	65.7%
	Proficient	Cutpoint	163.0	164.1	164.5	164.6
		SD	3.6	2.7	2.4	2.4
		%≥	**28.3%**	**25.6%**	25.0%	24.7%
	Advanced	Cutpoint	173.8	175.9	176.7	177.5
		SD	5.6	5.1	4.8	4.8
		%≥	**9.1%**	**6.7%**	6.1%	5.4%

Bold font represents data that were not presented to panelists.

[10] The data from one panelist were entered incorrectly for changes to Round 3 cutscores. These data were corrected after the ALS panels were adjourned. The corrected data have been used to calculate the cutpoints reported here as “final.”

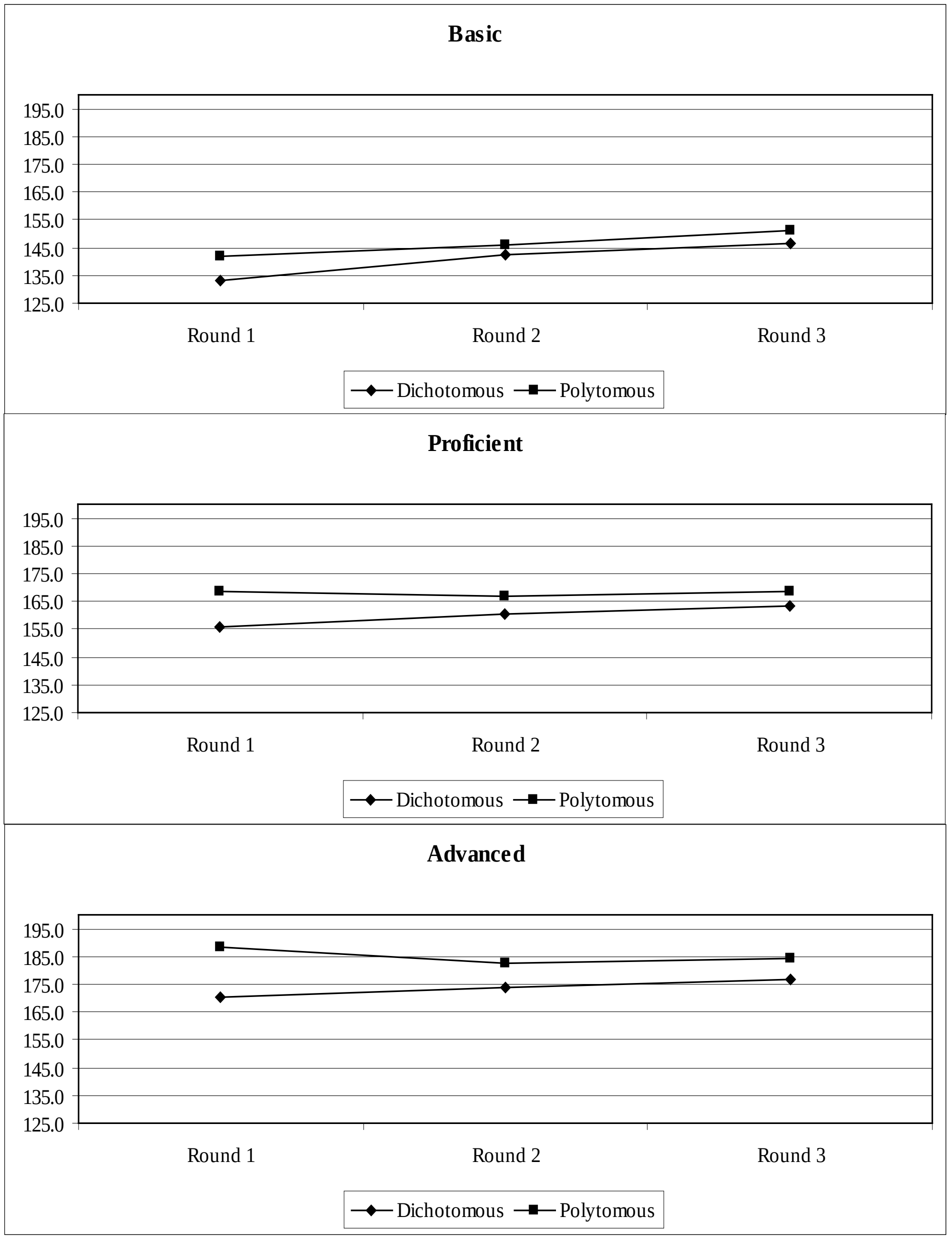
The variance associated with the Basic cutscores for Round 1 ratings generally was higher than that for the other cutscores. This is a typical pattern usually found in NAEP ALS data. In particular, Grade 4 judges demonstrated greater variability in their Basic ratings for Round 1 than grade 8 and grade 12 panelists. Panelists seem to experience relatively more difficulty in forming a clear concept of borderline Basic performance. This is evidenced by the tendency for relatively higher standard deviations of the Basic cutscores for all rounds of ratings for grade 4 and 8 panelists, and by panelists' responses to questions regarding their concept of borderline performance at each achievement level. Perhaps this difficulty stems from the fact that there is no definition of performance below the Basic level, so forming a concept of borderline Basic performance is relatively more difficult. Grade 12 panelists showed higher variability at the Advanced level. By Round 3, however, the standard deviation was low for *all* levels and grades.

COMPARISON OF CUTPOINTS BY ITEM TYPE

Differences in cutscores by item types—multiple-choice (dichotomous) items and constructed response (polytomous) items—have been an on-going interest in the NAEP ALS process. In general, the cutscores set for polytomous items are higher than those set for dichotomous items. Grade 8 and 12 panelists for the Civics Pilot Study set statistically significantly higher cutpoints for polytomous items than dichotomous items across achievement levels for all rounds of ratings. Grade 4 panelists, however, did not. Their ratings for dichotomous and polytomous items were *not* significantly different for Rounds 1 and 2. The standard deviations for their polytomous item ratings were quite high for those two rounds. Grade 4 raters who set their dichotomous cutscores higher than their polytomous cutscores produced these results, in part. The grade 4 Basic cutscores for polytomous and dichotomous items set by individual panelists do not *appear* to be more similar than those set by grade 8 and 12 individual panelists. Nor do these cutscores appear to be more similar than those for the Proficient and Advanced achievement levels. Detailed analyses of the cutpoints by item type are presented in Appendix J.

Panelists adjusted their ratings of polytomous and dichotomous items so that the differences in cutscores between the two types of items were reduced as the rounds of ratings progressed. The following figures (Figures 1-3) show cutpoints for each achievement level, across rating rounds, for each grade panel. Additional graphics (Figures 1- 18) in Appendix J reveal the pattern graphically, showing how the ratings for dichotomous and polytomous items moved closer together over rounds of ratings. Grade 4 panelists tended to lower their polytomous ratings and raise their dichotomous ratings across the three rounds of ratings. Grade 8 panelists generally raised their dichotomous ratings while keeping their polytomous ratings relatively constant. Like grade 8 panelists, grade 12 generally raised their dichotomous cutscores from Round 1 to Round 2. Unlike grade 8 panelists, they also adjusted polytomous ratings to slightly lower their cutscores. While the cutscores for the two item types were much closer after the initial round of ratings, polytomous item ratings still resulted in higher cutscores. Differences by item type were very small.

Figure 1: Grade 4 Cutpoints Averaged Across Panelists, by Item Type

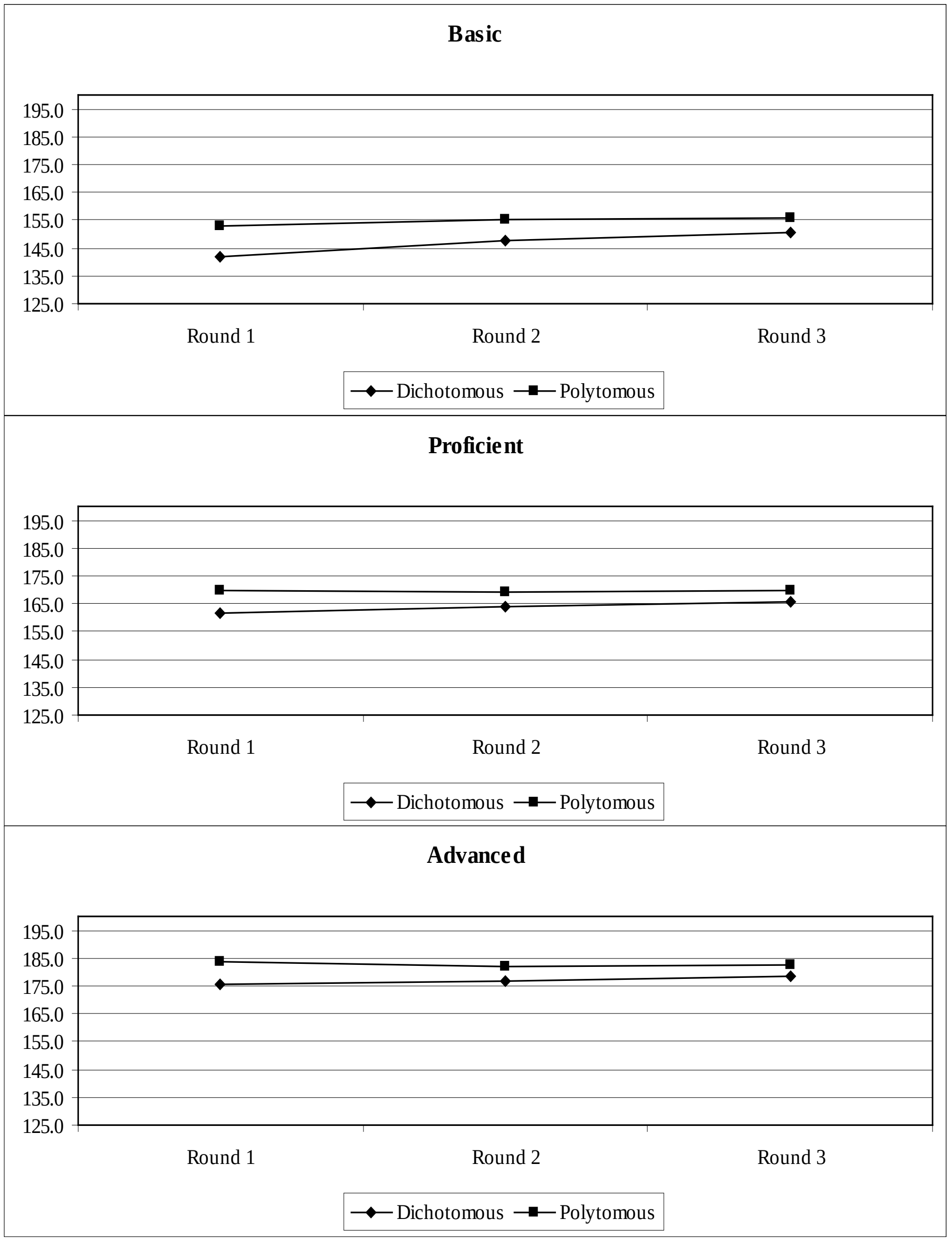


Basic

Proficient

Advanced

Figure 2: Grade 8 Cutpoints Averaged Across Panelists, by Item Type

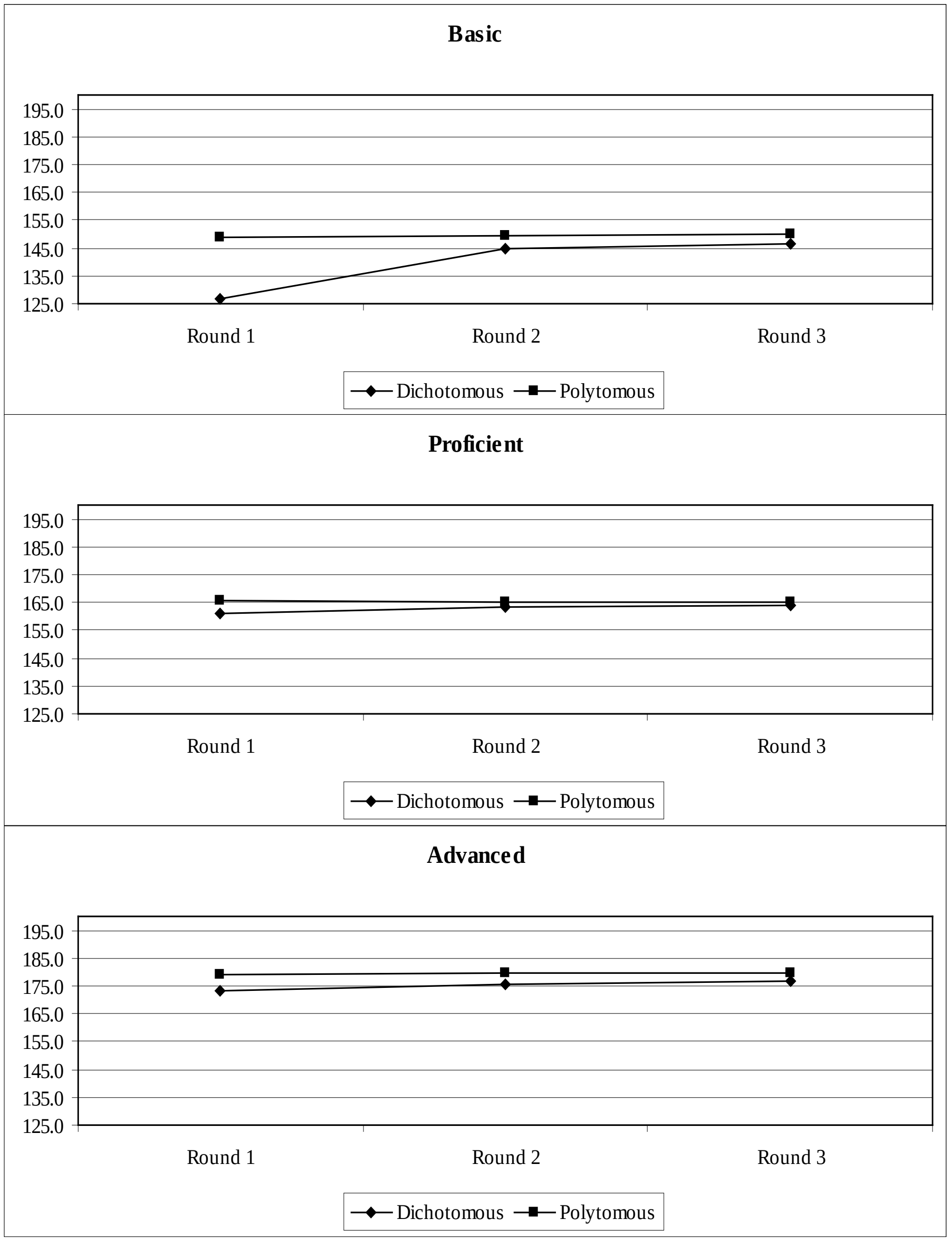


Basic

Proficient

Advanced

Figure 3: Grade 12 Cutpoints Averaged Across Panelists, by Item Type



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. After reviewing the feedback presented following Round 1, Civics 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. Several bar graphs show the percentages of items for which ratings were raised, lowered, and unchanged from one round to the next, by grade level (Figure 4) and the magnitude of average rating changes from one round to the next for different types of items, by grade level (Figure 5). Detailed analyses of rating changes are presented in Appendix K.

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

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	63	57	54	18	20	23
	B	39	32	30	10	13	18
8	A	30	24	22	11	12	11
	B	45	36	36	13	16	16
12	A	42	29	34	18	16	11
	B	48	42	46	19	17	16

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

		Raise			Lower		
Grade	Group	Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	53	52	55	5	6	10
	B	36	31	37	13	4	4
8	A	17	17	21	4	3	3
	B	31	28	28	3	3	3
12	A	15	14	19	4	3	4
	B	22	21	27	7	5	7

Figure 4
Average Percentage of Items for Which Ratings Were Raised, Lowered, or Unchanged Across Rounds

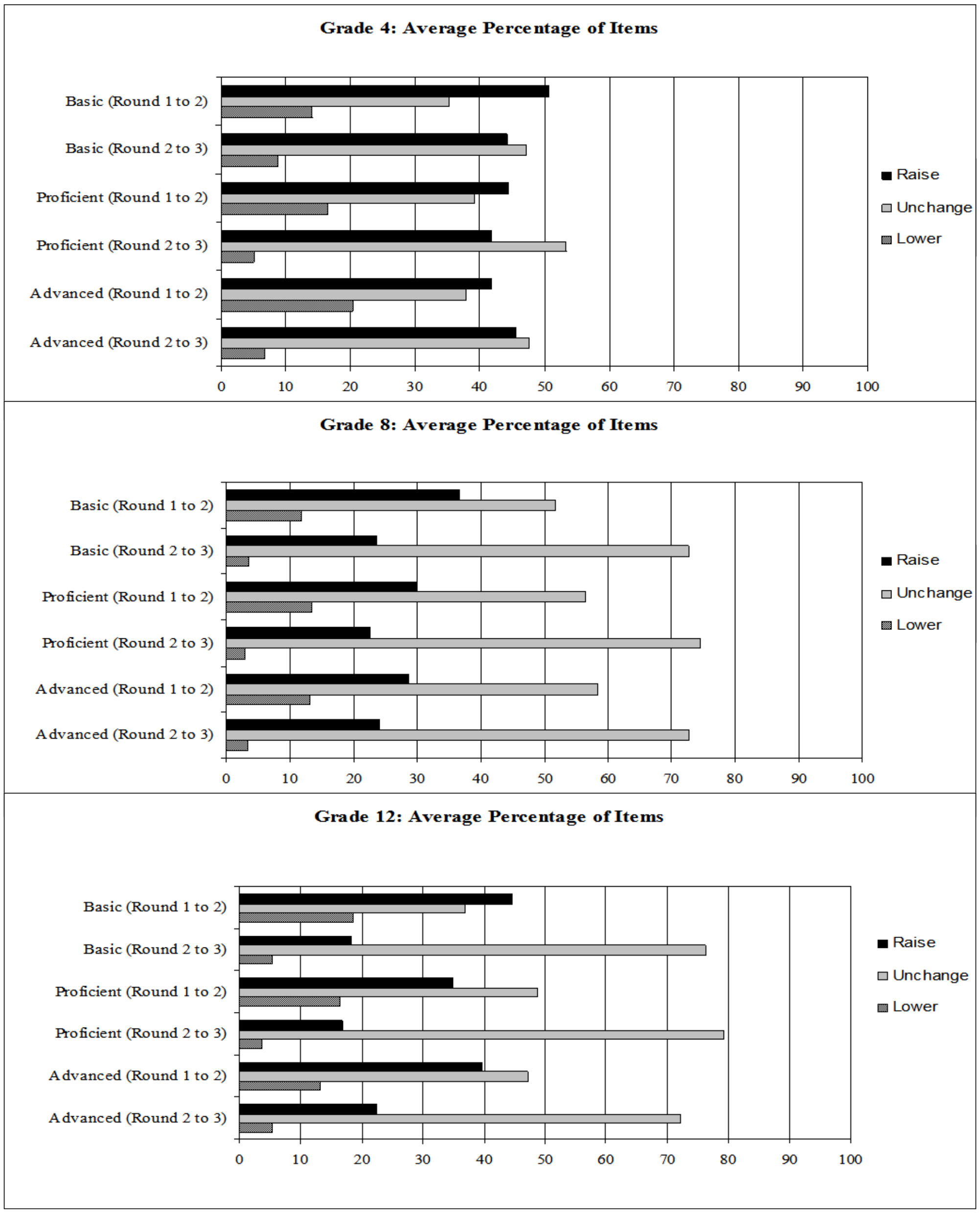
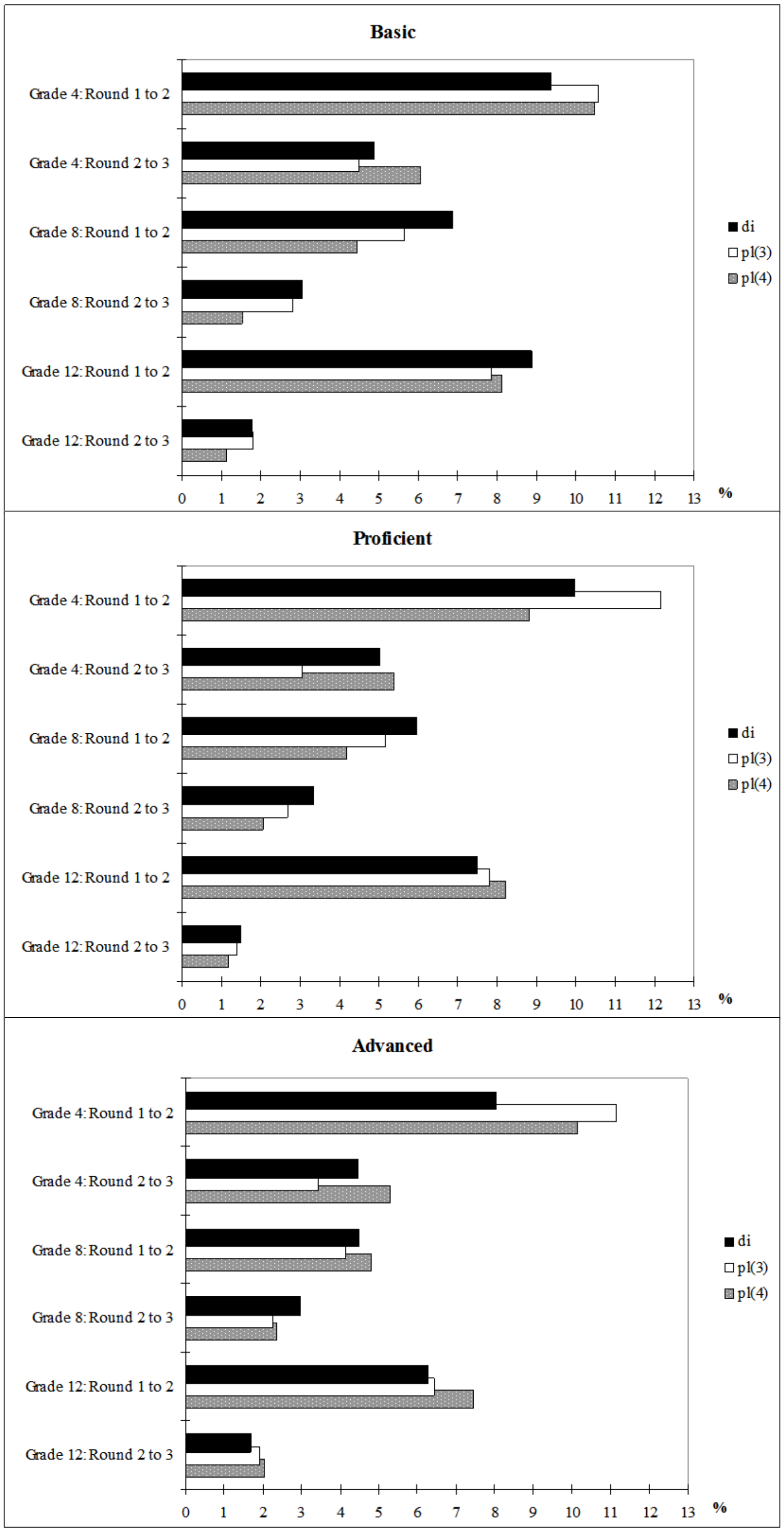


Figure 5
Magnitude of Average Rating Change



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. For the Civics Pilot Study, some judges changed almost every item rating from round 1 to round 2. Grade 4 panelists changed the largest proportion of item ratings from Round 1 to Round 2, and from Round 2 to Round 3. The changes were predominantly increases. Further, across all grades and achievement levels, panelists tended to raise their ratings for more items than they lowered them.

For all grades and all levels, the magnitude of average rating changes was greater between Rounds 1 and 2 than between Rounds 2 and 3. Grade 4 panelists made larger changes in their ratings for both polytomous and dichotomous items than panelists for grade 8 and grade 12. Grade 8 panelists made larger change in their ratings for dichotomous items than polytomous items at the Basic and Proficient levels. Grade 12 panelists made changes of about equal magnitude in their ratings for polytomous and dichotomous items. No noteworthy differences appeared when comparing changes in ratings by rating groups (group A and B) or by table groups.

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 Civics Pilot Study panelists as a step in the ALS process. Panelists transferred their ratings and 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 item type or category. Panelists repeated this analysis after Round 2.

It was anticipated that panelists who reported that the Reckase Charts were influential in forming their judgments would change their ratings to be similar to the chart values that were associated with the panelist's individual cutscores, or the grade level cutscores. Relationships between the use of Reckase Charts and rating patterns were explored with correlation analyses and t-tests. Rating patterns were examined by computing the differences between the individual judge's cutscores that resulted from his/her ratings from round to round (within-rater deviations), and the differences between individual cutscores and grade group cutscores. The measure of rater deviation from the group cutscore was the absolute value of the differences between the raters' cutscores and their group cutscores from the previous round. The extent to which the Reckase Charts influenced panelists was indicated by judges' responses to questions about the charts on the process evaluation questionnaires. Appendix N contains a summary of the results of these analyses.

Results of the analyses did not reveal a clear pattern. ACT anticipated that panelists would adjust their item ratings to be similar to the conditional p-values associated with either the rater location of the panelist or the grade level cutscore for panelists who reported that the Reckase Charts were important and influenced their judgment. Ratings for panelists at grades 8 and 12 generally followed expected patterns. For instance, judges for grades 8 and 12 who reported the charts to be *more important* showed smaller deviations from the group cutscores set in the previous round. Grade 4 judges, however, showed an opposite pattern. The more important the raters found the Reckase Charts, the more their cutpoints deviated from group cutpoints set in previous rounds.

Sample sizes were relatively small, and few statistically significant results were found. Those that were found generally supported the observations described above. Positive correlations between Reckase Chart *importance* and deviations from group cutpoints were found for grade 4 judges. Negative correlations were found between *importance* and group cutscore deviations for grade 8 judges. Grade 12 correlations were negative and generally non-significant. The exception for grade 12 was the correlation between *overall importance* and Basic level ratings for Round 3. These analyses have been summarized in Appendix L.

Relationships between the *importance* of the Reckase charts and within-rater variation were also examined. Grade 4 panelists' *perceived importance* of the charts was inversely related to the amount of within-rater deviation, although correlations were non-significant. Grade 8 judges' deviations, on the other hand, increased as *perceived importance* increased. Grade 12 judges' correlations were negligible and inconsistent.

T-tests were calculated using dichotomized *perceived importance* variables (charts were important/not important). These tests generally yielded outcomes that paralleled the results of the correlation analyses. A few unique pieces of information were found. For instance, Grade 12 panelists who considered Reckase Charts the most important or useful type of feedback had significantly lower deviations from group cutscores at the Proficient achievement level for Round 3 (p<.01).

ACT was concerned that the panelists would be more influenced by the Reckase Charts than by the achievement levels descriptions and other feedback data. Negative correlations between *perceived importance* of the Reckase Chart and deviations from group cutscores for grades 8 and 12 suggest that judges who found the charts important might have targeted their ratings to match group cutscores. The positive correlations between *perceived importance* and within-rater deviations support this possibility. If a rater is highly influenced by the Reckase Chart, his/her ratings might vary considerably to match the group cutscores. Unexpectedly, grade 4 judges exhibited a relationship opposite to those of grade 8 and grade 12 panelists. As *perceived importance* of the charts increased, deviations from the group cutpoints also increased.

Most raters appeared to understand how to use the charts, as indicated by mean *understanding* scores above 4.0 on a five-point scale. The majority of raters in all grades found the Reckase Charts to be the *most useful* piece of information. Raters also gave high ratings to the *importance* of the Reckase Charts, with means ranging from 3.75 to 4.1 on a five-point scale. These results should not be overinterpreted, however, because of the small sample size, the characteristics of grade group panels, and the restricted range for some variables.

EVALUATION OF CONSEQUENCES DATA

For the Civics Pilot Study, consequences data was introduced before collecting the final round of ratings. Panelists were told the percentage of students performing at or above their own cutscores for each achievement level as feedback from Round 3 ratings. They had the opportunity to adjust their cutpoints in response to those data. Comments were collected from panelists regarding their reactions to and opinions about the consequences of their cutscores.

When asked what they considered when making their cutscore recommendations, most panelists indicated that consequences data had little impact on their recommended cutscores. Although many panelists were quite concerned about the unexpectedly low performance of students relative to their cutscores, they generally seemed unwilling to make changes in their cutscores. They wanted to make some adjustments, but were reluctant to do so because they were satisfied with their item-by-item ratings and judgments. Several voiced concerns regarding the seemingly arbitrary nature of recommending cutscores, in contrast to the methodical collection of judgments during the item-by-item rating rounds.

In general, the effects of giving panelists individual consequences data appeared to be consistent with previous ACT research. That is, the data appeared to have 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, 38 of the 52 panelists (73%) answered "yes" and 14 panelists (26%) said "no." All 14 panelists recommended changes to the cutscores. Table 7 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 lower the cutscores for grade 4 at all levels, and to slightly raise the cutscores for grade 12 at all levels. For grade 8, the cutscores for Basic and Proficient were slightly lowered while the cutscore for Advanced was slightly raised.

Table 7
Number of Changes To Cutscores in Response to the Consequences Data Questionnaire, by Grade Groups

	Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)
Data Reflects Your Expectations?			
Yes	6	16	16
No	10	3	1
No Response	0	0	0
Recommend Changes to Cutscores			
Basic			
Raise	2	1	0
Lower	4	0	0
Proficient			
Raise	2	2	0
Lower	4	0	0
Advanced			
Raise	6	1	1
Lower	4	1	0
Recommend to NAGB			
Grade Cutscore as Set	13	14	14
Grade Cutscores Changed	2	2	1
Uninterpretable/No Response	1	3	2

Consequences data were computed again, based on the cutscores recommended in response to Round 3 consequences feedback. 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, 43 of 50 panelists (86%) answered "yes" and 7 panelists (14%) said "no." Two judges did not respond. Results of the recommendations have been presented in Table 8. As those data show, 4 of the 14 recommendations were to lower the cutscores for grade 8 Basic. Otherwise, the final recommendations were fairly evenly distributed between lowering and raising the cutscores across grades. These responses were collected to document panelists' evaluations of the final cutscores. No adjustments were actually made to the cutscores.

Table 8
Number of Changes Made to Cutscores in Response to the Consequences Data Questionnaire #2 by Grade Groups

	Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)
Data Reflects Your Expectations?			
Yes	14	14	15
No	2	4	1
No Response	0	1	1
Recommend Changes to Cutscores			
Basic			
Raise	1	0	0
Lower	1	4	0
Proficient			
Raise	1	0	0
Lower	1	1	1
Advanced			
Raise	1	0	0
Lower	1	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 Civics Pilot Study.

The responses of civics PS panelists have been presented for comparison with those of panelists for the geography and U.S. history ALS meetings. In general, the pilot study panelists were positive in their evaluation of the ALS process and their experience as participants. The responses of panelists to the process evaluation questionnaires have been presented by grade and by panelist type in Appendix N.

Understanding the Rating Process and Confidence in Ratings

Data reported in Table 9 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 4.68 – 4.94). 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 1.28 points, grade 8 increased 1.24 points, and grade 12 increased 1.41 points).

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.

Table 9
Civics Pilot Study Evaluation Questionnaires
Summary of Responses to Questions Related to Ratings

Questions	Round	Civics Pilot Study			Geography			U.S. History		
		Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)	Grade 4 (n=26)	Grade 8 (n=25)	Grade 12 (n=26)
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	4.19	4.21	4.05	3.75	3.89	3.76	4.46	4.10	3.96
	2	4.38	4.58	4.25	4.79	4.75	4.55	4.70	4.72	4.67
	3	4.94	4.68	4.81	4.81	4.57	4.66	4.63	4.68	4.48
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	4.19	4.32	3.82	3.86	4.07	4.14	4.29	4.10	4.89
	2	4.38	4.58	4.44	4.82	4.75	4.62	4.63	4.72	4.58
	3	4.81	4.79	4.75	4.81	4.71	4.83	4.61	4.69	4.59
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; 1=Far too Short; 3=Just Right)	1	3.13	3.37	3.13	2.93	2.93	3.21	3.64	3.34	3.11
	2	3.31	3.37	3.31	3.36	3.39	3.55	3.50	3.62	3.30
	3	3.47	3.56	3.65	3.44	3.68	3.55	3.64	4.21	3.44
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	3.19	3.26	3.18	3.21	3.14	3.62	4.54	4.14	4.07
	2	3.81	3.95	4.06	4.29	4.18	4.21	4.12	4.32	4.12
	3	4.47	4.50	4.59	4.48	4.32	4.52	4.39	4.48	4.11
5. The method for rating multiple-choice items was conceptually clear. (5=Totally Agree; 1=Totally Disagree)	1	4.19	4.26	3.81	4.07	4.00	3.76	4.11	3.97	4.22
	2	4.25	4.05	4.00	4.18	4.32	4.24	4.33	4.43	4.52
	3	4.50	4.21	4.59	4.44	4.36	4.4.1	4.52	4.43	4.59

Table 9 (continued)

Questions	Round	Civics Pilot Study			Geography			U.S. History		
		Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)	Grade 4 (n=26)	Grade 8 (n=25)	Grade 12 (n=26)
6. The method for rating multiple-choice items was easy to apply. (5=Totally Agree; 1=Totally Disagree)	1	4.00	4.00	3.56	3.82	3.71	3.83	3.89	3.83	3.89
	2	4.07	3.95	4.00	4.18	4.14	4.21	4.19	4.32	4.41
	3	4.38	4.26	4.56	4.33	4.14	4.38	4.44	4.32	4.44
7. The method for rating constructed-response items was conceptually clear. (5=Totally Agree; 1=Totally Disagree)	1	3.94	3.84	3.63	3.82	3.89	3.76	4.04	3.83	3.96
	2	4.00	3.74	3.81	4.25	4.25	4.10	4.33	4.17	4.22
	3	4.25	4.11	4.41	4.30	4.32	4.38	4.48	4.39	4.44
8. The method for rating constructed-response items was easy to apply. (5=Totally Agree; 1=Totally Disagree)	1	3.88	3.63	3.44	3.50	3.70	3.90	3.79	3.48	3.59
	2	4.13	3.74	3.94	4.04	3.89	4.03	4.11	4.07	3.89
	3	4.13	3.95	4.25	4.26	4.07	4.28	4.19	4.14	4.33

Table 10
Civics Pilot Study Evaluation Questionnaires
Summary of Responses to Questions Related to Achievement Levels Descriptions

Questions	Round	Civics Pilot Study			U. S. History ALS			Geography ALS		
		Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)	Grade 4 (n=26)	Grade 8 (n=25)	Grade 12 (n=26)	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)
1. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ 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	3.56	3.58	3.53	4.18	3.69	3.85	3.79	3.71	3.83
	2	4.13	4.05	4.06	4.44	4.21	4.35	4.29	4.11	4.21
	3	4.38	4.50	4.35	4.67	4.18	4.58	4.52	4.18	4.38
2. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ set of ratings my conception of Borderline Basic performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.25	3.37	3.29	3.71	3.55	3.78	3.50	3.25	3.59
	2	4.06	4.00	3.94	4.00	4.07	4.15	4.21	4.11	4.17
	3	4.47	4.44	4.41	4.52	4.14	4.38	4.56	4.18	4.38
3. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ 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	3.69	3.63	3.47	4.07	3.69	3.92	3.86	3.68	3.83
	2	4.00	4.16	4.13	4.37	4.19	4.33	4.33	4.11	4.21
	3	4.38	4.50	4.31	4.67	4.29	4.58	4.56	4.29	4.38
4. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ set of ratings, my conception of Borderline Proficient performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.31	3.42	3.29	3.64	3.55	3.78	3.46	3.21	3.55
	2	4.13	3.95	4.00	4.15	4.04	4.23	4.22	4.14	4.21
	3	4.50	4.44	4.41	4.59	4.29	4.46	4.63	4.18	4.38
5. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ 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	3.94	3.63	3.53	4.07	3.83	3.77	3.93	3.79	3.97
	2	4.19	4.11	4.13	4.48	4.14	4.31	4.44	4.11	4.17
	3	4.44	4.56	4.41	4.67	4.32	4.50	4.56	4.32	4.38
6. At the time I provided the $1^{st}/2^{nd}/3^{rd}$ set of ratings, my conception of Borderline Advanced performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.50	3.37	3.53	3.71	3.69	3.70	3.54	3.36	3.79
	2	4.13	3.89	4.00	4.15	4.10	4.23	4.22	4.07	4.17
	3	4.44	4.44	4.47	4.54	4.26	4.42	4.59	4.18	4.38

Understanding of the Achievement Levels Descriptions and Borderline Performance
The typical response pattern that has emerged from past ALS meetings was present for the Civics Pilot Study: achievement levels descriptions were generally better understood than the borderline descriptions. Panelists' understanding of both categories of performance increased over rounds so that the difference between the two diminished by Round 3. 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 4.31 to 4.56 by Round 3. Their conception of borderline performance approached *very well formed* for all achievement levels by Round 3, with the means ranging from 4.4 to 4.5. Table 10 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.

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. The civics pilot study panelists wrote borderline descriptions and modified them throughout the process. In contrast, the geography and U. S. history panelists discussed their concept of borderline performance with other panelists and used it in training exercises prior to rating items. Recall that Table 10 includes the responses of geography, U. S. history and civics pilot study panelists to questions about their understanding of achievement levels descriptions and their concepts of borderline performance. It was anticipated that the civics pilot study panelists would respond more positively when asked about their understanding of borderline performance than their understanding of ALDs, given that they wrote borderline descriptors but did not modify ALDs. Results indicated, however, that there were few differences between the two responses to these questions. Furthermore, when compared with responses for geography and U.S. history, the civics pilot study panelists' understanding of student performance at the borderline and at the three achievement levels was nearly the same. Civics pilot study panelists who wrote descriptions of borderline performance and discussed them did not reveal a clearer concept than geography and U.S. history ALS panelists who only discussed their concept of borderline performance.

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 = 4.56; grade 8 = 4.15; grade 12 = 3.94; 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 11). 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 11

	Grade 4	Grade 8	Grade 12
Most useful type of feedback information reported after Round 3 ratings			
p-value data	3.25	3.26	2.94
Rater location	3.75	3.21	3.44
Whole booklet	2.81	2.26	2.06
Reckase Charts	4.37	4.21	3.87

Comments about the Reckase Charts

Panelists' comments about the Reckase Charts were overwhelming positive, although participants offered many suggestions to simplify marking the charts and to improve instructions for using the charts. They indicated that although circling their ratings on the charts was a tedious task, the process of connecting their ratings to reveal the patterns in ratings was a highly effective learning tool. They strongly recommended that the ratings be marked electronically on the charts so that they only needed to connect them to examine the patterns. They suggested enlarging the numbers on the charts and printing them on white paper so the colored markings for each grade would be clear. A summary of panelists' diverse comments on these topics has been included in Appendix N.

RATINGS FOR MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEMS

In 1992 ACT found that the cutscores that would be set for multiple choice items were different than those for constructed response items scored for partial credit. Over the years, panelists have been asked for their input with regard to some possible reasons for this difference. Table 12 displays the data from the Civics Pilot Study panels related to this issue.

Table 12
Summary of Responses to Questions Related to
Multiple Choice/Constructed Response

		Civics Pilot Study		
Question	Round	Grade 4 (n=16)	Grade 8 (n=19)	Grade 12 (n=17)
1. If the ratings of student performance on multiple-choice items and constructed-response items are very different, this is most likely caused by different student behavior and performance on the items. 5=Totally Agree 1=Totally Disagree	1	3.53	3.21	3.06
	2	3.50	3.37	3.40
	3	3.64	3.53	3.80
2. If the ratings of student performance on multiple-choice items and constructed-response items are very different, this is most likely caused by the different rating methods. 5=Totally Agree 1=Totally Disagree	1	3.44	3.26	3.07
	2	3.20	3.00	2.94
	3	3.33	3.05	3.31
3. I think constructed-response items assess dimensions of knowledge and skills that are significantly different from those assessed by multiple-choice items. 5=Totally Agree 1=Totally Disagree	1	4.37	4.10	3.81
	2	4.37	4.16	4.18
	3	4.37	3.95	4.29

In general, panelists indicated a rather neutral response (3=half way between totally agree and totally disagree) when asked about the different rating methods (question 2) as the source of any differences in ratings for dichotomous and polytomous items. There was little agreement that the differences resulted from the rating methods. Panelists tended to agree that the source of any differences in their ratings was due to the fact that constructed response items assess dimensions of knowledge and skills that are significantly different from those assessed by multiple-choice items (question 3).

As has been the case in previous studies, panelists showed some changes in their responses to questions about dichotomous and polytomous items across rounds. As rounds of ratings progressed and more feedback data were evaluated, more panelists responded that differences in ratings for the two types of items were likely to result from student behavior and student performance on the items (question 1).

THE OVERALL ALS PROCESS

Data reported in Table 13 show the average responses to questions from the final questionnaire about the overall ALS process used for the Civics Pilot Study. Once again the responses indicate a generally positive reaction to the process. **All** of the judges were willing to sign a statement recommending the use of the achievement levels that resulted from the ALS procedures. The level of confidence in their ratings for the grade 4 panelists seemed noticeably lower than that for all other panelists, including Geography and U.S. History panelists (grade 4 mean = 3.8; all others between 4.0 and 4.4). Interestingly, the grade 4 Civics Pilot Study panelists responded most positively when asked about the ALS process producing levels that were defensible and reasonable. Also, grade 4 panelists were most willing to recommend the use of the achievement levels. Grade 4 panelists' lower confidence level in their ratings apparently did not undermine their enthusiasm for the overall ALS process.

In general, the Civics Pilot Study panelists tended to respond slightly less positively than ALS panelists for Geography and U.S. History to questions about the effectiveness of the ALS process and the opportunity to use their best judgment. Perhaps providing the Reckase Charts introduced information that challenged the panelists' judgments, therefore reducing the extent to which panelists used their best judgment in rating items. Reviewing the frequencies of responses to these questions indicated that the lower mean reflected only one negative response (response code 1 or 2) for each question from a grade 8 panel member. It would appear that nearly all of the panelists were satisfied with the effectiveness of the ALS process and the opportunity to use their best judgments, even though their mean scores were slightly lower for these questions than those of Geography and U.S. History ALS panelists.

Table 13
Panelists' Responses to Final Process Evaluation Questionnaire
Comparison of Civics Pilot Study to Geography and U.S. History ALS Responses

Questions		Civics Pilot Study			Geography			U.S. History		
		Grade 4 (n=16)	Grade 8 (n=18)	Grade 12 (n=16)	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)	Grade 4 (n=26)	Grade 8 (n=25)	Grade 12 (n=26)
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)		3.81	4.06	4.19	4.41	4.14	4.14	4.21	4.12	4.00
2. I would describe the effectiveness of this achievement levels-setting process as: (5=Highly Effective; 1=Not at all Effective)		4.00	3.67	3.88	4.04	3.75	4.11	3.75	3.74	3.92
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 Geography Assessment: (5=To a Great Extent; 1=Not at All)		4.19	4.17	4.00	4.30	4.29	4.21	4.14	4.28	4.15
4. I feel that his NAEP ALS process produced achievement levels that are defensible: (5=To a Great Extent; 1=Not at All)		4.31	4.17	4.25	4.11	4.18	4.14	3.61	3.96	4.14
5. I feel that this NAEP ALS process produced achievement levels that will generally by considered reasonable: (5=To a Great Extent; 1=Not at All)		4.31	4.00	4.25	4.19	4.29	4.21	3.75	3.64	3.84
6. I would be willing to sign a statement (after reading it, of course) recommending the use of achievement levels resulting from this ALS procedure:	Yes, definitely	75%	50%	50%	53.6%	60.7%	57.1%	17.9%	31.0%	51.9%
	Yes, probably	25.0	50.0	50.0	35.7	32.1	42.9	53.6	44.8	44.4
	No, probably not	0.0	0.0	0.0	7.1	7.1	0.0	21.4	6.9	0.0
	No, definitely not	0.0	0.0	0.0	0.0	0.0	0.0	7.1	0.0	0.0

EVALUATION OF THE SELECTION OF EXEMPLAR ITEMS

One of the primary outcomes of the NAEP ALS meeting is the identification of assessment items that illustrated the knowledge and skills associated with each achievement level to use in reporting NAEP results. Appendix H includes the items selected by pilot study panelists to serve as exemplar items for reporting the NAEP achievement levels.

Panelists reviewed and discussed the assessment items and responses that qualified statistically for consideration as exemplars. (Please see Table 14.) In general, there was a mix of both multiple choice and constructed response items for panelists' consideration in the exemplar item selection process. The exception was at the grade 4 Basic level where no constructed response item qualified for consideration as a Basic level exemplar.

Panelists were instructed to veto items or response scores that met statistical criteria but that did not meet substantive, content-related criteria. Judges were trained in the statistical criteria that had been used for selecting the items (described earlier in this report). In addition, panelists were instructed to use their knowledge of the achievement levels descriptions to evaluate each item on the list in terms of its quality as an illustrative or exemplar item. Panelists were to select items from the secondary list only if fewer than three of the items on the primary list were acceptable.

Table 14
Number of Assessment Items that Qualified and Were Selected as Exemplars for Civics Pilot Study

	Number Primary List*	Number Secondary List**	Number Selected
Grade 4			
Basic	7	4	6
Proficient	5	6	5
Advanced	4	2	5
Grade 8			
Basic	11	3	8
Proficient	11	4	9
Advanced	4	5	7
Grade 12			
Basic	10	8	7
Proficient	7	9	6
Advanced	3	3	3

*Items that met the statistical criteria for difficulty and discrimination.
**Items that met the difficulty criterion only.

The item blocks marked for release were identified in advance and used as common blocks for rating pools in each grade. All panelists rated these items, which were part of the grade-group discussions to train panelists in the paper selection exercise that was implemented prior to the first round of item ratings. Recall that in the Paper Selection Exercise, panelists were asked to select papers to represent the performance of students at the borderline of each achievement level. Thus, panelists were all familiar with all the items in the lists, and they had reviewed and discussed student responses to constructed response items.

The process was based on general agreement among panelists regarding whether each item should be used as an exemplar of performance at a specific achievement level. Facilitators allowed the group, as a whole, to determine whether items would be recommended or vetoed. In some instances, an item might be selected although a few people voted to veto the item. In other instances, an item may be vetoed because of intense opposition to the item—even by only a minority of panel members. 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, facilitators, and NAGB staff. Panelists were selected from each grade group and panelist type. Nine judges 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 was thanked for their work and the meeting was officially over. The following is a summary of the remarks made during the debriefing session.

About Writing Borderline Descriptions

Panelists at the debriefing session were asked about the usefulness of developing borderline descriptions. Several general public panelists indicated that writing borderline descriptions was a helpful activity for them. One grade 12 teacher, however, said writing borderline descriptions was of no help. One Native American panelist complained that the ALDs did not reflect the role of Native Americans in developing the nation.

About Understanding the Purpose of the Achievement Levels-Setting Process

When asked how well they understood the purpose of the achievement levels-setting process, one panelist said that s/he did not understand the purpose until after the first round of ratings.

About Improving the Organization of the Process

When asked where more time was needed to “digest” information, all of the participants felt that they were pressed for time and had too much to do. A few people were annoyed because they were exhausted and frustrated by the intensity of the process. The debriefing panelists all agreed that an added sixth day would reduce the time pressures and offer panelists the opportunity for networking and leisure activities.

About Working with the Reckase Charts

When asked about working with the Reckase Charts, panelists commented that they felt prompted to change their item ratings based on the information contained in the charts. Some panelists thought that the charts emphasized the importance of item ratings over panelists’ judgments. The group suggested that the significance of the Reckase Charts be made absolutely clear to judges during the ALS process. That is, the Reckase Charts should be conceptualized as only one of several sources of information available to panelists when forming their judgments. The information on the charts should not be regarded as more significant than well-formed judgments.

Evaluation of the Overall Pilot Study Process

One of the primary purposes of the civics pilot study was to test the procedures planned for implementation in the Civics ALS. The pilot study lead to the identification of necessary adjustments to the process so that it would be implemented smoothly in the actual ALS meeting. Although the civics pilot study was executed without any major problems, it became apparent while conducting the study that several procedures needed to be refined. The following section summarizes the procedural issues that needed improvement and the adjustments that were planned for the 1998 Civics NAEP ALS.

Agenda

Issue: Panelists wanted to lengthen the five-day meeting to allow more time to assimilate information, get to know other participants, and relax after a full day of intense work.

Adjustment: ACT will start the meeting earlier on Day 1 and add a social “mixer” to the activities planned at the end of the day. ACT would not extend the ALS meeting to six days because of concerns about the loss of participation in the process by outstanding panelists.

Achievement Levels Descriptions

Issue: ACT anticipated that some panelists would be troubled by the fact that they would be working with the finalized versions of the ALDs without the opportunity to modify the descriptions to reflect their own judgment of student performance.

Adjustment: ACT’s concern was unfounded. There was no evidence that panelists were troubled by not having the opportunity to modify the ALDs. None of the judges expressed a desire to revise the descriptions.

Issue: Participants commented that working with the ALDs would be easier if they were in bullet format.

Adjustment: Because the ALDs were written in a narrative format, the meaning of the ALDs would change if they were reformatted to bulleted statements. NAGB’s policy is to report ALDs in narrative format. No recommendation was made to NAGB to change the format.

Writing Descriptions of Borderline Performance

Issue: As a training exercise, panelists spent considerable time and effort writing descriptions of borderline student performance. There was no evidence to suggest that writing descriptors for borderline student performance enhanced panelists’ understanding of borderline performance more so than discussions of borderline performance.

Adjustment: ACT decided to retain this training exercise even though panelists did not respond more positively to the evaluation questions related to their level of understanding borderline performance.

A primary reason for including the exercise of writing borderline descriptions was to sharpen the focus of panelists on the ALDs. Since no modifications were allowed to the ALDs, this was the major motivation for panelists to study the ALDs carefully.

PREVIEW RATING SESSION

Issues: A Preview Rating Session had been added to the agenda to give panelists an understanding of how important the information about and training in borderline performance would be later in the process. Judges struggled greatly with rating NAEP items during the Preview Rating Session. ACT staff concluded that panelists had difficulty making the transition from applying the ALDs to evaluate assessment items, to applying the ALDs to estimate borderline student performance. Further, panelists were disturbed by the fact that they did not receive any type of feedback after the Preview Rating Session. Since training for the rating process had not been completed, ACT did not want to provide feedback to inform panelists about their ratings.

Adjustments: ACT will change the agenda so that panelists will be introduced to the rating process early in the meeting through a demonstration The introductory activity will illustrate the rating process by giving an example of each step in the procedure. Panelists will not participate in the rating of items at this early stage of the process so there will no longer be a question of providing feedback. This activity will be referred to as a training exercise, rather than the "Preview Rating Session."

PAPER SELECTION EXERCISE

Issue: The Paper Selection Exercise required judges to examine three student papers scored at each score point for all of the constructed response items in each panelist's item rating pool. The grade 4 group reviewed 138-141 student papers, the grade 8 group reviewed 165-168 papers, and the grade 12 group reviewed 171-174 papers. Although most of the student responses were very short answers, panelists complained that this exercise was tiring and caused them to feel fatigued before undertaking the first round of item-by-item ratings. They wanted more time to discuss their selections with other panelists in their group, and they wanted feedback from the exercise so they could judge how "accurately" they had selected papers.

Adjustments: ACT will reduce the number of papers that ALS panelists will examine during this exercise. Panelists will review the student papers only for the common blocks of items that all panelists rated in their grade group. This will provide time to discuss the selections and enhance the training process. As a result, 72 student papers will be the maximum number reviewed by grade 4 panelists, and about 65 papers will be the maximum number reviewed by grade 8 and 12 panelists. Scored papers for the remaining constructed response items in the rating pool will be available for judges to review prior to round 1 ratings. Further, ACT will give panelists feedback from the Paper Selection Exercise. Frequencies will be reported so that panelists will know how many panelists selected specific papers to represent performance at the borderline of Basic, Proficient, and Advanced achievement levels.

RECKASE CHARTS

Issues: Many panelists complained that they could not see the numbers on the Reckase Charts easily. They found that transferring their ratings to the charts by hand to be tedious work. Some suggested that their ratings should be marked electronically for them, which was an idea introduced by a facilitator when working with his grade group. A few panelists suggested that the Reckase Charts should be distributed earlier in the process rather than prior to the second round of ratings. Some panelists thought that the charts emphasized the importance of item ratings over panelists' judgments. They suggested that the significance of the charts be moderated by describing them as only one of many sources of information panelists could consider when forming their judgments.

Adjustments: Although the charts themselves could not be enlarged further, ACT will develop a computer presentation that will augment and display the Reckase Charts clearly during instructions. The presentation will include a “zoom-in” feature that will enable specific portions of the charts to be selected and enlarged for easier viewing by panelists. In addition, a computer program will be developed to electronically mark panelists’ item ratings on the charts. However, panelists will continue to connect their marked ratings by hand to assist them in identifying individual rating patterns. ACT decided not to distribute the charts earlier than prior to the second round of ratings because of the concern that panelists would receive information that would cause them to defer to data rather than to use their judgments regarding the ALDs. ACT also decided that the Reckase Charts would be conceptualized as a form of feedback for the ALS meeting, rather than as a distinct step in the rating process as it was implemented in the pilot study.

CONCLUSIONS DRAWN FROM CIVICS PILOT STUDY RESEARCH

The Civics Pilot Study was the final opportunity for ACT to evaluate procedures to be used for the operational Civics 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 Civics 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. Panelists used the Reckase Charts to evaluate their ratings for each item along several, important dimensions. For example, Reckase Charts showed panelists that the likelihood of students correctly answering an individual item is expected to increase as the overall performance of students increases. Reckase Charts also showed panelists that this was not always the case because of the lack of item discrimination within some ranges of performance.

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. Pilot study panelists advised, however, that the significance of the charts be moderated by describing them as only one of many sources of information panelists might consider when forming their judgments. The conceptualization of the charts was adjusted from being considered a step in the ALS process, to being described as an additional source of feedback information.

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 item 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 item 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 items 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 items. 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 items 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 Civics 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 DIFFERENCES BETWEEN MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEMS

The difference between the cutscores that would result from ratings of polytomous and dichotomous assessment items has been another persistent challenge to ACT's effort to refine the standard setting process for NAEP. As has been the case in past ALS meetings, panelists for the Civics Pilot Study generally set cutscores that were statistically significantly higher for polytomous items than dichotomous items[11]. Differences in ratings of multiple choice and constructed response items were one of many considerations brought to the attention of panelists when reviewing Reckase Charts. After studying the feedback, including the Reckase Charts, panelists adjusted their ratings for the subsequent round (Round 2 or Round 3). In general, the differences between multiple choice and constructed response ratings were reduced for all grades and all levels for the subsequent rounds. The cutscores set for polytomous items were usually higher than those set for dichotomous items. As has been the case for previous investigations in other NAEP ALS studies, differences in *cutscores* computed for items of the two types are greater than differences in *ratings* would suggest (ACT, 1997a). Although more research is needed to determine how judges perceive polytomous items relative to dichotomous items, the Reckase Charts appear to have been effective in helping to make panelists aware of differences when forming their judgments of student performance.

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 with reasonable accuracy (NAE, 1993: Shepard, 1995: Impara and Plake, 1998). ACT has collected considerable data during the Civics 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.

THE ISSUE OF PROVIDING CONSEQUENCES DATA BEFORE FINAL CUTSCORES ARE DETERMINED

NAGB has maintained a criterion-referenced achievement levels-setting procedure. They had not approved ACT's recommendations to provide consequences data to panelists before the final ratings were collected. ACT proposed to produce two sets of cutscores to share with NAGB for the 1998 Civics ALS Process. Cutscores would be computed, as usual, based on Round 3 ratings. These were not influenced by consequences data. Following the completion of the item-by-item rating process, consequences data were shared with panelists and they were given the opportunity to recommend different cutscores for each achievement level. Those recommendations were averaged to compute new, final cutscores. The final cutscores, based on panelists' recommendations, were informed by consequences data.

Panelists were given individual-level consequences data after the third round of ratings. They were provided with information about the consequences associated with each of their own cutscores and those of other panelists in their grade group. They had many data points to inform

[11] Grade 4 panelists did *not* set cutscores that were statistically significantly higher for polytomous items than dichotomous items for Rounds 1 and 2 for the Civics Pilot Study.

their recommendations of new cutscores. This was a new way of presenting consequences data to panelists, and it appeared to work well. Panelists seemed to be pleased to have this information, although they were somewhat reluctant to use it for changing their cutscores at this point. They asked questions about the data and questioned the wisdom of making "arbitrary" changes after the careful, detailed consideration given to individual item ratings. They were not inclined to make major adjustments in their cutscores as a result of the consequences data.

PLANNING FOR THE CIVICS ALS

The details of implementing the standard setting process were closely scrutinized and evaluated during the Civics Pilot Study. Panelists offered a wealth of information about adjusting the ALS process, from suggesting the color of markers to debating the impact of consequences data. As a result of extensive research conducted not only for the 1998 civics 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 (1997a). *Setting achievement levels on the 1996 NAEP in science: Final report, Volume IV: Validity evidence special studies*. Iowa City, IA: Author.

ACT (1997b). *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.

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.