Developing Achievement Levels for the 1998 NAEP in Civics Interim Report: Field Trials

Susan Cooper Loomis, Patricia L. Hanick, Luz Bay, and Jill D. Crouse
ACT, Inc.

December 2000

Developing Achievement Levels for the 1998 NAEP in Civics Interim Report: Field Trials

Susan Cooper Loomis, Patricia L. Hanick, Luz Bay, and Jill D. Crouse
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

OVERVIEW

Two field trials were conducted for the achievement levels-setting (ALS) process for the1998 Civics National Assessment of Educational Progress (NAEP). ACT proposed to conduct field trials as a means of collecting research information regarding new methods and procedures designed for the 1998 ALS process. ACT wanted to conduct the research involving panelists before the pilot study so that the pilot studies could be used to test the procedures selected for the ALS. Experiences with the 1994 pilot studies for geography and U.S. history led ACT to recommend this additional set of studies for research purposes so that pilot studies could be used for practice and fine-tuning.

ACT had proposed to conduct only two small-scale field trials. Once the design of the studies started to take shape, however, plans changed so that two field trials were planned for each subject—civics and writing—included in the 1998 NAEP ALS procedures. Further, the scale of the field trials expanded to 40 panelists in the second field trials.

This report documents two field trials conducted for civics. Taken together, the field trials research provided important information about various elements that constitute the ACT/NAGB standard-setting process. Findings from the field trials research greatly informed the procedures that were developed for the pilot study and implemented to set achievement levels in 1998.

CIVICS FIELD TRIAL 1

- ✓ *The purpose of the first civics field trial was to try out a new item-by-item rating method and compare the implementation and outcomes of the new method with the current ACT/NAGB method used since 1994.*
  The new method examined in the first field trial was called the Item Score String Estimation (ISSE) rating method, and that was compared to the results from the 1994 geography ALS using the Mean Estimation method.[1]

- ✓ *The method was selected for its simplicity and likely ease of implementation. The cognitive demand for judges using the ISSE method was expected to be less.*

- ✓ *Results of the first field trial in civics indicated that panelists were able to use the ISSE method without difficulty.*

- ✓ *The ISSE procedures were implemented with ease.*

- ✓ *The panelists expressed satisfaction with and confidence in the ISSE method and the outcomes of the process.*

[1] The 1994 geography ALS data were used for the field trials in civics since the Civics NAEP data would not be available before fall, 1998. The geography and civics assessments were similar in terms of the mix of multiple-choice and constructed response items. Further the two subjects are similar in terms of their role in the K-12 curriculum.

- ✓ *The cutpoints and their standard deviations did not differ greatly from those produced by geography ALS panelists using the Mean Estimation method, but the ISSE method resulted in higher Proficient and Advanced cutscores and lower percentages of students performing at those achievement levels than resulted from the Mean Estimation method.*

- ✓ *The ISSE method was found to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true score" or "true" judgment of the panelist (Reckase & Bay, 1999).*
  - *Because of this flaw, further research using the ISSE method was discontinued, and it was eliminated as a possibility for implementation in the civics ALS.*

CIVICS FIELD TRIAL 2

The original purpose of the second field trial was maintained: to study the effect of item maps and the timing of consequences data on the outcomes of the process. Since no alternative method had been selected, the Mean Estimation procedure was interfaced with these research features of the second field trial design. Mark Reckase developed a new method, and ACT studied this as an alternative to the current ALS method (Mean Estimation).

- ✓ *The fundamental purpose of the second field trial was to identify the method that would be used for the 1998 ALS process.* Because the alternative method tested in the first field trial had been found to be inherently biased, research continued to explore additional alternatives.

- ✓ *The second field trial was used to examine whether and how well an item mapping procedure could be interfaced with the item-by-item rating method and to study the relative effects of the timing of consequences data.*

- ✓ *Results of the second field trial indicated that panelists had no difficulty using either the new Reckase Method or the combination of the Mean Estimation method and item maps.*

- ✓ *Informing panelists of the consequences of their ratings earlier in the standard setting process, rather than later, did not significantly affect the outcomes.*

FINDINGS REGARDING ISSUES IN NAEP STANDARD SETTING

It is important to emphasize that the field trials addressed several unique and difficult technical challenges inherent to setting achievement levels for NAEP. These challenges have been an on-going concern for ACT in its effort to refine and improve the NAEP standard setting process.

THE ISSUE OF COGNITIVE COMPLEXITY

ACT has collected considerable data during the civics field trials and previous ALS research where panelists have reported their capacity to perform the tasks associated with estimating student performance with no significant cognitive difficulty.

- ✓ *Judges perceive that they are performing the estimation and judgmental tasks required by the method with relative ease.*

- ✓ *They report that they are confident in their judgments and satisfied with the results.*

- ✓ *There is no evidence to indicate that panelists are unable to make judgements regarding borderline performances of students relative to the achievement levels descriptions.*

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

One of the considerations for evaluating the various experimental methods explored in the field trials was based in part on indicators of "reasonable" 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. Judges were expected to adjust their ratings from round to round. 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, this was an indicator 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, this was also an indicator that they probably did not understand the rating method or the feedback.

- ✓ *The civics field trial panelists exhibited "reasonable" intrajudge consistency across rounds based on these indicators.*

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

ACT has been searching for an effective format to provide panelists with information about the relationship of their ratings to the performance of students on items. Although this information had been given to panelists in previous ALS meetings, there was no evidence to suggest that panelists either understood the information, or found it useful when forming their judgments about student performance.

One of the perceived advantages associated with the proposed ISSE method was that ACT had devised a format for providing panelists with intrajudge consistency data that appeared simple to read and easy to understand. When the ISSE method was eliminated from further consideration because of bias, the new design for providing intrajudge consistency feedback was also eliminated.

- ✓ *The Reckase method greatly advanced the effort to give panelists precise item-level information they could readily understand about ratings, relative to expected student performance.*

- ✓ *Panelists who used the Reckase method adjusted their ratings to be more similar to the IRT-based performance estimates of students at the cutscores. This finding was consistent for all three achievement levels.*

- ✓ *Evidence suggests that panelists considered the achievement levels descriptions and other forms of feedback in addition to the data in the Reckase Charts when forming their judgment of student performance.*

- ✓ *The Reckase method required panelists to select a scale score (a single row on the Reckase Charts) to represent their final ratings for each of the three achievement levels. When asked about their confidence in the cutscores that they selected, 20 of the 22 panelists using the Reckase method responded positively.*

THE ISSUE OF DIFFERENCES BETWEEN CUTSCORES FOR POLYTOMOUS AND DICHOTOMOUS 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. Differences in ratings of multiple choice and constructed response items were one of many considerations brought to the attention of panelists.

- ✓ *As has been the case in previous ALS studies, panelists in the civics field trials set cutscores that were statistically significantly higher for polytomous items than dichotomous items.*

- ✓ *Panelists using the Reckase method were instructed to examine ratings for dichotomous and polytomous items to determine whether their ratings reflected any patterns of differences.*

The differences between multiple choice and constructed response ratings were reduced in 24 of the 66 opportunities for adjustments (22 judges and 3 achievement levels). Differences increased between multiple choice and constructed response ratings in 19 of the 66 opportunities for adjustments. Differences remained unchanged in 23 opportunities for adjustments. This finding suggests that panelists did not lower their ratings for polytomous items, even when they saw that the ratings were associated with relatively higher performance scores. Apparently they judged the performance of students to be inconsistent with the ALDs. Clearly this is not an oversight by the panelists indicating a flawed process.

- ✓ *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 making panelists aware of these differences when forming their judgments of student performance.*

THE ISSUE OF TIMING CONSEQUENCES DATA

The impact of consequences data on outcomes has been a topic of considerable interest in setting standards. The impact of receiving consequences data on the final cutpoints seemed moderate. About an equal number of comments were made by panelists about the positive impact of consequences data as were made about the limited impact. Judges, in general, found the consequences data informative and useful, but cutpoints were not greatly influenced by the data.

- ✓ *No compelling differences were found in cutscores produced by judges who received consequences data early in the process compared with late in the process.*

- ✓ *Most panelists indicated that consequences data were only one of many factors they considered when forming their judgments about student performance.*

PLANNING FOR THE CIVICS PILOT STUDY

- ✓ *TACSS recommended that ACT eliminate item maps from further consideration because of the unresolved difficulties associated with selecting a response probability to serve as the statistical criterion for mapping items.*

- ✓ *The Reckase Method, per se, was eliminated, but the Reckase Charts would be used in the pilot studies, with the expectation that they would also be used for the ALS.*

- ✓ *The Reckase Charts, introduced in the second field trial for civics, were judged to be a promising addition to the ALS process designed for NAEP. The charts appeared to have improved and strengthened the ALS process.*

- ✓ *The Reckase Charts would be presented to panelists as a step in preparation for the rating process. This modified role of the Reckase Charts would be implemented in the ALS for civics, unless findings in the pilot study suggested otherwise.*

- ✓ *There was little evidence that the consequences data impacted judgments of panelists enough to effect the cutscores they set.*

ACT and TACSS have consistently recommended that NAGB allow consequences data to be included as feedback in the ALS process. The findings from field trial research indicated that outcomes of the process would not be significantly different if consequences data were provided as feedback.

- ✓ *Recognizing that NAGB wishes to have the NAEP achievement levels set via a criteria-referenced methodology, TACSS recommended that consequences data not be presented to panelists until after three rounds of ratings.*

- ✓ *TACSS recommended implementation of a design that would produce two sets of cutscores: one set based solely on ratings, and one modified as a result of consequences data. The latter were to be the cutscores used for selecting exemplar items and, presumably, to be recommended to NAGB.*

Developing Achievement Levels on the 1998 National Assessment of Educational Progress in Civics Interim Report: Field Trials

Susan C. Loomis, Patricia L. Hanick, Luz Bay, and Jill D. Crouse

INTRODUCTION

Setting achievement levels for the National Assessment of Educational Progress (NAEP) is a judgmental process. Panelists' estimates of how students would perform on each item of the assessment determine the cutpoints that assign student performance to a given level of achievement. Therefore, the method for collecting and summarizing these judgments is of great importance in setting achievement levels for NAEP.

For the 1998 NAEP achievement levels-setting (ALS) process, ACT proposed several studies designed to improve procedures for collecting and summarizing judgments. Four field trials were designed to explore new methods for setting achievement levels: two for civics and two for writing[2]. It was anticipated that the results of the first set of field trials would lead to identification of the method that would be used for the 1998 ALS process. It was expected that the second set of field trials would then focus on additional research issues important to the development of the ALS process.

Because the results of the first field trial did not lead to identification of the method that would be used for the 1998 ALS process, additional exploratory research was conducted in the second field trial in an effort to identify new methods for setting achievement levels. In particular, the field trials explored topics that related to the cognitive complexity of the item-rating task, methods of computing cutscores, timing of consequences data,[3] differences between polytomous and dichotomous item ratings and cutscores, and effective formats for presenting intrajudge consistency feedback[4] to panelists. Taken together, research from the field trials has provided important information about various elements that typically constitute the standard-setting process. Findings from the field trials research greatly informed the procedures developed for the pilot study. The following report is a full account of the methodologies used in the first and second civics field trials and the outcomes of those methods.

[2] A report by Loomis, Bay, Yang, and Hanick (1999) provides a comprehensive overview of procedures and findings from the field trials in both civics and writing.

[3] "Consequences" refers to the percentage of students scoring at or above each achievement level based on the cutpoints panelists set.

[4] "Intrajudge consistency" refers to the internal consistency of a judge's ratings. Indicators of intrajudge consistency include measures of the extent to which individual item ratings are consistent with a judge's overall item ratings at a given level of achievement, as well as measures of rating changes from round to round.

BACKGROUND INFORMATION FOR FIELD TRIALS RESEARCH

Since 1992, ACT has experimented with various item-by-item rating methods in the NAEP ALS process to determine the cutpoints for each achievement level (Loomis & Bourque, in press). The specific method used since 1994 has been called "Mean Estimation" (ME) because judges estimate the average or mean score of polytomous items for a student performing at the borderline of each achievement level. Critics expressed concern that the item-by-item rating method could not produce valid cutpoints because panelists were incapable of performing the cognitively complex task of estimating probabilities with reasonable accuracy (NAE, 1993; Shepard, 1995; Impara and Plake, 1998). The Technical Advisory Committee on Standard Setting (TACSS) presented evidence to the contrary and refuted the claims of the critics (Hambleton, et al., 2000).

In response to this concern, ACT proposed to explore a new method for setting NAEP achievement levels that was thought to be less cognitively complex than the Mean Estimation method. This method was based on the work of Impara and Plake (1997). Their studies compared the modified-Angoff method with the "Yes/No" method.[5] Instead of estimating probabilities, the "Yes/No" method required judges to select "yes" if they thought a student performing at the borderline of each achievement level would respond to the assessment item correctly. Panelists selected "no" if they judged that a student performing at the borderline would not respond correctly. It was argued that panelists could perform the task easily of selecting "yes/no," whereas estimating percentages was a task requiring a level of accuracy that panelists could not attain.

The results of the studies by Impara and Plake (1997) indicated that the two methods produced similar cutpoints, but the "Yes/No" method had several advantages over the modified-Angoff method. The "Yes/No" method produced greater homogeneity of panelists' ratings (interjudge consistency) than the modified-Angoff method. When panelists were asked to compare the two methods, they reported that the "Yes/No" method was easier to use and easier to understand, and they were more comfortable using it than the modified-Angoff method. Finally, judges' level of confidence in the standards that they set was as high as the confidence level of panelists who used the modified-Angoff method.

ACT proposed to further explore the "Yes/No" method. The research by Impara and Plake involved only dichotomous items. Because NAEP includes both dichotomous and polytomous items, the "Yes/No" method by itself was not suitable for setting achievement levels on NAEP. If the procedure were to be implemented for NAEP, it needed to be expanded to address rating polytomous items.

IMPLEMENTING THE ITEM SCORE STRING ESTIMATION (ISSE) METHOD

ACT needed to identify a rating method that could be easily used to rate both dichotomous and polytomous items and that would produce reliable cutscores. The 1998 Civics NAEP included a large number of both dichotomous and polytomous items, while the 1998 Writing NAEP included only two writing tasks for each student. ACT created a new procedure that combined two rating

[5] This method is actually the procedure first described by Angoff (1971).

methods that had been reported in the literature. It merged the "Yes/No" method for dichotomous items described by Impara and Plake (1997) with the extended-Angoff method for polytomous items described by Hambleton and Plake (1995). This combination of rating methods formed the basis of the new Item Score String Estimation (ISSE) method.

The ISSE method treated ratings as if they were an "item score string" for a series of student responses that included both dichotomous and polytomous items. The ISSE method required judges to indicate (yes/no) whether students performing at the borderline of each achievement level were likely to respond correctly to each dichotomous item. In contrast, the item-by-item method for rating dichotomous item required panelists to estimate the probability (e.g., 65%) that students performing at the borderline of each achievement level would respond correctly to an item. For polytomous items, the ISSE method required panelists to indicate the most likely score (e.g., 1, 2, or 3 on a scale of 1-3) for students performing at the borderline of each achievement level. The rating would be recorded as a whole number. In contrast, the Mean Estimation method for polytomous items required panelists to estimate the average score (e.g., 2.4 on a scale of 1-3) for students performing at the borderline.

COMPUTING CUTSCORES USING THE ISSE METHOD

The first issue to be addressed in determining whether the ISSE method could be used to set NAEP achievement levels was how to combine ratings to produce one cutpoint that defined the achievement standard. To explore this issue, simulation studies were conducted using data collected from previous ALS meetings. The research examined different computation methods that aggregated panelists' judgments collected through implementing the ISSE method (Chen, 1998). The EM-algorithm by Sanathanan and Blumenthal (1978) was used for these computations. Please see Appendix A for documentation of this procedure.

Results of the simulation studies indicated that the new ISSE method under consideration was feasible to use for setting NAEP achievement levels. That is, the method produced reasonable results, relative to previous ALS procedures. The algorithm could be used successfully to aggregate ISSE ratings to produce numerical cutpoints. Further, the ISSE ratings for a relatively small number of judges and items could be aggregated and still produce statistically acceptable results. This was a particularly important finding because the ALS process typically includes 30 panelists per grade. The writing NAEP was comprised of only 20 writing prompts per grade. Our concern was that far more panelists would be needed to produce reliable results with only 20 prompts. That concern was eliminated through the simulation studies. Further, the ISSE method could be used for both the civics and writing NAEP because it accommodated ratings of dichotomous *and* polytomous items. Given these encouraging outcomes, ACT decided to continue investigating the merits of the ISSE method through field trial research.

KEY ASPECTS OF FIELD TRIAL 1 CIVICS

The purpose of the first civics field trial was to study the ISSE method relative to the Mean Estimation (ME) method. The criteria for evaluating the ISSE method were measures of the

reasonableness of the cutpoints, the level of interjudge consistency,[6] the level of intrajudge consistency, and the reactions of panelists to the procedures. ACT has implemented many ALS processes for NAEP, and that experience would serve as a basis for comparing the ISSE method with the Mean Estimation method. ACT would not choose a less reliable method or one that produced cutscores that were very different from those produced from the Mean Estimation method. In addition to these measures, panelists' reactions to the rating method and its outcomes were of central importance. Panelists' comments and responses regarding the ease of using the rating method, their understanding of the rating method, and their satisfaction with and confidence in the outcomes of the process, for example, were all included in the criteria used to evaluate the ISSE method. The amount of time required to complete the task and the observations regarding panelists' reactions were also taken into account. If the ISSE met the statistical requirements and appeared to be as easy to use as the Mean Estimation method, it would be used for the 1998 Civics ALS process.

ACT had proposed to introduce consequences data *before* the final round of ratings. ACT had collected panelists' reactions to consequences data provided at the end of the ALS process for geography, U.S. history, and science, but consequences data had never been provided when panelists were actually allowed to make changes to cutpoints after reviewing the data. This field trial was one of several opportunities planned for collecting data to study the effect of providing consequences data to panelists before the final cutscores were set.

The first field trial for civics was conducted February 7-8, 1997 (Saturday-Sunday) at the Tyler Conference Room on the ACT campus in Iowa City, Iowa. The study was conducted only for grade eight. The NAEP ALS Project Director facilitated the two-day study.

DATA, ACHIEVEMENT LEVELS DESCRIPTIONS, AND ITEM RATING POOLS

The 1994 NAEP Geography data were used for the civics field trials. Administration of the 1998 Civics NAEP had not been completed; consequently, 1998 civics data were not available to use for the field trial. The choice of geography data seemed reasonable because the assessment characteristics of the Geography NAEP were very similar to the characteristics of the Civics NAEP, and both subjects are social studies. Further, the two disciplines are regarded about equally with respect to status in the curriculum of most schools and districts.

Four blocks of items from the 1994 NAEP Geography survey were used for the field trial. The same blocks had been selected for the Similarities Classification Study (SCS) that was conducted as validation research for the 1994 Geography ALS (ACT, 1995). The blocks for the SCS had been selected to maximize the representativeness of the assessment features of the entire grade 8 item pool. Item format, content coverage, reliability and item characteristics of difficulty, discrimination, and so forth were the features represented (Carlson, 1995). Given the limitation that the civics field trial was planned for two days in contrast to the five-day ALS meeting, rating four blocks of items seemed the minimum number for comparing the two methods in the allotted amount of time.

[6] "Interjudge consistency" refers to the extent to which an individual judge's ratings are consistent with the ratings of the other judges in the grade-level rating group.

RECRUITMENT AND SELECTION OF PANELISTS

The study design called for 20 panelists, 10 for each of the two rating methods to be compared. Requests were sent to 186 key persons in local communities near ACT national headquarters asking for nominations of "outstanding" persons who were familiar with student work at the 8th grade level. The nominators included middle school principals, government officials, community leaders, school superintendents, and geography faculty from The University of Iowa. ACT was unable to recruit the planned number of panelists and only 8 participated in the study. The composition of the panel was to meet the same criteria required for ALS panels as nearly as possible (see Table 1). The panel was to consist of 55% teachers (TR), 15% nonteacher educators (NT), and 30% members of the general public (GP). Panelists were paid an honorarium of $100 for the two days. Breakfast and lunch were provided on both days.

Table 1
Composition of Civics Field Trial 1 Panel

Panelist Type			Gender		Race		Total
TR	NT	GP	Male	Female	White	Minority	
2	3	3	3	5	7	1	8

Although the study was originally designed to compare the Mean Estimation and ISSE methods directly, only the ISSE method was implemented due to the small number of panelists recruited for the study. It was reasoned that past Geography ALS results produced by using the Mean Estimation method could be compared with the field trial results using the ISSE method. The comparison would be limited, given the differences between the two studies, but useful nonetheless.

Although the number of panelists was small, their credentials were outstanding and the quality of the panelists was judged to be quite high. Four panel members had taught or were currently teaching geography/global studies courses in local junior high schools, and four had a past or current affiliation with the geography department at the University of Iowa.

THE ALS PROCESS DESIGNED FOR CIVICS FIELD TRIAL 1

STEP 1: BRIEFING MATERIALS

Before the meeting, all panelists were mailed a set of materials that contained important background topics and introductory information on setting achievement levels. Panelists were asked to avoid reviewing published reports that described results of the 1994 Geography NAEP. The briefing materials included the following information:

- 1994 NAEP *Geography Framework*;
- 1998 NAEP Geography Achievement Levels Descriptions;
- *Multiple Challenges*, a booklet on the 1998 NAEP;
- NAGB brochure;
- *The NAEP Guide*;
- Cover letter with instructions for preparing for the field trial;

- Item-Use and Nondisclosure Agreement;
- Check Request Form;
- Request for Taxpayer I.D. Number and Certification;
- Directions and a map to the meeting.

STEP 2: GENERAL ORIENTATION AND TRAINING EXERCISES (DAY 1)

Although the study only lasted two days, all elements of the five-day achievement levels-setting process were covered, at least to some extent. A copy of the meeting agenda has been included as Appendix B. Panelists were provided an abbreviated orientation to the achievement levels-setting process and the procedures planned for the field trial. The orientation session included a general introduction to the NAEP program and an overview of the method used to develop NAEP achievement levels presented by a member of the NAGB staff. Panelists participated in several training exercises that were the same as those included in the training of ALS panelists.

TAKING A FORM OF THE NAEP

Panelists were administered a form of the Geography NAEP, and they reviewed their own responses relative to the scoring guides.

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS (ALDS)

Panelists were given time to work with the ALDs and to discuss them. They developed an understanding of the achievement levels descriptions and reached a common agreement on their meaning. Panelists studied the items and the scoring guides for each item. They were not given the opportunity to revise the ALDs. Panelists applied the ALDs to student test booklets requiring responses for multiple choice and constructed response items. The final version of the grade 8 geography achievement levels descriptions (ALDs) was used for the field trials.

UNDERSTANDING BORDERLINE PERFORMANCE

Panelists also received brief training in the concept of borderline performance. They did not write borderline descriptions as part of their training, however.[7] Panelists discussed borderline performance and reached a common understanding of what constituted borderline performance at each achievement level.

STEP 3: THE ITEM RATING PROCESS

ROUND 1 RATINGS (END OF DAY 1)

After orientation and general training exercises, the panelists were trained in the rating method by the facilitator.

[7] Panelists had been asked to write descriptions of borderline performance as part of the 1996 Science ALS process. Process evaluations by panelists revealed no significantly higher level of understanding of borderline performance for science panelists than for panelists in earlier studies that did not include written borderline descriptions.

The ISSE Rating Task

The ISSE rating task for multiple-choice items required judges to indicate whether or not students performing at the borderline of each achievement level would answer the item correctly by marking "yes" or "no" on the rating form. The rating forms for the ISSE method displayed a "Y/N" (yes/no) beside the spaces provided for rating multiple-choice items. The ISSE rating task for constructed response items required judges to indicate the most likely score (whole numbers) of those described in the scoring guide. Either "1-3" or "1-4" which identified the score range, appeared beside the spaces provided for rating constructed response items. Please see Appendix C for detailed instructions to panelists on the ISSE rating method.

FEEDBACK AFTER ROUND 1 (BEGINNING OF DAY 2)

At the start of Day 2 in a general session, panelists were given feedback data resulting from their first round of ratings. They were provided with cutpoints and their standard deviations, rater location charts, student performance data, and whole booklet feedback prior to the second round of ratings. Training in the use of each form of feedback data was provided. Copies of the different types of feedback based on the first round of ratings are included in Appendix D.

Cutpoints

The cutpoints were computed from the combined ratings of all raters and all items for each achievement level. The cutpoints were presented on the ACT NAEP-like scale, which is a linear transformation of the NAEP score scale. This transformation guarded the confidentiality of the results and decreased the potential for achievement level data from the Geography NAEP or other NAEP subjects to influence panelists in the field trial. Item parameters produced by a 3-parameter IRT model were used in computing the cutscores and other feedback data.

Standard Deviations of the Cutpoints

The standard deviation is the indicator of the level of variability around each cutscore. The cutscores were computed as the mean score over all items and raters, and the standard deviation is relative to that mean score value. The standard deviation is computed as the variability of individual raters' cutscores relative to the overall group cutscore. (See the description of Rater Location Feedback Charts.)

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 were 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 were used so the cutscores for each panelist remained confidential. (In fact, most panelists openly and freely discussed their rater location data.) The graphs indicate the cutscores that resulted 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 was produced to display rater locations for each of the three achievement levels.

Student Performance Data

Panelists received information about overall student performance on each item in the rating pool. The proportion of students who gave the correct answer was listed as the actual "p-value" for each dichotomous item. The mean (average) score was reported for each polytomous item, along with the percentage of student responses scored at each rubric score point. The data also reported various categories of no response for each item. Student performance data served as a reality check because it showed how students actually performed on each item. The data indicated how easy or difficult the items were for all students who took the 1994 Geography NAEP. They did not indicate how easy or difficult the items were for students at different achievement levels.

Whole Booklet Feedback

Whole booklet feedback was 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. The whole booklet feedback reported the percent of total possible points that a student needed to earn in an assessment booklet in order to meet the minimal requirements for performance at the cutscore of each achievement level. For example, the whole booklet feedback report might state: "Based on the cutscore for your grade, students performing at the borderline Basic level are expected to get 49% of the total possible score points for this booklet." A similar statement was given for each achievement level. This feedback was based on the cutpoints the panelists had set during the first round of ratings, and was updated after subsequent rounds of ratings.

Whole Booklet Exercise

As part of Round 1 feedback, the panelists participate in a whole booklet exercise, which was an extension of whole booklet feedback. Panelists participated in this exercise only after Round 1. They were shown actual student booklets with scores near the cutpoints that had been set by Round 1 ratings. The booklets were the same form used for the training exercise "Taking a Form of the NAEP." Panelists evaluated booklets scored within 2% above or below the total possible points associated with each cutpoint. They were asked to examine the responses of the student to all items in the booklet as a whole, and to determine if the responses represented their expectations of student performance at the lower borderline of Basic, for example. If they perceived a discrepancy between the expected performance and the observed performance in the booklets scored at the cutpoint, they discussed the achievement levels descriptions and borderline performances again with other panelists to try to understand the cause for this discrepancy. Performance higher than expected would signal that they had set their cutpoints too high. Performance lower than expected would signal that they had set their cutpoints too low.

ROUND 2 RATINGS (DAY 2)

After reviewing and discussing the results and feedback from round 1 ratings, panelists again rated the same assessment items using the same methodology. Panelists could change all, some or none of their ratings for any or all achievement levels.

FEEDBACK AFTER ROUND 2 (DAY 2)

Feedback information was distributed to panelists after being updated using the cutpoints set from the second round of ratings. Except for the Whole Booklet Exercise, panelists received the same forms of feedback after Round 2 that were presented after Round 1. Feedback information based on the second round of ratings has been included in Appendix D.

STEP 4: CONSEQUENCES DATA (DAY 2)

In addition to the feedback data resulting from their second round of ratings, panelists were given information about the consequences of the ratings that they provided. That is, panelists were told the percentage of students scoring at or above each achievement level based on the cutpoints they set for the second round of ratings. Panelists were not given consequences data after Round 1, so they were trained in the use of this information after Round 2.

Consequences data were explained to panelists, and they were asked to complete a questionnaire in which they were given the opportunity to recommend new cutpoints that would raise or lower the percentages of students performing at or above each level. A sample questionnaire has been included as Appendix E. The recommended cutpoints were averaged and new cutpoints and consequences data were presented. For panelists who chose to recommend unchanged cutpoints, the new average was computed using the values of the group cutpoints from Round 2.

STEP 5: EVALUATIONS THROUGHOUT THE PROCESS

Panelists responded to process evaluation questionnaires throughout the study. They completed one evaluation at the end of the first round of item-by-item ratings, which ended the first day's activities. They completed another evaluation at the end of the second round of ratings, and a final process evaluation questionnaire at the conclusion of the study on the second day. A more detailed description of the results of the evaluation questionnaires has been included later in this report.

OUTCOMES OF THE CIVICS FIELD TRIAL 1

The purpose of the civics field trial was to study the ISSE rating method and compare results to those from the Geography ALS using the Mean Estimation rating method. In general, panelists' responded positively to the procedures implemented for the civics field trial. The ISSE method did not pose any particular problems for panelists. The criteria for evaluating the ISSE method were judgments of the reasonableness of the cutpoints and their standard deviations, the level of interjudge and intrajudge consistency, the ease with which the method was implemented by panelists, and judges' satisfaction with and confidence in the process and the results of the

process. Logistic concerns and the reactions of panelists were particularly important criteria for this first panel study. If the ISSE method seemed as easy or easier than the Mean Estimation method for panelists to understand and use as a rating method, and if the results generally seemed reasonable—both to panelists and to ACT and the technical advisors—then the plan was to go forward with the ISSE method for the next field trial.

EVALUATION OF THE CUTPOINTS AND THEIR STANDARD DEVIATIONS, AND RESULTING CONSEQUENCES DATA

Table 2 contains the cutpoints and their standard deviations computed from the two rounds of ISSE ratings. For comparison purposes, the cutpoints resulting from the geography ALS using the Mean Estimation rating method are included in Table 2. The cutpoints from the geography ALS reported in Table 2 are computed from ratings for the four item blocks included in the field trial.

It should be noted that the cutpoints produced by the field trial panel were not entirely comparable to those produced by the geography ALS panels because of significant differences in the processes. For example, the geography ALS process lasted five days while the field trial lasted two days. Panelists for the geography ALS process were drawn from nationally representative samples, whereas the panelists for the field trial were recruited locally. Field trial panelists rated four blocks of items selected to be maximally representative of the full grade-level item pool rated by ALS panelists. Because many elements of the processes varied, the cutpoints for the two studies are not wholly comparable. However, the data do provide the opportunity to compare the results of the ISSE rating method with the results of the Mean Estimation rating method using items from the same item pool as the basis for the ratings.

Table 2
Comparison of the Outcomes from the ISSE Method for Civics Field Trial 1
With the Mean Estimation Method for Geography ALS Using the Same Item Rating Pool

Round	Achievement Level	Civics FT#1 (ISSE)		Geography ALS Ratings for FT#1 Item Pool (ME)	
		Cutpoint (SD)	%≥	Cutpoint (SD)*	%≥
1	Basic	149.62 (7.93)	67.2%	149.22	68.2
	Proficient	171.47 (6.73)	11.6	163.47	28.8
	Advanced	189.75 (8.66)	0.2	173.09	8.5
2	Basic	152.19 (9.79)	61.6%	151.81	61.6
	Proficient	171.47 (4.67)	11.6	164.62	25.5
	Advanced	187.33 (4.00)	0.3	174.47	7.0
3	Basic	Only two Rounds of Ratings		152.20	60.7
	Proficient			165.21	24.5
	Advanced			175.69	5.5

* The Geography ALS cutpoints were computed from ME ratings for the 65 items included in the civics FT rating pool. Standard deviations of the cutpoints cannot be computed because it would necessitate the computation of individual cutpoints of some ALS panelists that are based only on ratings for one block of items.

CUTPOINTS FROM THE FINAL ROUND

When comparing the cutpoints computed from the final round of ratings for the two groups[8], the Basic cutpoint was about the same for the two panels. The cutpoints for Proficient and Advanced, however, were considerably higher using the ISSE method than those calculated using the Mean Estimation method. As would be expected, the percentages of students performing at the Proficient and Advanced levels were much lower for ISSE ratings than the Mean Estimation ratings.

The geography ALS percentages in Table 2 were slightly different from those reported by NAGB in 1994 for *The Nation's Report Card in Geography*. The source of this difference was due to the fact that the Mean Estimation ratings for only those items that were included in the civics field trial were used to compute cutpoints for this comparison. The 1994 grade 8 geography cutscores that were reported in *The Nation's Report Card* resulted in 71% of the students scoring at or above the Basic level, 28% at or above the Proficient level, and 4% at or above the Advanced level.

Table 3 reports the percent correct score, which is the average over all panelists of the sum of scores for each panelist's ratings for each achievement level divided by the total possible points for correctly answering all the items. Data are presented separately for computations based on dichotomous item ratings and those based on polytomous item ratings. A comparison of the data for the two rounds provides an indication of the rating changes made by each panelist from round 1 to round 2, as well as the differences in ratings by item type.

The percent correct data for dichotomous item ratings computed from ISSE ratings are simply the proportion of 48 multiple-choice items the panelist estimated that a student performing at the borderline of each achievement level would answer correctly. There were 17 short constructed response items with 3 score levels and 4 extended constructed response items with 4 score levels. The total number of points for the polytomous items was 67, and the percent correct data reported in Table 3 were computed by dividing the sum of score estimates for the polytomous items by the total possible points (67).

[8] The "final round" for the field trial panel using the ISSE method was Round 2, whereas for the ALS panel using the Mean Estimation method it was Round 3.

Table 3
Civics Field Trial 1
Percent Correct Score at Each Cutpoint Computed for Individual Panelists for Rounds 1 and 2 Using the ISSE Rating Method

Panelist	Items	Round 1			Round 2		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
1	MC	40.4%	74.5%	97.9%	53.2%	93.6%	97.9%
	CR	39.7	62.1	79.3	53.4	75.9	86.2
	All	40.0	67.6	87.6	53.3	83.8	91.4
2	MC	36.2	78.7	95.7	83.0	95.7	100.0
	CR	36.2	69.0	87.9	60.3	86.2	93.1
	All	36.2	73.3	91.4	70.5	90.5	96.2
3	MC	38.3	80.9	100.0	29.8	72.3	100.0
	CR	43.1	65.5	91.4	37.9	60.3	86.2
	All	41.0	72.4	95.2	34.3	65.7	92.4
4	MC	17.0	42.6	80.9	31.9	74.5	97.9
	CR	46.6	70.7	87.9	55.2	79.3	93.1
	All	33.3	58.1	84.8	44.8	77.1	95.2
5	MC	63.8	95.7	100.0	48.9	85.1	100.0
	CR	60.3	82.8	96.6	56.9	79.3	98.3
	All	61.9	88.6	98.1	53.3	81.9	99.0
6	MC	53.2	80.9	100.0	27.7	72.3	95.7
	CR	46.6	84.5	98.3	34.5	60.3	81.0
	All	49.5	82.9	99.0	31.4	65.7	87.6
7	MC	66.0	95.7	100.0	57.4	83.0	100.0
	CR	50.0	82.8	98.3	43.1	62.1	84.5
	All	57.1	88.6	99.0	49.5	71.4	91.4
8	MC	40.4	76.6	100.0	34.0	72.3	100.0
	CR	53.4	69.0	91.4	51.7	70.7	93.1
	All	47.6	72.4	95.2	43.8	71.4	96.2

EVALUATION OF CUTSCORES BY ITEM TYPE

Cutscores for dichotomous item ratings have generally been found to be lower than those for polytomous item ratings. As these data show, however, the estimates of performance for dichotomous items are generally more demanding than those for polytomous items when the comparison is based on "raw ratings." This pattern is particularly clear for ratings at the Proficient and Advanced levels. Most panelists expected students to perform less well on constructed response items than multiple choice items. For two rounds of ratings, 8 panelists generated estimates for 48 individual cutscores—a total of 96 when cutscores are subdivided by item type. Of those, ratings for multiple choice items were higher than for constructed response items in about 71% of the cases. These raw rating data are not good estimates of the actual cutscores because they do not take account of item difficulty. They do, however, provide an indication that panelists' expectations of performance for polytomous items are not necessarily more demanding than for dichotomous items.

EVALUATION OF INTERJUDGE CONSISTENCY

Interjudge consistency is an important criterion in judging the reasonableness of the achievement levels-setting process. The standard deviations of the cutscores are indicators of interjudge consistency. In general, the standard deviations decrease across rounds of ratings, which indicates that the judges' ratings become more similar and less variable. This pattern occurred for the standard deviations of the cutscores set by the civics field trial panelists at the Proficient and Advanced levels. At the Basic level the standard deviations actually increased.[9] The standard deviations for the civics field trial were fairly high because of the relatively small number of panelists and the few rounds of ratings.

Another indication of interjudge consistency is the data from the rater location charts. During the process, ACT staff checked the Rater Location Charts for irregularities, such as exceptionally low or high cutscores, and cutscores that were nearly identical for different achievement levels. If these problems occurred, staff discussed them with panelists and cleared up any confusion about the process. TACSS examined interjudge consistency data as part of their overall review of the research findings.

ACT staff inspected rater location charts to determine whether individual panelists adjusted their cutscores in directions and magnitudes that appeared to be logical and reasonable (see Appendix D for rater location charts). In general, the changes in cutpoints made by most panelists were judged to indicate that they were responding to feedback in a logical manner. Visually inspecting the rater location charts for the first civics field trial revealed that panelists tended to set cutscores that were spread across the ACT NAEP-like scale. This was particularly true for ratings at the Basic level for Round 1 and Round 2. Most of the cutscores for Proficient and Advanced for Round 2 were more clustered with less spread than for Round 1 (see Table 4.)

Table 4
Range of Individual Cutscores Displayed on Rater Location Charts for Civics Field Trial 1

	Range of Cutscores	
	Round 1	Round 2
ISSE Method (n=8)		
Basic	139 – 160	137 – 166
Proficient	161 – 181	164 – 177
Advanced	177 - 199	181 - 192

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 (e.g., Berk, 1994). Indicators of intrajudge consistency across rounds include both the magnitude of change in item ratings from round to round, and the number of item ratings changed from round to round. If panelists make extreme adjustments to their ratings it would indicate that they had little confidence in their

[9] Comparable standard deviations of the cutscores for the Geography ALS could not be computed for these items because Geography ALS panelists in different rating groups rated the blocks of items included in the field trial.

ratings. If panelists make changes of great magnitude to item ratings, or change a large number of item ratings after Round 2 for example, this would signal that panelists were confused or that they had not understood the process. Indicators of intrajudge consistency within rounds inform panelists about the consistency of their individual item ratings relative to their overall estimate of student performance at the cutscore.

INTRAJUDGE CONSISTENCY ACROSS ROUNDS

Panelists' individual consistency across rounds can be evaluated using the percentages of items for which the ratings were changed from round 1 to round 2. Panelists typically make far more changes to their item ratings from Round 1 to Round 2 than they make from Round 2 to Round 3. Panelists had only two rounds of ratings for the field trials, so this measure is somewhat less useful as an indicator of intrajudge consistency. After reviewing the feedback presented following round 1, panelists were given the opportunity to change their ratings for round 2. The results of these changes have been presented as percentages of item rating changes in Table 5. The largest percentages of ratings were unchanged at *all* levels for *all* panelists. The largest number of changes occurred at the Basic and Proficient levels, with ratings for fewer items changed at the Advanced level**.** Note that three panelists made no changes in their Advanced item ratings for round 2, and *no* panelist raised the Advanced item ratings at round 2.

Table 5
Percentages of Changes in Item Ratings from Round 1 to Round 2 Using the ISSE Method
Civics Field Trial 1

	Basic			Proficient			Advanced		
Panelist	Raise	No Change	Lower	Raise	No Change	Lower	Raise	No Change	Lower
1	31.9	60.9	7.2	27.5	71.0	1.4	0.0	98.6	1.4
2	52.2	47.8	0.0	26.1	51.0	0.0	0.0	97.1	2.9
3	0.0	89.9	10.1	2.9	84.1	2.9	0.0	95.7	4.3
4	21.7	75.4	2.9	31.9	66.7	1.4	0.0	100.0	0.0
5	0.0	87.0	13.0	1.4	87.0	1.4	0.0	100.0	0.0
6	8.7	56.5	34.8	8.7	59.4	31.9	0.0	82.6	17.4
7	1.4	85.5	13.0	0.0	71.0	29.0	0.0	88.4	11.6
8	4.3	85.5	10.1	5.8	88.4	5.8	0.0	100.0	0.0
Average	15.0	73.6	11.4	13.0	72.3	9.2	0.0	95.3	4.7

Note: There were a total of 69 items in the rating pool. Percentages are based on 69 item ratings.

INTRAJUDGE CONSISTENCY WITHIN ROUNDS

ACT has considered various means of providing panelists with feedback about their ratings and expected student performance during the standard setting process. A satisfactory method for presenting these data to panelists has been challenging to develop. Several different formats have been used for the ALS processes from 1992 through 1994. There was little evidence, however, that panelists understood this feedback or how to use the information when forming their judgments about student performance.

One of the major attractions of the ISSE method to ACT was the opportunity to produce intrarater consistency feedback for panelists in a format that would be informative to each panelist, regardless of the panelist's level of intrarater consistency. One version of intrajudge consistency

feedback has been included in Appendix F. Although panelists were not shown these data during the field trial, the sample format illustrated one way these data could be presented.

The format listed the assessment items from the most difficult to the least difficult with a brief description of the content of the item. Each chart listed the ranked items and identified each as a multiple choice (MC) or constructed response (CR) item. Following the items were columns of coded ratings for Basic, Proficient and Advanced levels. Ratings for a multiple-choice item were coded "0" for an incorrect response, and "1" for a correct response. NAEP uses a partial credit model for scoring constructed response items. The constructed response items were coded "1" for an incomplete response, "2" for a partial response, and "3" for a complete response, for example. Scores on the ACT NAEP-like score scale were listed sequentially from highest to lowest.

By studying the chart, panelists were expected to be able to understand the relationship of their ratings of the items, the relative level of item difficulty, and the ACT NAEP-like scale score associated with each**.** Easier item would likely be rated as "correct" or "complete" at all three achievement levels. As the degree of item difficulty increased, the item rating pattern would show patterns to reflect this increasing difficulty. For example, a relatively more difficult item would be rated as correct or complete at the Proficient and Advanced levels, but not at the Basic level. The most difficult items would be rated correct or complete primarily at the Advanced level. Similarly, "correct" ratings and higher expected score ratings would cluster at the highest scale scores and decrease as the scale scores decreased.

To produce this type of intrajudge consistency feedback, each item in the rating pool was mapped onto the performance scale; i.e., the ACT NAEP-like score scale. The response probability criterion of 65% was used for mapping items. Each multiple choice item was mapped to the ACT NAEP-like scale using a response probability of 74%; that is, 65% plus correction for guessing.[10] Each response score for each constructed response item was mapped to the ACT NAEP-like scale where the expected score was 65%. The figures in Appendix F illustrate these data as intrajudge consistency feedback charts for each panelist. Figures 1a-1i were based on the ratings that each panelist provided in round 1, and Figures 2a-2i were based on ratings for round 2. Although ACT reviewed the intrarater consistency feedback using this format as part of data analysis, these data were not used as feedback in the field trial.

EVALUATION OF PANELISTS' COMMENTS AND RESPONSES TO PROCESS EVALUATION QUESTIONNAIRES

To better understand their perceptions of the rating processes, panelists were asked to respond to three process evaluation questionnaires. For comparison purposes, the responses of panelists who participated in the geography and U. S. history ALS have been included in Tables 6 and 7. The complete set of responses to the process evaluation questionnaires has been included in Appendix G.

[10] RP=.74 is the mapping criterion used for multiple choice items with four response choices. This reflects a correction for guessing for four-choice items when RP=.65 is the mapping criterion. The formula is RPc = RP + (1-RP)/k, where RP is response probability used as the mapping criterion (65%), RPc is the response probability corrected for guessing, and k is the number of alternatives in multiple choice items.

Table 6
Summary of Responses to Questions Related to Ratings
ISSE Civics

Questions	Round	Responses						Geography*			U.S. History*		
		5	4	3	2	1	Mean	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 first/second rating session were:	1	1	3	4	0	0	3.63	3.75	3.89	3.76	4.46	4.10	3.96
	2	5	3	0	0	0	4.63	4.79	4.75	4.55	4.70	4.72	4.67
(5=Absolutely Clear; 1=Not at all Clear)	3							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 first/second rating session was:	1	1	6	1	0	0	4.00	3.86	4.07	4.14	4.29	4.10	4.89
	2	5	3	0	0	0	4.63	4.82	4.75	4.62	4.63	4.72	4.58
(5=Totally Adequate; 1=Totally Inadequate)	3							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 first/second rating session was:	1	1	4	3	0	0	3.75	3.21	3.14	3.62	4.54	4.14	4.07
	2	3	4	1	0	0	4.25	4.29	4.18	4.21	4.12	4.32	4.12
(5=Far too Long; 1=Far too Short)	3							4.29	4.18	4.21	4.12	4.32	4.12
4. The most accurate description of my level of confidence in the ratings I provided to represent the three achievement levels during the first/second rating session is that I was:	1	1	4	3	0	0	3.75	3.21	3.14	3.62	4.54	4.14	4.07
	2	3	4	1	0	0	4.25	4.29	4.18	4.21	4.12	4.32	4.12
(5=Totally Confident; 1=Not at all Confident)	3							4.48	4.32	4.52	4.39	4.48	4.11
5. The method for rating multiple-choice items was conceptually clear.	1	1	4	2	0	0	3.86	4.07	4.00	3.76	4.11	3.97	4.22
	2	5	3	0	0	0	4.63	4.18	4.32	4.24	4.33	4.43	4.52
(5=Totally Agree; 1=Totally Disagree)	3							4.44	4.36	4.41	4.52	4.43	4.59
6. The method for rating multiple-choice items was easy to apply.	1	2	5	0	0	0	4.29	3.82	3.71	3.83	3.89	3.83	3.89
	2	4	4	0	0	0	4.50	4.18	4.14	4.21	4.19	4.32	4.41
(5= Totally Agree; 1=Totally Disagree)	3							4.33	4.14	4.38	4.44	4.32	4.44
7. The method for rating constructed response items was conceptually clear.	1	1	5	1	0	0	4.00	3.82	3.89	3.76	4.04	3.83	3.96
	2	2	4	1	1	0	3.88	4.25	4.25	4.10	4.33	4.17	4.22
(5=Totally Agree; 1=Totally Disagree)	3							4.30	4.32	4.38	4.48	4.39	4.44

Table 6 (continued)

Questions	Round	Responses						Geography*			U.S. History*		
		5	4	3	2	1	Mean	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)	Grade 4 (n=26)	Grade 8 (n=25)	Grade 12 (n=26)
8. The method for rating <u>constructed response</u> items was easy to apply.	1	1	4	2	0	0	3.86	3.50	3.70	3.90	3.79	3.48	3.59
	2	2	4	1	1	0	3.88	4.04	3.89	4.03	4.11	4.07	3.89
(5=Totally Agree; 1=Totally Disagree)	3							4.26	4.07	4.28	4.19	4.14	4.33

* These data are included for your information.

17

Table 7
Summary of Responses to Questions About the ALS Process

Questions	Responses 5	4	3	2	1	Mean	Geography Grade 4 (n=30)	Geography Grade 8 (n=28)	Geography Grade 12 (n=31)	U.S. History Grade 4 (n=26)	U.S. History Grade 8 (n=25)	U.S. History 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)	1	5	2	0	0	3.88	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)	0	5	3	0	0	3.63	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 by best judgment in rating items to set achievement levels for the NAEP assessment. (5=To a Great Extent; 1=Not at All)	4	4	0	0	0	4.50	4.30	4.29	4.21	4.14	4.28	4.15
4. I feel that this NAEP ALS process produced achievement levels that are defensible: (5=To a Great Extent; 1=Not at All)	3	3	1	0	0	4.29	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 be considered reasonable: (5=To a Great Extent; 1=Not at All)	2	4	1	0	0	4.14	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 use of achievement levels resulting from this ALS procedure:	Yes, definitely					4	53.6%	60.7%	57.1%	17.9%	31.0%	51.9%
	Yes, probably					2	35.7	32.1	42.9	53.6	44.8	44.4
	No, probably not					1	7.1	7.1	0.0	21.4	6.9	0.0
	No, definitely not					1	0.0	0.0	0.0	7.1	0.0	0.0

Given that the essential elements of the five-day ALS process were condensed into an intense two-day process for the field trial, it was reasonable to expect somewhat less positive responses from the field trial panelists than the ALS panelists. However, the overall responses of the field trial panelists were positive when evaluating the processes. No response patterns emerged to suggest significant irregularities in panelists' reactions to the ISSE method or the process for setting achievement levels using the method.

When asked to evaluate the instructions for the rating sessions, the responses of panelists using the ISSE method were similar to those of geography and U.S. history ALS panelists using the Mean Estimation method (ISSE = 4.63; ME 4.75 (geography) and 4.72 (history) when 5 = *absolutely clear* and 1 = *not at all clear*). Panelists indicated that the amount of time they had to complete the tasks during the rating sessions was slightly more than they needed (3.6 and 3.5 when 5 = *far too long*; 1 = *far too short*), suggesting that they did not feel rushed during the rating sessions.

In general responses became more positive after Round 1, a finding consistent with the pattern typically evidenced by ALS panelists. That is, panelists' responses have generally reflected an increase in understanding and confidence as the rounds of ratings progressed. The exception to this general trend was the question that referred to the clarity of conceptualizing the rating method to constructed-response items. The responses of the field trial panelists using the ISSE method were slightly less positive for Round 2 (mean = 3.88) than Round 1 (mean = 4.00). Further, these responses were less positive than those of the ALS panelists using the Mean Estimation method (ISSE = 3.88; ME = 4.25 (geography) and 4.17 (history)). When panelists were asked if the method for rating constructed response items was easy to apply, the responses of the field trial panelists were about the same for Round 1 (mean = 3.86) and Round 2 (mean = 3.88). Typically, responses become more positive. After the final round (Round 2), the civics field trial panelists responses were noticeably lower than those of the ALS panelists after the final round (Round 3) (ISSE = 3.88; ME = 4.07 (geography) and 4.14 (history)).

Additional questionnaire data indicated panelists' level of confidence in the results produced by the ISSE method, which was lower than that of panelists using the Mean Estimation method (ISSE = 3.88; ME = 4.14 (geography) and 4.12 (history) when 5 = *totally confident* and 1 = *not at all confident*). When asked if they would be willing to sign a statement recommending the use of achievement levels resulting from this standard setting procedure, six panelists responded positively and two negatively.

EVALUATION OF THE COGNITIVE COMPLEXITY OF THE RATING TASKS

As mentioned earlier, major criticisms of the NAEP ALS process implemented by ACT have focused on the item-by-item rating methods. The estimation tasks were thought to be too cognitively complex for panelists to perform accurately (NAE, 1993; NRC 1999). Research findings related to this issue (Impara & Plake, 1997; 1998) have indicated that the level of cognitive complexity would be less for the ISSE method than the modified-Angoff method. It was argued that panelists would find it easier to select a "yes" or "no" to indicate if students would answer an item correctly than to select a percentage to indicate the likelihood of students answering an item correctly.

The issue of cognitive complexity was included as an integral part of evaluating the ISSE rating method. Responses to the evaluation questionnaires reflected the perceptions of panelists regarding the level of cognitive complexity required to perform the rating task, among other questions. The responses of field trial panelists when asked about their understanding of the ISSE rating method were similar to responses of geography and U.S. history panelists using the Mean Estimation method (ISSE = 4.63; ME 4.75 and 4.72 when 5 = *totally adequate* and 1 = *totally inadequate*). There was no evidence to suggest that panelists were unable to perform the estimation tasks involved in the process.

EVALUATION OF PANELISTS' RESPONSES TO CONSEQUENCES DATA QUESTIONNAIRES

In addition to the feedback data resulting from their second round of ratings, panelists were given information about the consequences of the ratings they had provided. That is, panelists were told the percentage of students scoring at or above each achievement level based on the cutpoints they set for the second round of ratings.

Panelists engaged in a lengthy group discussion about the consequences information. There was little interest in changing the cutpoints. (Please see Table 8.) Five of the eight panelists recommended that the cutpoints for Basic and Proficient be reported as set. Six recommended that the cutpoint for Advanced be reported as set. In general, the adjusted Basic cutscore was slightly higher, the adjusted Proficient cutscore remained about the same, and the adjusted Advanced cutscore was slightly lower than the previous cutscores.

Table 8
Frequencies of Recommended Changes to Cutpoints
After Receiving Consequences Data

Recommended Cutpoints (n=8)	Basic	Proficient	Advanced
As set	5	5	6
Lower	1	1	1
Higher	2	2	1

New consequences data were computed based on the recommendations collected after the first set of consequences data described above. These data were distributed to panelists who were allowed to discuss the adjusted cutpoints and updated consequences data. They were asked to state the final cutscores that they would recommend to NAGB, if they had the opportunity. During the discussion they decided that the group cutscores were to be their final "recommendation to NAGB." Only one person suggested a change from the group cutscores. He felt that the cutscores should each be lowered by one standard deviation because he thought this would better reflect "borderline performance." After the discussions, panelists were again asked to fill out a questionnaire regarding the consequences data and their final recommended cutscores. Only one panelist recommended that the percentages of students performing at all levels should be larger. Table 9 displays the consequences data that resulted from adjusting cutscores.

Table 9
Percentages of Students Performing at or Above Each Achievement Level Based on Cutscores Adjusted During the ALS Process for Civics Field Trial 1

Achievement Level	%≥ Round 1 Cutscores	%≥ Round 2 Cutscores	Final Recommendations
Basic	67.2	61.6	64.3
Proficient	11.6	11.6	13.6
Advanced	0.2	0.3	0.4

DETECTION OF BIAS IN ISSE CUTSCORES

The Technical Advisory Committee on Standard Setting (TACSS) reviewed the outcomes of the first civics field trial and expressed some reservation about the cutscores computed from the ISSE ratings. The ISSE method was found to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true scores" or judgments of panelists (Reckase, 1998b; Bay, 1998; Reckase & Bay, 1999). The authors explained:

> A simple example can illustrate the cause of the bias. Suppose ten multiple-choice items possessed identical statistical characteristics and measured the same construct. Suppose further that a judge intended to set the standard as 8 on the raw score scale. For the ISSE method, the judge would assign a 0 or 1 to each of the ten items depending on whether or not a student performing at the standard would answer the item correctly. The likelihood of a correct response would have to be .8 for a person at the standard, producing a total score string of 8. But the ISSE method does not permit this option, forcing a 1 to be the most likely score. If the judge accurately assigned the most likely score to each of the ten items, s/he would produce a string of ten 1s, resulting in a standard of 10, rather than 8. The bias would be positive if the standard is set at more than 50% correct and negative if the standard is set at less than 50% correct. The bias becomes more extreme as the standard deviates from 50% correct (Reckase & Bay, 1999, p. 3).

As a result of Reckase's work and the findings from the first set of field trials in both civics and writing, TACSS recommended that ACT discontinue further research using the ISSE method. The method was eliminated as a possibility for implementation in the Civics NAEP achievement levels-setting meeting.

DEVELOPING A NEW RATING METHOD: THE RECKASE METHOD

Having eliminated the ISSE method, two achievement levels-setting methods were considered for further study in a second civics field trial: the Mean Estimation method with item mapping added for the final "rating" round, and a methodology developed by TACSS member Mark Reckase, now called the Reckase method (Reckase, 1998a). In keeping with ACT's intent to explore new procedures for collecting and summarizing judgments, the Reckase method showed promise in meeting the unique challenges inherent to setting achievement levels for NAEP. "The method is particularly designed to fit with the scaling of the NAEP assessments and using the procedures

typically used by NAGB to define the scores that set the boundaries for achievement levels." (Reckase, 1998a, p. 1).

The Reckase method addressed two major issues associated with setting NAEP achievement levels. The first issue related to providing a format for intrajudge consistency feedback. The Reckase charts were designed to meet the challenge of giving panelists information about the relationship of their ratings to item characteristics in a technically accurate and readily understandable way. The method greatly advanced efforts to inform panelists of the correspondence between their ratings and expected student performance. The second issue related to differences in cutscores by item types, e.g. constructed response and multiple choice. The design of the Reckase chart would also reveal any pattern of differences in ratings of the kinds of items included in the assessment. Like other methods considered by ACT for the 1998 ALS process, the Reckase method also involved a simple way for panelists to adjust cutscores for their final round of ratings. TACSS reviewed the proposed Reckase method and recommended that ACT investigate its merits through continued field trial research.

KEY ASPECTS OF FIELD TRIAL 2 FOR CIVICS

The second civics field trial was conducted at the Radisson Hotel, Highlander Plaza in Iowa City, Iowa on Wednesday and Thursday, June 24 and 25, 1998. The primary purpose of the second field trial was to identify the procedures that would be used for the 1998 ALS process. The ISSE method had been eliminated from further consideration and ACT was committed to researching new methods to collect panelists' judgments. For this field trial, ACT studied the new Reckase standard setting method (Reckase, 1998a) as an alternative to the Mean Estimation method. In addition, ACT studied a procedure that combined item mapping with mean estimation. Item maps are used to represent cutscores relative to items in the "bookmark" method used by CTB-McGraw. The "bookmark method" has gained popularity as an easy-to-use standard setting method. ACT has studied item maps previously, but not as a standard setting method. The plan was to study the "interface" of item maps with the Mean Estimation method. The interface between item mapping and the Mean Estimation method is of interest because item maps offered an easy-to-use method for adjusting cutscores without evaluating each item one by one. And finally, ACT also studied the effect of consequences data given to panelists during the item rating process. Appendix H shows the diagrams of the various research designs proposed for Field Trial 2.

The same criteria were used to evaluate and compare both methods, that is the Reckase method (MR) and the Mean Estimation method with Item Maps (ME/IM). The methods were evaluated on the basis of judgements of the reasonableness of the cutpoints and their standard deviations; the level of interjudge and intrajudge consistency; the ease with which the method was implemented by panelists; and panelists' satisfaction with and confidence in the process and the results of the process. Logistic concerns and the reactions of panelists were particularly important criteria for evaluating the methods during this stage of the process. The two methods would be compared and evaluated, and a decision would be reached regarding the method to be implemented for the pilot study and the Civics ALS process. A variety of information would contribute to the choice of method(s) to recommend to NAGB.

PANELISTS

Because four groups (two "methods" groups each divided into two "consequences" groups) were to be used in the field trial, TACSS recommended that no fewer than 40 panelists be recruited for the study. Panelists were assigned to one of the four experimental groups consisting of at least 10 panelists each. As nearly as possible, the composition of the panels was to meet the same criteria as that required for the ALS panels: 55% teachers (TR), 15% nonteacher educators (NT), and 30% members of the general public (GP).

RECRUITMENT

The first effort to recruit panelists for the second civics field trial around the Iowa City area began in January of 1998. The response from educators was poor, primarily due to heavy teaching responsibilities. The second field trial was originally scheduled for Saturday and Sunday, March 14-15, 1998. It was postponed until the end of the school year when educators said they would be available to participate.

A second effort to recruit panelists began in May without much improvement in participation rates. To better understand the cause of the low response rate, ACT conducted a small telephone survey of school district educators who suggested that participation would increase if the study were held during the week, rather than the weekend. In an effort to improve participation rates, the dates of the study were changed to weekdays.

As an added incentive, TACSS recommended increasing the honorarium from $100 to $300. NAGB authorized this change. Each participant was paid an honorarium of $300 and travel expenses. ACT had no problem recruiting the panelists after the honorarium was raised and the two-day meeting was rescheduled for weekdays.

School district administrators and Area Education Agencies in nearby areas of Iowa, Wisconsin and Illinois were asked to identify outstanding teachers, educators, and qualified members of the general public as possible panelists. Elected officials in Johnson and surrounding counties in Iowa were also asked to nominate panelists. Many individuals were nominated from the membership of the Geographic Alliance of Iowa.

Nominees were accepted as panelists if they fit the general description of "panelist type," were familiar with the knowledge and skills of eighth graders in the area of geography/social studies, and were willing to participate in the study. The nominees were confirmed as panelists as they were recruited until the goal was reached of 44 committed panelists. Later one panelist dropped out. About 74% of the 43 panelists were teachers (TR), 12% were educators who were not teaching in the classroom (NT), and 14% were knowledgeable members of the general public (GP). Sixty-three percent of the panelists were men and 37% were women. All of the panelists were white. Please refer to Table 10 for more details about the composition of the panel.

Table 10
Composition of Civics Field Trial 2 Panel

	Type			Gender		
Group	TR	NT	GP	M	F	Total
A	8	1	1	6	4	10
B	8	1	2	7	4	11
C	8	2	1	7	4	11
D	8	1	2	7	4	11
	32	5	6	27	16	43

TABLE DISCUSSION GROUPS

Appendix H shows the diagrams of the various designs proposed for Field Trial 2. As the diagrams illustrate, panelists within each rating method group were further divided into two different groups: those who would receive consequences data after each round of ratings and those who would not. There were two table groups seated at tables of about five persons per table. There were approximately 10 panelists in each of the four treatment groups. The demographic attributes of panelists were considered when assigning members to treatment groups and to the table groups within each treatment group. The goal was to have each group as equal as possible with respect to panelist type, gender and race. To the extent possible, panelists from the same geographic location (e.g., city or school) were assigned to different table groups.

DATA, ACHIEVEMENT LEVELS DESCRIPTIONS, AND ITEM RATING POOLS

NAEP Geography data from 1994 were used for field trial 2 because administration of the Civics NAEP had not been completed and, consequently, civics data were not available. The choice of geography data was made for the same reasons that led to this choice for field trial 1.

The grade 8 Geography NAEP consisted of 125 items divided into 7 blocks. Only four blocks were used for the field trial, totaling 69 items. A fifth block of 16 items was used for practice ratings. The same 4 blocks of items had been selected for the Similarities Classification Study (SCS) that was conducted as validation research for the 1994 Geography ALS (ACT, 1995). The blocks for the SCS had been selected to maximize the representativeness of the assessment features of the entire grade 8 item pool. Item format, content coverage, reliability and item characteristics of difficulty, discrimination, and so forth were the features represented (Carlson, 1995). Given the limitation that the field trial was planned for two days, in contrast to the five-day ALS meeting, rating four blocks of items seemed the maximum number of blocks to use for rating. Given that the purpose was to compare the two rating methods, however, four blocks seemed the minimum number of blocks to use.

THE ALS PROCESS DESIGNED FOR CIVICS FIELD TRIAL 2

The recommended design divided panelists into four groups (A-D) composed of ten members each. Groups A and B used the Mean Estimation method in combination with item maps that were introduced at Round 3 for panelists to mark their cutscore. Groups C and D used the Mean Estimation method in combination with the Reckase method that introduced the Reckase Charts at round 2 and required panelists to mark their cutscores for Round 3. Groups A and C were given

consequences data after each round of ratings in the process, whereas groups B and D were given consequences data beginning after round 3 in the process.

The third rating session did not involve item-by-item ratings for either method. For round 3, panelists using the Reckase method were asked to choose a single row or scale score to represent their ratings for each achievement level. Panelists were instructed to draw a horizontal line through the row at the selected cutscore. The IRT-based performance estimates for each item associated with each cutscore were, in effect, their round 3 ratings.

Similarly, panelists using the Mean Estimation method were given item maps and lists of items ordered on difficulty. Panelists were asked to mark their Round 3 cutscore for each achievement level directly on the item map.

DESCRIPTION OF THE RECKASE METHOD

The Reckase method used the Mean Estimation (ME) item-by-item rating procedures for the first round of ratings. Panelists estimated the percentage of students performing at the borderline of each achievement level that would correctly answer each multiple-choice item. They also estimated the average (mean) score that those students would get for their response to each constructed response item. Those estimates were then transferred onto charts, now called Reckase Charts. The Reckase Charts displayed columns that represented the IRT-based performance estimates for each assessment item in the rating pool, and the rows represented IRT-based performance estimates for each score point on the ACT NAEP-like scale. For dichotomous items, the performance value was the probability of correct response at each scale score point for each item. For polytomous items the performance value was the expected score for each item at each scale score. The expected performance across scale score points could be observed, as could the expected performance across items for students scoring at a particular scale score. A sample Reckase Chart has been included in Appendix I. The accompanying instructions can be found in Appendix C.

Panelists marked their charts by circling the expected performance score that corresponded to their item-by-item ratings for the first round of judgments. They marked their ratings for each item and each achievement level on the charts. They then connected the circles to see their rating patterns graphically. They also marked the group cutscore and their own cutscore for each achievement level on the charts. By examining the charts, the judges were able to consider the relationship between their estimates of student performance for each item and the expected student performance at the cutscore. Further, judges considered any observable patterns in their ratings, such as patterns of different estimates for multiple choice and constructed response items, or patterns of different estimates for items assessing different content areas, or patterns of ratings above or below their own or the group cutscore(s) for a particular block of items. Panelists were instructed that if their judgments of students performing at the borderline of each achievement level exactly fit the estimates generated by the model based on actual student performance, all of their ratings would fall along a single row. In other words, if panelists' ratings were on a single row, their ratings perfectly matched estimated student performance. They were reminded that their judgments were to be based on the achievement levels descriptions and how students at the borderline of each achievement level would perform on the items.

Panelists considered this information along with other types of feedback when rating items a second time. After the second round of ratings, a new set of Reckase Charts was distributed to panelists, and they transferred their Round 2 item ratings from their rating forms to the Reckase Charts and evaluated them again. After evaluating all of their Round 2 feedback, panelists using the Reckase method were asked to decide on a cutscore and mark it on the chart. The model-based estimates were, in effect, their Round 3 ratings. Rather than engage in a third round of item-by-item ratings, however, panelists selected three scores on the ACT NAEP-like scale to represent their cutscores for the three achievement levels.

DESCRIPTION OF THE MEAN ESTIMATION AND ITEM MAPPING METHOD

Panelists participated in two rounds of item-by-item ratings using the Mean Estimation procedure. This was the same procedure used by ACT in several earlier studies and the procedure used for the item-by-rating tasks for the first two rounds in the Reckase method. Panelists estimated the probability that students performing at the borderline of each achievement level will correctly respond to each multiple-choice (dichotomously-scored) item. They estimate the average (mean) score of students performing at the borderline of each achievement level for each constructed-response (polytomously-scored) item. After two rounds of item-by-item ratings, panelists worked with item maps. ACT had extensive experience with the rating method, and several research studies had been conducted with item maps. This field trial was an opportunity to determine how well the two methods could be used together in a single standard-setting process.

Panelists were given a list of the assessment items in a column, from the least difficult to the most difficult, followed by a brief description of the content of the item. The map itself was a graphic representation of where items were located or "mapped" onto the ACT NAEP-like scale based on a response probability value. Panelists could read the actual items and determine if the collection or "domain" of items matched the knowledge and skills students should have, as described by the achievement level descriptions. Panelists were instructed that a few items might be included within a domain that would be judged to be inappropriate, relative to the ALD for that level. Such instances should generally be disregarded. Panelists were urged to look at the domain of performance for each achievement level to judge whether the cutscores should be adjusted. Panelists could adjust their cutscores for Basic, Proficient, or Advanced by raising or lowering their ACT NAEP-like scores on the item map. Panelists marked their cutscore for each achievement level on the item map, and they recorded the score on the form. Please see Appendix K for a sample item map. A copy of the instructions presented to panelists regarding the use of item maps can be found in Appendix C.

STEP 1: BRIEFING MATERIALS

Before the meeting, all panelists were mailed a set of materials that contained important background topics and introductory information on setting achievement levels. The briefing materials included the following information:

- 1994 NAEP *Geography Framework*;
- 1994 NAEP Geography Achievement Levels Descriptions;

- *Multiple Challenges*, a booklet on the 1998 NAEP;
- NAGB brochure;
- *The NAEP Guide*;
- Cover letter with instructions for preparing for the field trial;
- Item-Use and Nondisclosure Agreement;
- Check Request Form;
- Request for Taxpayer I.D. Number and Certification;
- Directions and a map to the meeting.

STEP 2: GENERAL ORIENTATION AND TRAINING EXERCISES (DAY 1)

Although the study only lasted two days, all elements of the five-day achievement levels-setting process were covered, at least to some extent. The agenda for the Civics Field Trial 2 have been included in Appendix B. One agenda was for the groups that used the Mean Estimation method with Item Maps (groups A and B), and the other was for the groups that used the Reckase method (groups C and D). All groups received the same orientation, training, and instructions to prepare them for the first round of ratings.

Panelists were provided an abbreviated orientation to the achievement levels-setting process and the procedures planned for the field trial. The orientation session included a general introduction to the NAEP program and an overview of the method used to develop NAEP achievement levels presented by a member of the NAGB staff. Panelists participated in several training exercises that were the same as those included in the training of ALS panelists.

TAKING A FORM OF THE NAEP

Panelists were administered a form of the Geography NAEP, and they reviewed their own responses relative to the scoring rubrics.

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS (ALDS)

Panelists were given time to work with the ALDs and to discuss them, but they were not given the opportunity to revise them. The final version of the grade 8 geography achievement levels descriptions (ALDs) was used for the field trial. Panelists participated in exercises to help them become familiar with ALDs. They examined the scoring rubric and assessment items, both multiple choice and constructed response. Panelists were asked to think about the performance required for students to answer the item correctly, or to score at each rubric point with respect to the descriptions of geography knowledge and skills associated with the ALDs. They then identified each item with one of the achievement levels descriptions.

In a similar exercise, panelists were given several student booklets to review. The booklets represented various performance levels. Panelists were asked to think of knowledge and skill levels represented in each booklet, as a whole, and to classify the booklets according to the ALDs.

The combination of exercises focused panelists on the achievement levels descriptions and caused them to think of how the assessment items measured knowledge and skills described in each. The first exercise helped panelists to think about how the achievement levels descriptions applied to both multiple-choice and constructed-response items. The second exercise helped panelist to think of overall performance in a holistic manner. They saw examples that the Proficient level of performance, for example, could be achieved through a variety of performances. The goal was to help panelists view student performance in a compensatory manner rather than to think that Proficient performance required correct answers to all multiple choice items, for example.

UNDERSTANDING BORDERLINE PERFORMANCE

Panelists also received brief training in the concept of borderline performance. They did not write borderline descriptions as part of their training, however.[11] Panelists discussed borderline performance and reached a common understanding of what constituted borderline performance at each achievement level.

STEP 3: THE ITEM RATING PROCESS

ROUND 1 RATINGS (DAY 1)

After orientation, panelists were trained in the Mean Estimation (ME) rating method. All panelists used this rating method for the first two rounds of ratings. Please refer to Appendix C for detailed instructions to panelists on the Mean Estimation method. The rating forms included a percentage sign (%) beside the spaces provided for rating multiple-choice items to remind panelists to record their probability estimate of correct response. For constructed response items, the Mean Estimation method required judges to estimate the average score (e.g., 2.4 on a scale of 1-3) of students performing at the borderline of each level. The score range (e.g., "1-3" or "1-4") appeared on the rating form beside the spaces provided for rating constructed response items. For round 1 ratings, panelists read each item, considered what constitutes a correct response, checked the scoring guide, and marked their rating forms. After rating all items in their item pools, panelists completed a process evaluation questionnaire. That completed Day 1 activities.

Round 2 Ratings (Day 2)

After the groups reviewed the various types of feedback based on their Round 1 ratings, the panels were reconvened into rating-method groups. Groups A and B stayed in one room where they completed Round 2 ratings, using the same procedures as for round 1. They were instructed that they could change their ratings for all, some, or no items at any or all achievement levels. Groups C and D were moved to another room and received instructions in the Reckase method. They used the Mean Estimation (ME) item-by-item rating procedures for the second round of ratings as well. In preparation for that, however, panelists evaluated their Round 1 ratings on Reckase Charts. First, they had to transfer their ratings to Reckase Charts. They also marked

[7] Panelists had been asked to write descriptions of borderline performance as part of the 1996 Science ALS process. Process evaluations by panelists revealed no significantly higher level of understanding of borderline performance for science panelists than had been found for earlier studies that did not include written borderline descriptions.

their charts with their own Round 1 cutscores for each achievement level and with the group cutscores for each level. The Reckase Charts used to review Round 1 ratings did not include scale score values. Instead, letters of the alphabet were used to label the rows. This was done because of concerns that panelists would place too much emphasis on the cutscore and not enough on the achievement level descriptions as their guide for ratings. After evaluating these charts, the panelists using the Reckase method engaged in the Mean Estimation procedure to provide their item-by-item judgments of performance for Round 2. They could change their ratings for all, some, or none of the items at any or all achievement levels.

Round 3 Ratings (Day 2)

The third round of judgments did not require item-by-item performance estimates from panelists in either rating-method group. Panelists in each rating-method group were trained in separate rooms for Round 3 ratings. After discussing and considering the information gained from reviewing the various forms of feedback after round 2, Groups A and B were introduced to item maps. Item maps were presented as an easy and efficient way for panelists to adjust cutscores without going through the item-by-item rating process a third or fourth time. Judges were instructed in how to read the item maps and how to use the information when forming their judgments of student performance. By studying the item maps, judges could understand the relationship of their ratings to the items, the relative level of item difficulty, and the ACT NAEP-like scale score associated with each item.

Panelists were given a list of the assessment items in a column, from the least difficult to the most difficult, followed by a brief description of the content of the item. The map itself was a graphic representation of where items were located or "mapped" onto the ACT NAEP-like scale based on a response probability value. Panelists could read the actual items and determine if the collection or "domain" of items matched the knowledge and skills students should have, as described by the Proficient achievement level descriptions. Panelists were instructed that a few items might be included within a domain that would be judged to be inappropriate, relative to the ALD for that level. Such instances should generally be disregarded. Panelists were urged to look at the domain of performance for each achievement level to judge whether the cutscores should be adjusted. Panelists could adjust their cutscores for Basic, Proficient, or Advanced by raising or lowering their ACT NAEP-like scores on the item map. If no adjustment was made for a cutscore, the grade group cutscore for that level was used in computing the Round 3 cutscores. Please see Appendix K for a sample item map and a copy of the instructions presented to panelists regarding the use of item maps.

Panelists using the Reckase method transferred their Round 2 ratings to the charts and evaluated them again. The ACT NAEP-like scale score values were included on the Reckase Charts used for marking Round 2 ratings. After marking their charts and evaluating their Round 2 ratings on the charts, panelists in Groups C and D were ready to give their Round 3 judgments. They were instructed to select a cutscore for each achievement level and draw a horizontal line through the performance estimates for each item associated with each cutscore. One row was selected to represent the Basic cutscore, one the Proficient, and one the Advanced. The expected IRT-based performance estimates for each item associated with each cutscore were, in effect, their round 3

ratings. The cutscores for all panelists in each group (C and D) were averaged to compute the Round 3 cutscore for each achievement level.

STEP 4: FEEDBACK AFTER EACH ROUND

FEEDBACK AFTER ROUND 1 (BEGINNING OF DAY 2)

At the start of Day 2, panelists were given feedback data based on their round 1 ratings. Panelists received this information together in a general session at the start of the day. Four sets of feedback were produced: one set for each treatment group A, B, C, and D. The feedback data were distributed to panelists in a general session at the start of the second day of the field trial. The various forms of feedback were described, and panelists were instructed in how to interpret and use each type. They were provided with cutpoints and their standard deviations, rater location charts, student performance data, and whole booklet feedback prior to the second round of ratings. (Please refer to the Civics Field Trial 1 report for definitions of these types of feedback.) None of the experimental conditions had been varied for Round 1; all groups had experienced identical procedures. Nonetheless, feedback was not shared across treatment groups. Feedback data were discussed among panelists in table groups after they went to their separate rooms for each rating method. Copies of the feedback based on the first round of ratings are included as Appendix J.

Panelists using the Reckase method received instructions in the Reckase Charts *after* completing their review and discussion of feedback data from Round 1.

FEEDBACK AFTER ROUND 2 (DAY 2)

After the second round of ratings, all panelists were given updated feedback data using the cutpoints set from the second round of ratings. Four sets of feedback data were again prepared and distributed. Panelists received this information together in a general session prior to the third round of ratings. They were again provided with cutpoints and their standard deviations, rater location charts, student performance data, and whole booklet feedback. Panelists reviewed and discussed the data with other members of their treatment group after returning to the meeting room for their rating group. Copies of the feedback based on the second round of ratings have been included in Appendix J.

After reviewing the standard feedback, panelists in Groups A and B were trained in the item mapping procedure. They reviewed the item information provided on the item maps, and marked their cutscores. Panelists in Groups C and D again marked their ratings from the previous round on their Reckase Charts and evaluated them in preparation for the next round of ratings. They were instructed to mark their cutscores on their Reckase Charts.

Feedback After Round 3 (End of Day 2)

Cutpoints for Round 3 were computed by averaging the cutscores recommended by each panelist in each group. The same standard forms of feedback data for each of the four groups were again

prepared and distributed to panelists. Panelists using the Reckase method were not given Reckase Charts for marking their Round 3 ratings.

STEP 5: CONSEQUENCES DATA

Although consequences data are a type of feedback, the timing of when panelists received consequences data was an experimental condition that defined treatment groups. Consequences data had never been provided to panelists during the NAEP ALS process. ACT proposed to use this field trial as an opportunity to study the impact on cutscores and on the overall process of providing consequences data during the process. For that reason, it is being presented as a separate step in the process.

CONSEQUENCES DATA AS FEEDBACK AFTER EACH ROUND

ROUND 1

Following the general session at the start of Day 2 when all panelists were instructed in the various forms of feedback data, groups A and C were moved to another room where they were given information about the consequences of their ratings for round 1. Panelists in groups B and D had a brief break in another location. To be informed about the consequences associated with their cutscores, panelists were told the percentages of students scoring at or above the cutpoint set for each achievement level based on the first round of ratings. After consequences data were explained and displayed in bar graphs and pie charts, panelists were given a chance to discuss this information with persons in their rating group. There were approximately 10 people in each rating group (A and C) who received consequences data. After the discussion, they completed a questionnaire in which they were given the opportunity to recommend new cutpoints that would raise or lower the percentage of students performing at or above each level. These recommendations were collected as part of ACT's research on the impact of consequences data on ratings. The recommendations had no other role in the process. A copy of the questionnaire appears in Appendix E.

ROUND 2

After the feedback data for Round 2 were described and distributed, Groups A and C were moved to another room where they received consequences data for the second time. Groups B and D did not receive consequences data after round 2. These data were updated to reflect the rating group's cutscores based on round 2 ratings. Panelists were given the chance to discuss this information with persons in their rating group, and then they were asked to complete a second questionnaire in which they could again recommended new cutpoints that would raise or lower the percentages of students performing at or above each achievement level. These recommendations were again collected for research purposes only. The recommendations had no other role in the process. The questionnaire was the same as that used with Round 1 consequences feedback.

ROUND 3

In addition to the standard feedback resulting from the cutpoints marked on the item maps and Reckase charts for Round 3, all panelists received consequences data based on the Round 3

cutscores. Groups A and C received consequences data for the third time while groups B and D received consequences data and instructions for their use for the first time.

Panelists were given the chance to discuss the consequences data with members of their treatment group, i.e., with other panelists who had used the same rating method and who had received consequences data together. Panelists in each group (A, B, C, and D) were encouraged to reach agreement on a group cutscore for each achievement level. Judges were asked to complete a questionnaire in which they recommended new cutpoints that would raise or lower the percentages of students performing at or above each achievement level. All panelists were given a questionnaire to record their reactions to and recommendations regarding the consequences associated with round 3 cutpoints. Panelists were informed that *these recommendations would be used* to compute a final cutscore for each achievement level. If a panelist recommended no adjustment to the Round 3 cutscore, the group cutscore for that level was used in computing the new, final cutscore.

FINAL

Groups A and C received consequences data for the fourth time, which were updated by averaging the cutscores panelists recommended on consequences data questionnaire for Round 3 cutscores. Groups B and D received consequences data for the second time, which were updated by averaging the cutscores panelists recommended on the consequences data questionnaire. Panelists were given another consequences questionnaire that included the same questions and allowed the same recommendations. These recommendations were again collected for research purposes only.

STEP 6: EVALUATIONS THROUGHOUT THE PROCESS

Panelists completed five process evaluation questionnaires throughout the meeting. One questionnaire was completed after each round of ratings and one was completed at the end of the entire process. When the last evaluation was finished, the panelists were thanked for their work and the meeting was adjourned. Responses to the questionnaires have been summarized and appear later in this report.

PROCEDURAL IRREGULARITIES

Two irregularities occurred during Civics Field Trial 2. The first was a delay in distributing Consequences Questionnaire #1 to Groups A and C after round 1. The omission was realized about five minutes after the form should have been handed out. Panelists in group A were interrupted while they were completing Process Evaluation Questionnaire #2 and instructed to fill out Consequences Questionnaire #1. Panelists in group C were interrupted from marking their Reckase Charts to complete the questionnaire. Once they finished filling out the consequences questionnaire, they returned to their work. This disruption in procedures did not appear to cause a problem for panelists.

The second unexpected irregularity had to do with the assignment of panelists to table groups. At the start of the meeting, panelists were assigned to one of four groups. There were two tables of

about five persons per table, for each of the four groups. One particular table seemed quite slow in completing the training tasks. In an effort to diminish this problem, panelists were re-assigned to table groups. This time they were assigned on the basis of panelist identification numbers. Table group assignments are made to create the most even distribution possible of panelists by type, gender, and race. By using the panelist ID to assign table groups, the balance was inadvertently eliminated. The arrangement resulted in teachers only sitting at one of the two tables for each of the four groups.

OUTCOMES OF THE CIVICS FIELD TRIAL 2

The field trials were planned as opportunities to try out methods and procedures and collect data based on panelists' reactions to the processes. The process that usually required five days at an ALS meeting was condensed into two days. The field trial probably did not allow adequate time for panelists to digest all the information given to them. Although the number of panelists in each rating group and the length of time involved in this field trial were equal to or greater than those frequently used in other standard setting studies, the numerical results of this field trial do not meet the requirements for a "real" NAEP ALS process.

EVALUATION OF CUTSCORES, STANDARD DEVIATIONS, AND RESULTING CONSEQUENCES DATA

Because of initial misunderstandings of the instructions, four raters (3 in group A and 1 in group B) had extremely low round 1 cutpoints for the Basic level. The instructions were discussed with these panelists, and ratings for subsequent rounds were more typical. In order to avoid losing data, the round 1 mean cutpoint for Basic for the appropriate group (A or B) was computed without these outliers. The resulting value was used to replace the extreme cutpoints of the four raters. Subsequent analyses were performed on the altered data set.

Table 11 displays information about the treatment groups' cutpoints, their standard deviation (SD), and percentage of students performing at or above each achievement level across rounds. The same data from the Geography ALS have been included for comparison purposes. Although there were variations in cutscores across groups, the *relative* consistency across experimental conditions was a striking result. Not only were cutscores across the four groups in this field trial similar to one another, they were also very similar to those computed from ratings in the Geography ALS for the same item blocks.

The overall findings of the second field trial indicated that there were very few statistically significant differences in the cutscores between the Mean Estimation method and the Reckase method overall. Panelists who used the Reckase method tended to raise cutscores more frequently than panelists using the Mean Estimation method. Further, there were no compelling differences in cutscores based on the timing of consequences data. The cutscores produced by groups A and C, who received consequences data early in the process, were not significantly different from the cutscores produced by groups B and D, who received consequences data later in the process. Graphic representations of cutscores and their standard deviations by groups over rounds have been included in Figures 1-4. A brief report on the analyses of differences by rating group and by treatment group is included in Appendix L.

Table 11
Comparison of the Outcomes from the Four Experimental Groups for Civics Field Trial 2 and the Geography ALS

		Civics Field Trial 2 Groups								Geography ALS	
		A ME/IM (n=7)		B ME/IM (n=10)		C ME/MR (n=9)		D ME/MR (n=9)		ME Method	
Round	Level	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)**	%=>
1	Basic	136.4 (25.5)	89.8	146.5 (12.2)	***74.3****	145.9 (9.3)	75.1	138.6 (11.3)	***87.4***	149.2	73.5
	Proficient	157.6 (5.1)	46.3	164.5 (5.1)	***26.6***	163.2 (4.4)	29.7	162.5 (4.5)	***32.0***	163.5	31.0
	Advanced	168.9 (5.8)	16.0	175.3 (4.1)	***6.0***	173.9 (3.6)	7.9	174.1 (5.0)	***7.4***	173.1	8.5
2	Basic	146.7 (7.8)	73.5	151.3 (5.1)	***63.5***	148.4 (7.4)	70.0	143.5 (8.6)	***79.5***	151.8	67.1
	Proficient	161.0 (4.8)	35.8	166.6 (3.1)	***21.6***	165.6 (2.7)	23.5	162.5 (4.7)	***32.0***	164.6	28.8
	Advanced	171.8 (5.9)	10.9	177.3 (2.7)	***4.0***	177.0 (3.1)	4.4	175.5 (4.2)	***6.0***	174.5	6.5
3	Basic	149.7 (4.1)	67.1	153.2 (3.4)	58.6	149.2 (6.0)	68.2	146.5 (8.6)	73.5	152.2	66.3
	Proficient	163.2 (2.9)	29.2	167.1 (2.8)	19.6	164.3 (3.3)	26.6	163.2 (4.1)	29.7	165.2	26.6
	Advanced	174.5 (4.3)	7.0	178.0 (2.3)	3.4	176.7 (4.8)	4.4	177.4 (5.1)	4.0	175.7	5.5
Final	Basic	149.7 (0.0)	67.1	153.1 (2.0)	58.6	145.9 (3.8)	75.1	147.6 (4.0)	71.8	Only three rounds of ratings	
	Proficient	163.2 (0.0)	29.2	167.2 (1.1)	19.6	163.7 (0.7)	28.8	162.6 (1.3)	32.0		
	Advanced	174.5 (0.0)	7.0	177.9 (1.1)	3.7	175.0 (1.9)	6.5	177.2 (0.7)	4.4		

* Data printed in ***bold italics*** were not presented to panelists in the process

** The Geography ALS cutpoints were computed from ME ratings for the 65 items included in the Civics FT rating pool. Standard deviations of the cutpoints cannot be computed because it would necessitate the computation of individual cutpoints of some ALS panelists that are based only on ratings for one block of items.

ROUND 1 CUTSCORES AND THEIR STANDARD DEVIATIONS

Round 1 cutpoints for all levels were noticeably lower for group A than the other three groups. (Please see Table 11.) The cutscores for Proficient and Advanced were about the same for groups B, C, and D. The cutscores for Basic were about the same for groups B and C, and they were higher than the cutscores for Basic level for groups A and D. The standard deviations (SD) of the cutpoints for Basic were considerably higher for all four groups than the SD for Proficient and Advanced. This is typically the case. Perhaps the absence of a description for performance below the Basic level contributes to the higher level of variability in panelists' ratings at the Basic level. Copies of the feedback based on the first round of ratings have been included as Appendix J.

ROUND 2 CUTSCORES AND THEIR STANDARD DEVIATIONS

All of the cutpoints for Round 2 were the same or higher than the cutpoints for Round 1. This observation held true for all levels and all groups. The cutscores for group A no longer were considerably lower than those for the other groups, because all panelists in that group understood how to record their judgments. In general, the standard deviations of the cutscores decreased for round 2, although the SD for Basic remained the highest of the three levels, as in round 1.

ROUND 3 CUTSCORES AND THEIR STANDARD DEVIATIONS

Round 3 cutpoints for all levels were noticeably more similar across all groups than the cutpoints for Round 1 and Round 2. In particular, the cutscores for Advanced were very similar for all groups (range 174.5 – 178.0). The standard deviations (SD) of the cutpoints for Basic were higher for groups B, C and D than the SD of the cutpoints for Proficient and Advanced. The SD of the Basic cutscore and the Advanced cutscore were about the same for group A. Standard deviations for the Reckase method groups showed an unusual pattern. Standard deviations typically decrease from round to round, and that was generally the case for cutpoints for all groups when Rounds 1 and 2 are compared. For Round 3, however, the standard deviation of cutscores increased for panelists who marked their cutscores on the Reckase charts.

FINAL CUTSCORES

Final cutpoints for all levels and groups were very similar to Round 3 cutpoints. Group C made the largest change in the Basic cutpoint (Round 3 = 149.2 and Final 145.9). This was somewhat surprising since Group C had received consequences data throughout the process and since panelists had the opportunity to select their individual Round 3 cutscores directly on the Reckase Charts. The SD was zero for the cutpoints set by Group A for all levels because all panelists selected the group cutscores as their final recommendations. The SD for the Basic cutpoints set by Groups C and D were higher than the SD for cutscores set for Proficient and Advanced.

EVALUATION OF METHODS FOR ROUND 3 RATINGS

The panelists in all groups used an item-by-item rating procedure for the first two rounds. The rating method for groups A and B were exactly the same as those in previous ALS procedures for geography, U.S. history, and science. The ratings, per se, for groups C and D also required the same procedure as used before, although working with the Reckase Charts added considerably to the information of panelists and to their activities. Transferring ratings and cutscores to the charts took quite a long time. Prior to Round 3, panelists estimated the expected performance of students at the borderline of each achievement level for each item. Round 3 ratings were different. Panelists in groups A and B marked item maps, and panelists in groups C and D marked Reckase Charts.

ADJUSTING CUTSCORES USING ITEM MAPS

Panelists responded positively to the item maps. Recall that item maps were introduced as an easy and efficient way for panelists to adjust cutscores without going through the item-by-item rating

process a third or fourth time. The original plan for field trial 2 was to examine the interface of the two methods (Mean Estimation and Item Maps) to determine whether the two could be used in one ALS process without creating a sense of disruption in the process. Judges appeared to experience no real problem in switching to the item maps to mark their round 3 cutpoints. All 21 panelists using the item maps adjusted at least one of their cutscores for round 3 after studying the maps. Of the 63 opportunities for adjustment (21 judges and 3 achievement levels) 37 adjustments were made to raise the cutscores, 16 were made to lower the cutscores, and no changes were made in 10 cases. In 20 cases the adjustments were relatively large—5 score points or more. It would appear that judges were able to adjust their cutscores readily using the item maps without repeating the item rating procedures. None appeared to find this task difficult. The fact that all panelists chose to adjust at least one of their cutscores suggested that a rough fit or interface existed between the Mean Estimation method and the item maps that required further smoothing Additional analyses of panelists' reactions to the item maps have been summarized in the section on process evaluation questionnaires.

ADJUSTING CUTSCORES USING RECKASE CHARTS

Panelists also responded positively to the Reckase method. Recall that the Reckase method required judges to select a single scale score from the Reckase Chart to represent their ratings for each achievement level for round 3. They appeared to experience no problem when instructed to perform this task. All 22 panelists using the Reckase method adjusted at least two of their cutscores for round 3. Of the 66 opportunities for adjustment (22 panelists and 3 achievement levels) 31 adjustments were made to raise the cutscores, 27 were made to lower the cutscores, and no changes were made in 8 cases. In 9 cases the adjustments were 5 score points or more. Overall, panelists using the Reckase method made adjustments of smaller magnitudes by round 3 than panelists using the item maps. Further analyses of panelists' reactions to the Reckase method have been summarized in the section on process evaluation questionnaires.

EVALUATION OF PANELISTS' COMMENTS AND RESPONSES TO PROCESS EVALUATION QUESTIONNAIRES

To better understand panelists' perception of the rating process, they were asked to respond to process evaluation questionnaires using Likert-type scale items, for the most part. The questionnaires were administered throughout the meeting: four to the ME/IM rating groups and five to the ME/MR rating groups. The additional questionnaire was designed specifically to collect evaluations of instructions in and reactions to the use of Reckase Charts. The complete set of responses to the process evaluation questionnaires has been included in Appendix M.

Overall, the response of panelists to the ALS procedures was generally positive. Given that the essentials of a five-day process were condensed into an intense two-day process, it was reasonable to expect panelists to be only moderately positive about their ALS experience. However, panelists perceived the standard-setting procedures as a positive experience, regardless of their experimental group. No response patterns emerged to suggest significant negative or unusual reactions by panelists to any of the procedures implemented.

A consistent finding from past ALS research has been that panelists' responses generally tend to become more positive as the rounds of ratings progress. That is, judges usually indicate an increase in their understanding of the rating tasks and an increase in confidence in their ratings as they gain more experience with the ALS process. This general trend was not always consistent with the responses of the different groups over rounds in this field trial, however. Many panelists indicated *less* positive responses to questions related to rating method after round 3 than for earlier rounds. Recall that round 3 required judges to mark their cutscores on either the item maps or Reckase Charts. Further, groups B and D received consequences data for the first time after round 3.

In particular, group A was unusual in that some of their responses were least positive after round 3. For example, when group A was asked if the method for rating multiple-choice items was conceptually clear, the mean response (5=totally agree; 1=totally disagree) for round 1 was 4.10, for round 2 was 4.60, but dropped for round 3 to 3.50. A similar response pattern was produced when group A was asked if the method for rating multiple-choice items was easy to apply (round 1=4.0; round 2=4.5; round 3=3.5). Apparently, they experienced more difficulty with item maps than with the Mean Estimation method.

When examining panelists' responses to questions about the item mapping method, all responses were at least somewhat positive (3.0 or greater). For groups A and B who used the item maps, responses were consistently higher for group B than group A. Recall that group B received consequences data for the first time after round 3 and group A received consequences data after round 1. None of the panelists who used the item maps appeared to experience problems related to the maps. All of the panelists who used item maps expressed confidence in their ability to use them.

When examining panelists' responses to questions about the Reckase method, all responses were also at least somewhat positive (3.0 or greater). For groups C and D who used the Reckase method, responses were similar, regardless of when the groups received consequences data. Twenty of the 22 panelists expressed confidence in their selection of one row on the Reckase Charts to represent their ratings for each achievement level. Eighteen of the 22 panelists agreed that the cutscores they set using the Reckase Charts were closer to their concept of borderline performance than their previous ratings.

Panelists' responses to questions regarding rating methods and consequences data were quite varied by group. Few panelists responded negatively (scores 1-2) to these questions, and nearly all responded at least somewhat positively (scores 3-5). When asked if they would be willing to sign a statement recommending the use of achievement levels resulting from their ALS procedure, all but two of the forty panelists responded positively.

COGNITIVE COMPLEXITY OF THE PROCEDURES

The issue of cognitive complexity was included as an integral part of evaluating the experimental ALS methods. Responses to process evaluation questionnaires reflected the perceptions of panelists regarding the level of cognitive complexity required to perform the rating task, among other questions. By round 3, all but one of the 43 panelists indicated that they understood the

tasks they were to accomplish during the session. Eighteen responded that their level of understanding was "totally adequate." By round 3, all but two panelists indicated that the method for rating multiple-choice and constructed response items was conceptually clear. They did not appear to have difficulty performing any of the tasks involved in the process. Panelists reported that both their overall satisfaction with the rating methods and their confidence in the results produced by the methods were high. Responses indicated that they were comfortable with the level of cognitive complexity required of them.

EVALUATION OF PANELISTS' RESPONSES TO CONSEQUENCES DATA QUESTIONNAIRES

Rather surprisingly, no compelling differences in the cutscores resulted from the timing of providing consequences data. (Please see Table 11.) Groups A and C received consequences data after every round. Groups B and D were given consequences data for the first time after round 3. Receiving consequences data earlier or later in the process did not seem to have an effect on the final cutscores and percentages of students performing at or above each achievement level (see Table 12). The final percentages of students performing at the cutscores for the three achievement levels recommended by the early consequences groups (A & C) were more similar than those made by the late consequences groups (B & D).

Table 12
Percentages of Students Performing at or above Each Achievement Level Based on Cutscores for Field Trial 2 Civics ALS Process

	Early Consequences Data		Late Consequences Data	
	Group A	Group C	Group B	Group D
Basic				
Round 1	89.8	75.1	N/A	N/A
Round 2	73.5	70.0		
Round 3	67.1	68.2	58.6	73.5
Final	67.1	75.1	58.6	71.8
Proficient				
Round 1	46.3	29.7	N/A	N/A
Round 2	35.8	23.5		
Round 3	29.2	26.6	19.6	29.7
Final	29.2	28.8	19.6	32.0
Advanced				
Round 1	16.0	7.9	N/A	N/A
Round 2	10.9	4.4		
Round 3	7.0	4.4	3.4	4.0
Final	7.0	6.5	3.7	4.4

Group A (early consequences data with item maps), tended to make the greatest changes in cutscores and the subsequent percentages of students performing at or above each level between round 1 and round 2, with smaller changes between round 2 and round 3, and no changes between round 3 and round 4. Some group A panelists did not understand the rating process for round 1 and produced extremely low Basic cutpoints. The magnitude of change in ratings from round 1 to round 2 was largely a result of this misunderstanding.

Group C (early consequences with Reckase Charts), tended to make substantial changes between round 1 and round 2, with smaller changes between round 2 and round 3, and additional changes between round 3 and round 4.

Group B (late consequences with item maps), made fewer changes in the percentages of students performing at or above each achievement level for their final recommendations than group D (late consequences with Reckase Charts) made to their final recommendations. For a complete report of the reactions of different groups to the consequences data, please see Appendix M.

Panelists were asked to explain the impact, if any, of consequences data on their cutscores. Table 13 is a summary of the frequencies of their responses. About an equal number of comments were made about the positive impact of consequences data as were made about the limited impact of consequences data. Many comments indicated that consequences data caused judges to reevaluate and reconsider their cutscores. Most panelists commented that they found the consequences data helpful in their understanding of the overall ALS process. Only one judge indicated that the consequences data were highly influential when forming his/her judgments of student performance. Many panelists responded that they had used consequences data to adjust their cutscores, along with other sources of information.

Table 13
Frequencies of Responses Regarding the Impact of Consequences Data in Civics Field Trial 2
Summary of Response Types by Groups

	Group A	Group B	Group C	Group D	
Response	ME early CD	ME late CD	MR early CD	MR late CE	Total
Positive impact; Improved confidence	3	1	2	5	11
Moderate Impact	0	0	1	0	1
Caused me to reevaluated my cutscores	1	4	5	1	11
Caused me to reconsider my cutscores relative to cutscores of others	2	1	2	2	7
No impact; Limited impact	3	2	3	4	12
No Comment	2	4	0	0	6

Changes Made to Cutscores in Response to Consequences Data

The effect of consequences data appeared to be more a function of method than of timing. Table 14 displays the number of changes panelists made to their cutscores in response to consequences data. Panelists using the Reckase method tended to recommend more changes in response to the consequences data than panelists using the Mean Estimation method. That is, it seemed that the adjustments in ratings made by panelist in the Mean Estimation rating group were less numerous and persistent than those made by panelists in the Reckase group. These differences, however, were not statistically significant.

Table 14
Number of Changes Made to Cutscores in Response to the Consequences Data by Group for Civics Field Trial 2

Round	Level	Group A (ME)	Group B (ME)	Group C (MR)	Group D (MR)
1	Basic Proficient Advanced	9 3 1	No Consequences Data	3 4 1	No Consequences Data
2	Basic Proficient Advanced	3 1 0	Distributed During Rounds	3 4 8	Distributed During Rounds
3	Basic Proficient Advanced	0 0 0	3 2 2	5 6 7	6 3 2
4	Basic Proficient Advanced	0 0 0	0 0 0	0 0 0	3 1 0
	Totals	**17**	**7**	**41**	**15**

SUMMARY OF FIELD TRIAL RESEARCH

The field trials were designated as opportunities to explore new methods and procedures for collecting and summarizing judgments used in setting achievement levels for NAEP. Taken together, the field trials research provided important information about various elements that constitute the standard-setting process designed by ACT. Findings from the field trials research greatly informed the procedures that were developed for the pilot study.

CIVICS FIELD TRIAL 1

The purpose of the first civics field trial was to compare the Item Score String Estimation (ISSE) rating method with the Mean Estimation (ME) method. Results of the first field trial in civics indicated that panelists were able to use the method without difficulty. The ISSE cutpoints and their standard deviations appeared to be reasonable when compared with those produced by the Mean Estimation method. The panelists expressed satisfaction with and confidence in the ISSE method and the outcomes of the process. The ISSE procedures were implemented with ease. However, the ISSE method resulted in higher Proficient and Advanced cutscores and lower percentages of students performing at all three achievement levels than the modified-Angoff method. The ISSE method was found to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true score" or "true" judgment of the panelist (Reckase & Bay, 1999). Because of this flaw, further research using the ISSE method was discontinued, and it was eliminated as a possibility for implementation in the civics ALS.

CIVICS FIELD TRIAL 2

The fundamental purpose of the second field trial was to identify the method that would be used for the 1998 ALS process. ACT studied the effect of item maps and the timing of consequences data on the outcomes of the process. ACT also studied the new Reckase standard setting method.

Results of Field Trial 2 indicated that panelists had no difficulty using the new methodologies involving item maps and Reckase Charts. Informing panelists of the consequences of their ratings earlier in the standard setting process, rather than later, did not significantly affect the outcomes. Panelists using the Reckase method tended to change their item ratings to be more similar to the IRT-based performance estimates of students at the cutscores. In addition, panelists using the Reckase method usually modified extreme ratings to fall within a band of values near the cutscores generated by the rating groups or the individual panelist.

FINDINGS REGARDING ISSUES IN NAEP STANDARD SETTING

It is important to emphasize that the field trials addressed several unique and difficult technical challenges inherent to setting achievement levels for NAEP. These challenges have been an on-going concern for ACT in its effort to refine and improve the NAEP standard setting process.

THE ISSUE OF COGNITIVE COMPLEXITY

The charge has been made that the item-by-item rating method used in the NAEP ALS process 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 & Plake, 1998). ACT has collected considerable data during the civics field trials and previous ALS research where panelists have reported their capacity to perform the tasks associated with estimating student performance. Judges perceive that they are performing the estimation and judgmental tasks required by the method with relative ease. They report that they are confident in their judgments and satisfied with the results. There is no evidence to indicate that panelists are unable to make complex judgements.

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

One of the considerations for evaluating the various experimental methods explored by the field trials was based in part on indicators of "reasonable" 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. Judges were expected to adjust their ratings from round to round. 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, this was an indicator 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, this was also an indicator that they probably did not understand the rating method or the feedback. The civics field trial panelists exhibited "reasonable" intrajudge consistency across rounds based on these indicators.

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

ACT has been searching for an effective format to provide panelists with information about the relationship of their ratings to the performance of students on items. Although this information

had been given to panelists in previous ALS meetings, there was no evidence to suggest that panelists either understood the information, or found it useful when forming their judgments about student performance. The Reckase method greatly advanced the effort to give panelists precise item-level information they could readily understand about ratings, relative to expected student performance. Panelists who used the Reckase method, in general adjusted their ratings for round 2 to be more similar to the IRT-based performance estimates of students at the cutscores. This finding was consistent for all three achievement levels. 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. This did not happen, which 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 performance estimates of student performance. Responses to the process evaluation questionnaires support this assumption.

After adjusting their ratings for round 2, the Reckase method required panelists to select a single row or scale score to represent their ratings for each achievement level for round 3. When asked about their confidence in the selection of one row to represent their ratings, 20 of the 22 panelists responded positively indicating that they were at least somewhat confident in the selection.

Panelists were asked to respond to the hypothetical situation of selecting a range of rows instead of a single row on the Reckase Charts to represent their ratings. Panelists did not prefer selecting a range of rows, nor did they prefer averaging the scores of the rows instead of selecting a single row.

THE ISSUE OF DIFFERENCES BETWEEN CUTSCORES FOR POLYTOMOUS AND DICHOTOMOUS 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 field trial set cutscores that were statistically significantly higher for polytomous items than dichotomous items. Differences in ratings of multiple choice and constructed response items were one of many considerations brought to the attention of panelists using the Reckase method. After studying the feedback, including the charts, panelists adjusted their ratings for round 2. The differences between multiple choice and constructed response ratings were reduced in 24 of the 66 opportunities for adjustments (22 judges adjusting 3 achievement levels). Differences increased between multiple choice and constructed response ratings in 19 of the 66 opportunities for adjustments. Differences remained unchanged in 23 opportunities for adjustments. This finding suggests that panelists did not lower their ratings for polytomous items. Apparently they judged the performance of students to be inconsistent with the ALDs. Clearly this is not an oversight by the panelists indicating a flawed process. Although more research is needed to determine how judges perceive polytomous items differently than dichotomous items, the Reckase Charts appear to have been effective in making panelists aware of these differences when forming their judgments of student performance.

THE ISSUE OF TIMING CONSEQUENCES DATA

The impact of consequences data on outcomes has been a topic of considerable interest in setting standards. No compelling differences were found in cutscores produced by judges who received consequences data early in the process compared with late in the process. The impact of receiving consequences data on the final cutpoints seemed moderate. About an equal number of comments were made by panelists about the positive impact of consequences data as were made about the limited impact. Judges, in general, found the consequences data informative and useful, but cutpoints were not greatly influenced by the data. Most panelists indicated that consequences data were only one of many factors they considered when forming their judgments about student performance.

PLANNING FOR THE CIVICS PILOT STUDY

The Reckase charts, introduced in the second field trial for civics, were judged to be a promising addition to the ALS process designed for NAEP. The charts appeared to have improved and strengthened the ALS process. After reviewing the results of both the civics and writing field trials, it was agreed that research would continue on the Reckase charts. TACSS recommended using a modified version of the Reckase method for further research in the civics pilot study. The Reckase method, per se, was eliminated but the Reckase charts would be used in the pilot studies, with the expectation that they would also be used for the ALS. Panelists in future studies would omit selecting a single row on the chart to represent student performance at the cutscore for each achievement level. The Reckase charts would be presented to panelists as a step in preparation for the rating process. This modified role of the Reckase charts would be implemented in the ALS for civics, unless findings in the pilot study suggested otherwise.

There was little evidence that the consequences data impacted judgments of panelists enough to effect the cutscores they set. ACT and TACSS have consistently recommended that consequences data be included as feedback in the ALS process. The findings from field trial research indicated that outcomes of the process would not be significantly different if consequences data were provided as feedback. Recognizing that NAGB wishes to have the NAEP achievement levels set via a criteria-referenced methodology, however, TACSS recommended that for the pilot study, consequences data not be presented to panelists until after three rounds of ratings. TACSS recommended implementation of a design that would produce two sets of cutscores: one set based solely on ratings, and one modified as a result of consequences data. The latter were to be the cutscores used for selecting exemplar items and, presumably, to be recommended to NAGB.

REFERENCES

ACT (1995). *Research studies on the achievement levels set for the 1994 NAEP in Geography and U.S. History*. (Unpublished).

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

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, DC: American Council on Education.

Bay, L. (1998). *An investigation of the bias of cutpoints resulting from item score string estimation (ISSE) ratings.* A report prepared for the meeting of the 1998 NAEP Achievement Levels-Setting Project Technical Advisory Committee for Standard Setting (TACSS).

Berk, R.A. (1994). Standard setting—the next generation. In M.L. Bourque (Ed.), *Proceedings of joint conference on standard setting for large-scale assessments, Vol. II*. Washington, DC: National Assessment Governing Board.

Carlson, J. (1995). *Estimation of reliability of NAEP IRT proficiency score estimates.* Technical Memorandum, ETS.

Chen, W. (1998). *Setting achievement level standards for NAEP using item score judgment: a simulation study*. A paper prepared for presentation at the annual meeting of the National Council on Measurement in Education, San Diego.

Hambleton, R.K. & Plake, B.S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education*, 8, 41-56.

Hambleton, R.K., Brennan, R.L., Brown, W.J., Dodd, B., Forsyth, R.A., Mehrens, W.A., Nellhaus, J., Reckase, M.D., Rindone, D., van der Linden, W.J., & Zwick, R. (2000). A response to "Setting reasonable standards" in the National Academy of Sciences' Grading the Nation's Report Card. *Educational Measurement: Issues and Practice*, 19(2), 5-14.

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. & Bourque, M.L. (in press). Tradition to innovation: Standard-setting on the National Assessment of Educational Progress. In Cizek, G.J. (Ed.), *Setting performance standards: Concepts, methods and perspectives*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Loomis, S.C., Bay, M.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 for civics and writing.* Paper presented at the annual meeting of the National Council on Measurement in Education. Montreal.

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.

National Research Council (1999). *Grading the Nation's Report Card: Evaluating NAEP and Transforming the Assessment of Educational Progress*, James W. Pellegrino, Lee R. Jones, and Karen J. Mitchell, eds. Committee on the Evaluation of National and State Assessments of Educational Progress, Board on Testing and Assessment. Washington, DC: National Academy Press.

Plake, B.S. (1995). An integration and reprise: What we think we have learned. *Applied Measurement in Education, 8,* 85-92.

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

Reckase, M.D. (1998b, April). *Methods for collecting test-based judgments.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

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.

Reckase, M.D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT.* Iowa City, IA: ACT.

Sanathanan, L. & Blumenthal, S. (1978). The logistic model and estimation of latent structure. *Journal of the American Statistical Association*, 73, 794-799.

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.