Developing Achievement Levels for the 1998 NAEP in Civics: Final Report

Susan Cooper Loomis and Patricia L. Hanick
ACT, Inc.

December 2000

Developing Achievement Levels for the 1998 NAEP in Civics: Final Report

Susan Cooper Loomis and Patricia L. Hanick
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 OF THE REPORT

Achievement levels are an important part of the National Assessment of Educational Progress (NAEP). There are three components to the NAEP achievement levels. Achievement levels descriptions state what students should know and be able to do; cutscores identify the performance levels on the NAEP score scale and serve as the bases for reports of the proportion of students who score at or above each; and exemplar items show what students who score within each achievement level category can do.

This report describes the process for setting the achievement levels. A summary of the process used to develop the final versions of the achievement levels descriptions is included, as well as a detailed account of the operational achievement levels-setting study that produced the numerical cutscores and exemplar items recommended to NAGB for adoption as the 1998 NAEP Civics achievement levels. The report also describes the Web-based process used by ACT to collect public opinion and comments evaluating the reasonableness and usefulness of the Civics NAEP achievement levels that resulted from the achievement levels-setting process.

Procedural validity is a necessary—but not sufficient—condition for a valid achievement levels-setting (ALS) process. This report documents the process used to set the Civics NAEP achievement levels and provides evidence for the procedural validity of the civics ALS process. In addition, it provides an overview of two field trials and one pilot study conducted to determine and refine the design for the 1998 civics ALS process. A brief description is provided of the procedure by which the achievement level descriptions were modified and finalized prior to implementing the pilot study. These studies are reported elsewhere, and brief summaries are presented here.

THE PANELISTS

The ALS panelists were nominated and selected through a carefully planned design that incorporates principles of sampling[1]. NAEP achievement levels are set by panels of broadly representative persons who are well-qualified with respect to knowledge in the subject area and knowledge of students at the 4th, 8th, or 12th grade. Panels are composed of teachers (55%), other educators (15%), and representatives of the general public (30%). School districts serve as the basic sampling unit for the panelist nomination and selection process, although they are supplemented for identification of nominators in postsecondary institutions and for nominators in other specific positions. A total of 843 nominators were contacted to participate in the process of identifying Civics ALS panelists, and 426 candidates were nominated. The goal was to select 90 panelist from the pool of nominees such that there would be 30 panel members for each of the three grade levels. A total of 87 panelists participated in the ALS, and they represented 22 states and Guam.

[1] The entire process is described in detail in the *Design Document* for the 1998 NAEP ALS Process (ACT, 1997b).

THE PROCESS

The ALS process lasted five days. The NAEP ALS Project Director served as the primary process facilitator and led all general sessions. A process facilitator directed the activities within each grade group panel. Panelists in each grade group worked with a content facilitator who had been a member of the framework development committee. Great care and attention were directed to training facilitators in the process and to assuring that the process was duplicated in each grade group.

Training and preparation for the achievement levels-setting task were scheduled throughout the process. The item-by-item rating method used to collect judgments for setting the cutscores was implemented after more than three full days of training and practice. Three rounds of item-by-item ratings were collected, and feedback was provided following each round to help prepare panelists for the following rounds of ratings.

Panelists completed seven extensive evaluations of the process. The process evaluation questionnaires were administered throughout the process—generally at the end of each day, with some additional evaluations following significant steps in the process. Panelists were very positive about the process and the outcomes of the process.

OUTCOMES OF THE PROCESS

Both during and after the ALS process, extensive analyses are conducted of feedback and other data to ascertain how the process functioned and to understand what led to the outcomes of the process. The procedures designed and implemented by ACT to set 1998 Civics NAEP achievement levels proved to be a highly effective process, as determined by the evaluation criteria.

The following ALS process outcomes were evaluated, and both the evaluations and results are presented in the report:

- Cutpoints and their standard deviations
  - Comparison of Cutpoints by Item Type
- Intrajudge consistency
  - Consistency across rounds (changes in ratings from round to round)
  - Reckase Chart analyses
- Consequences questionnaire data
  - Individual consequences data
  - Grade-level consequences data
- Process evaluation questionnaire data
- Selection of exemplar items

The ALS process was evaluated on the basis of whether the following outcomes were achieved:

- reasonable cutscores;
- relatively low standard deviations of those cutscores;
- reasonable levels of judges' item rating consistency, between and within rounds;
- high level of positive responses to the process evaluation questionnaires;
- adequate number of exemplar items selected;
- patterns of results consistent with previous studies; and
- absence of extreme reactions to consequences data.

The process implemented in the Civics ALS was designed to be compatible with the psychometric attributes of NAEP, to meet NAGB's policies and guidelines for setting achievement levels, and to be consistent with best procedures and practices for standard setting known to ACT and TACSS. The result was a highly effective and successful standard setting process. Panelists were able to carry out the process without observed or self-reported difficulty, and their reaction to the procedures was very positive.

The ACT/NAGB NAEP ALS Process includes a design feature that provides a measure of the reliability of the process. Each grade level panel is randomly divided into two rating groups that are as equally matched as possible. Similarly, each grade level item pool is randomly divided into two rating pools such that the items rated by each rating group are as equally matched as possible.

- ✓ *The cutscores of the two rating groups were found to be very similar, indicating that other similar panels would produce statistically similar results.*

While few modifications were made to the ALS process implemented for the 1998 NAEP, the modifications represented important changes to the ALS process. The most significant changes were:

- finalizing the achievement levels descriptions before the ALS panels were convened;
- introducing consequences data during the rating process, rather than after the cutscores were set; and
- providing the Reckase Charts as a means of informing panelists about item-level student performance and intrarater consistency.

FINALIZING THE ALDS BEFORE CONVENING THE ALS PANELS

Developing the achievement level descriptions has been an important part of the standard setting process for NAEP. A strong logical connection links NAGB's policy definitions of achievement in general to the operational definitions of achievement in civics. These operational definitions of achievement are the basis of training panelists, and they guide the item rating process. Useful and reasonable outcomes of the ALS process depend upon useful and reasonable achievement levels descriptions.

Prior to convening the 1998 ALS panels, the achievement level descriptions had been carefully crafted and thoroughly reviewed in a well-documented process. The revised achievement levels descriptions were compared not only to the Civics Framework and to the policy definitions, but they were also compared to the item pools for each grade level. The procedure for evaluating and modifying the ALDs prior to the operational ALS studies was judged to be a considerable improvement over previous practices.

While the plan to finalize the ALDs was generally judged to be a positive change in the process, there was concern that panelists would be less committed to the ALDs and to the standards they set since they had no role in writing the descriptions of what students should know and be able to do.

- ✓ *Although they had no role in writing or modifying the statements of what students should know and be able to do, panelists evaluated their understanding of ALDs as positively as had been the case in previous procedures.*

The typical response pattern that has emerged from past ALS meetings was present for the Civics ALS: Achievement levels descriptions were generally better understood than the borderline descriptions. Panelists' understanding of both categories of performance increased over rounds so that the difference between the two diminished by Round 3. A comparison of evaluations by the civics panelists who wrote descriptions of borderline performance and geography panelists who only discussed the concept indicated that there was little difference between the two.

- ✓ *Writing borderline descriptions does not markedly improve the clarity of panelists' concepts of borderline performance.*

PROVIDING CONSEQUENCES DATA DURING THE PROCESS

Determining when and how much information to provide to panelists has been a continuing concern for the design of the ALS process. Of considerable debate has been the provision of consequences data to judges. The goal has been to provide the best balance of information to panelists so that their judgments will be both realistic and based on the ALDs. For the 1998 ALS study, NAGB agreed to allow panelists to review consequences data during the process of setting cutscores. Accordingly, panelists first reviewed consequences data after their second round of item-by-item ratings. They were provided consequences data again after the third round of ratings, and they made recommendations for final cutscores based on their evaluation of those data. Interestingly, the consequences data were regarded by most panelists as just one among many sources of information for their consideration.

- ✓ *The concern that consequences data would dominate panelists' judgments was unfounded.*

- ✓ *Informing panelists of the consequences of the cutscores they set increased confidence in the credibility of the outcomes of the process.*

- ✓ *The finding that panelists' responses to the consequences did not lead to significant modifications in cutscores increased confidence in the process, in general.*

USING THE RECKASE CHARTS AS FEEDBACK

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

- ✓ *All three grade panels for the Civics ALS ranked the Reckase Charts as the most helpful feedback given to them.*

A concern associated with incorporating the Reckase Charts into the ALS process was that panelists would rely on the chart data to the exclusion of other sources of relevant feedback, possibly deferring their judgment to the statistical data shown on the chart. There was no evidence of undue influence based on observations of panelists working with the charts, panelists'

responses to questionnaire items, and extensive follow-up analyses of individuals' Reckase Charts.

- ✓ *Although panelists were greatly impressed by the usefulness of the charts and the ease of using them, they indicated that they considered other forms of feedback as well when forming their judgments*.

- ✓ *The Reckase Charts did not overly influence panelists when modifying their ratings, to the exclusion of other types of feedback.*

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

One persistent challenge to improving the ALS process has been to find a way to provide panelists with information about the relationship between their individual item ratings and student performance. *After panelists studied the Reckase Charts, they generally adjusted their ratings to be more similar to the IRT-based performance estimates of students at the cutscores—either their own cutscores or the grade-level cutscores.*

- ✓ *Intrajudge consistency increased across all three grades.*

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

- ✓ *Although the Reckase Charts were rated as the most useful feedback, findings show that panelists did not rely solely on the Charts and disregard other feedback data.*

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

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

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

THE ISSUE OF DIFFERENCES BETWEEN MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEMS

The difference between the cutscores that would result from ratings of polytomous and dichotomous assessment items has been another persistent challenge to ACT's effort to refine the standard setting process for NAEP. As has been the case in past ALS meetings, panelists for the Civics ALS 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 when reviewing Reckase Charts. After studying the feedback, including the Reckase Charts, panelists adjusted their ratings for the subsequent round (Round 2 and Round 3). In general, the differences between multiple choice and constructed response ratings were reduced for all grades and all levels for the subsequent rounds. The cutscores set for polytomous items were still higher than those set for dichotomous items. As has been the case for previous investigations in other NAEP ALS studies, differences in *cutscores* computed for items of the two types are greater than differences in *ratings* would suggest (ACT, 1997a).

- ✓ *Although more research is needed to determine how judges perceive polytomous items relative to dichotomous items, the Reckase Charts appear to have been effective in helping to make panelists aware of differences and to modify their judgments of student performance.*

- ✓ *The rating data and the process evaluation response data, indicate that panelists' ratings were based on the achievement levels descriptions.*

- ✓ *Student performance simply fell short of panelists' informed judgments of what was reasonably expected of the students who met the standard of the NAEP achievement levels.*

THE ISSUE OF PROVIDING CONSEQUENCES DATA DURING THE RATING PROCESS

The impact of consequences data on outcomes has been a topic of considerable interest to NAEP standard setting. No compelling differences were found in cutscores produced by the Civics ALS judges who received consequences data for the first time after Round 2 and judges from other ALS studies who received consequences data after ratings were completed and cutscores set.

- ✓ *Judges, in general, found the consequences data informative and useful, but their item ratings and cutscores did not appear to be greatly influenced by the data.*

- ✓ *When given the opportunity to change their own cutscores after learning of the consequences, few panelists chose to make changes. Those who did tended to adjust their cutscores by only a few points.*

THE ISSUE OF COGNITIVE COMPLEXITY

ACT has collected considerable data during the civics ALS studies and previous research where panelists have reported their capacity to perform the tasks associated with estimating student performance.

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

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

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

ACHIEVEMENT LEVELS SET BY NAGB

ACT's decision to recommend the cutscores, achievement levels descriptions, and exemplar items to NAGB was based on a large amount of information collected from several sources.

- ACT ALS Project Staff have extensive experience with the NAEP ALS Process. Their observations and first-hand involvement with implementing the process and analyzing the results supported the conclusion to recommend the results to NAGB.
- Panelists had been very positive about the process and the outcomes of the process, and they had recommended adoption.
- The public opinion survey indicated that the achievement levels were useful and reasonable, and this outcome supported the conclusion to recommend adoption.
- The Technical Advisory Committee on Standard Setting, a very distinguished and highly experienced team, had carefully analyzed and evaluated the data and recommended adoption.

During their regularly scheduled meeting in May 1999, NAGB approved the 1998 Civics NAEP Achievement Levels, as recommended.

CONCLUSION

This comprehensive evaluation of the outcomes of the process revealed remarkable consistency, agreement, and overall satisfaction at every stage. Given that the NAEP standard setting process is based on judgments by broadly representative panels of individuals, such consistency is an impressive accomplishment.

Developing Achievement Levels for the 1998 NAEP in Civics: Final Report[2]

Susan Cooper Loomis and Patricia L. Hanick
ACT, Inc.

INTRODUCTION

Achievement levels are an important and integral part of the National Assessment of Educational Progress (NAEP). Analyses of how students perform on NAEP relative to statements of what they "should know and be able to do" represent the primary means of reporting NAEP results. NAEP Achievement Levels communicate this information about student performance to a variety of constituencies in an effort to improve education in the United States. Achievement levels provide an answer to the question: "How good is good enough?"

There are actually three components of NAEP Achievement Levels: achievement levels descriptions, cutscores, and exemplar items. Achievement levels setting (ALS) in NAEP refers to the overall process through which these three components are produced. The term also refers to the process through which achievement levels descriptions are translated onto the NAEP reporting scale as cutscores for each level.

The National Assessment Governing Board (NAGB) established policy definitions that provide the general descriptions of the three NAEP achievement levels: Basic, Proficient, and Advanced. The policy definitions are used to formulate operational definitions based on the assessment framework to describe what students should know and be able to do in that subject area at each grade assessed in NAEP (4^{th}, 8^{th}, and 12^{th}) and at each level of achievement. These operational definitions are called *achievement levels descriptions* (ALDs). *Cutscores* are numerical representations of student performance at each achievement level. Cutscores represent the lower boundary of performance for each achievement level—specifically, the minimal score on NAEP that represents performance at each level of achievement. Student performance relative to achievement levels is reported to educators, policymakers, parents, and the general American public. In order to make these reports more meaningful, NAEP items are selected to illustrate the kinds of tasks, the knowledge, and the skills required for performance at each level. *Exemplar items*, the third component of NAEP Achievement Levels, provide concrete examples of student performance at the Basic level, the Proficient level, and the Advanced level.

All U.S. school districts that receive Title I funding are required to report student achievement using some form of standards. Most states have set performance standards against which to judge the educational achievement of their students. This increase in the importance of reporting results in terms of performance standards is accompanied by an increase in attention to the process of setting cutscores. Questions and concerns have emerged regarding psychometric and standard-setting issues related to the NAEP achievement levels. ACT has attempted to address these issues openly and frankly through extensive research conducted prior to the 1998 Civics NAEP

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

achievement-levels setting (ALS) study. ACT has incorporated improvements and enhancements to the ALS process not only from this research, but also from experiences gained since 1992 during six NAEP achievement levels-setting efforts (Hambleton, et al., 2000). As a result of these efforts, the NAEP ALS process is regarded by many as the model to follow in standard setting for large-scale assessments. The 1998 NAEP ALS process includes a comprehensive training component, multiple rounds of ratings that produce cutscore estimates, and extensive feedback that is both item specific and holistic. Panelists' experiences with and reactions to the process are well documented through a comprehensive series of questionnaires administered throughout the process.

This report provides an overview of the various stages of research leading up to the operational ALS study. It describes the process used to develop the final versions of the achievement levels descriptions and provides a detailed account of the operational achievement levels-setting study that produced the numerical cutscores and exemplar items recommended to NAGB for adoption as the 1998 NAEP Civics achievement levels. The report also describes the process used by ACT to collect public opinion and comments evaluating the reasonableness and usefulness of the Civics NAEP achievement levels that resulted from the achievement levels-setting process.

RESEARCH CONDUCTED PRIOR TO THE CIVICS ALS

ACT carried out two field trials and one pilot study each for the 1998 Writing and Civics NAEP. All six of those studies were completed and reviewed by ACT's technical advisory committees prior to convening the panels for the 1998 Civics ALS meetings.[3] Taken together, the field trials and pilot study research provided important information about various elements that constitute the standard-setting process designed by ACT (Loomis & Hanick, 2000a; Loomis, Hanick, Bay & Crouse, 2000a and 2000b).

FIELD TRIAL #1

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

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

Results of the first field trial in civics indicated that panelists were able to use the ISSE method without difficulty. The panelists expressed satisfaction with and confidence in the ISSE method and the outcomes of the process. The procedures were implemented with ease. The ISSE cutpoints and their standard deviations appeared to be reasonable when compared with those produced by ratings of the same items with the Mean Estimation method.[4] The ISSE method resulted in higher cutscores for the Proficient and Advanced levels than those from the Mean Estimation method. The cutscores for the Basic level were approximately the same for the two methods. The ISSE cutscores resulted in lower percentages of students performing at or above all three achievement levels, particularly at the Proficient and Advanced levels. Further research showed the ISSE method to be biased in such a way that cutscores would be higher for the Advanced level and lower for the Basic level when compared with the "true" scores or "true" judgments of the panelist (Reckase & Bay, 1999). Because of this flaw, further research using the ISSE method was discontinued, and it was eliminated as an alternative for implementation in the Civics ALS.[5]

FIELD TRIAL #2

The fundamental purpose of the second field trial was to identify the procedures that would be used for the 1998 ALS process. ACT's goal was to complete the research phase prior to the pilot study, and the second field trial was the final opportunity to conduct research with panelists before the pilot study. ACT planned to study whether or not item maps could be incorporated into the rating process as a means for panelists to adjust their cutscores without a third round of item-by-item ratings. In addition, ACT wanted to study the effect of providing consequences data to panelists during the rating process. ACT studied both the effect of item maps and the timing of providing consequences data on the outcomes of the ALS process.

The ISSE method had been eliminated from further consideration, and no final decision had been made regarding the rating method to use in the 1998 ALS process. In field trial 2, ACT implemented the new Reckase method (Reckase, 1998) as an alternative to the Mean Estimation method for setting cutscores.

Results of the second field trial indicated that panelists had little difficulty with either the item maps or the Reckase method. There was no statistically significant difference between cutscores set by groups of panelists who were informed of the consequences of their ratings after the first round of ratings and cutscores set by panelists who were not informed until after the last round of item ratings.

PILOT STUDY

After reviewing the results of the field trials, it was agreed that research would continue on the use of Reckase Charts for setting NAEP achievement levels. The Reckase Charts were introduced

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

[5] For more detailed information about the civics field trials, please refer to Loomis, Hanick, Bay & Crouse, 2000a.

in the second field trial and judged to be a promising addition to the ALS process designed by ACT. They appeared to have added substantially to panelists' understanding of the process without a significant increase in the cognitive demand. Rather than follow the design of the Reckase method as originally described (Reckase, 1998), the charts were used with the Mean Estimation method throughout the rounds of ratings. The charts were used in the civics pilot study, with the expectation that they would also be used for the civics ALS.

The pilot studies for the 1998 ALS process had been planned as a "dry run" for the operational ALS to determine whether modifications to training, instructing, timing, and so forth were needed. The civics pilot study was an opportunity to continue studying and refining the incorporation of Reckase Charts into the NAEP ALS process. ACT was particularly interested in evaluating panelists' reactions to the ALS process with the Reckase Charts included as feedback since this modification had not yet been implemented. The pilot study was a final check on procedures to assure a successful operational Civics NAEP ALS.

Throughout the pilot study, ACT collected information about the reactions of panelists to the ALS process and the Reckase Charts. Their suggestions lead to adjustments in the process to assure smooth implementation of the methodology when used for the operational ALS meeting. Pilot study panelists strongly recommended that ACT develop a method to transfer electronically each panelist's ratings on to his/her Reckase Charts. The method would have to be very fast and efficient because each of the 90 ALS panelists rated 60-100 items at each of the three achievement levels, and each rating had to be marked on the charts. Panelists also recommended increasing amount of feedback during the training exercises.

Findings from the pilot study were unusual in that cutscores for all grades and all levels *increased* from round to round, while the standard deviations of the cutpoints *decreased* for each round of ratings. It has been common for the standard deviation to decrease from round to round, but the uniform pattern of increasing cutpoints has never been observed in data generated from previous NAEP studies. As was the case in the second field trial for civics, the Reckase Charts seemed to help panelists adjust their item ratings to be more consistent with IRT-based estimates of student performance with respect to the panelist's own cutscore and the grade-level cutscore. Panelists typically modified their extreme ratings to fall within a band of values nearer the grade-level cutscore or nearer their own cutscore.[6]

DEVELOPING FINAL VERSIONS OF THE ACHIEVEMENT LEVEL DESCRIPTIONS[7]

Preliminary achievement levels descriptions were developed as part of the 1998 Civics NAEP Framework (NAGB, 1998). The preliminary achievement levels descriptions were reviewed extensively, and they were finalized prior to convening the ALS panels for civics. The process of transforming the preliminary ALDs into the final ALDs involved three steps:

- Convening focus groups to review the preliminary ALDs and make recommendations to improve them;

[6] For more detailed information about the civics pilot study, see Loomis & Hanick, 2000a.

[7] For a complete account of the process used to finalize the 1998 NAEP ALDs, please refer to Loomis & Hanick, 2000b.

- Convening an Expert Review Panel to consider the recommendations from the focus groups and to modify the descriptions appropriately;
- Collecting informed opinions regarding the revised ALDs.

FOCUS GROUPS

The focus groups involved a broad segment of the population to study and evaluate the preliminary achievement levels descriptions. One focus group was convened in each of four geographic regions. Participants were selected to represent the three categories of persons who serve on the NAEP achievement levels-setting panels; teachers, nonteacher educators, and general public. The purpose of the review was to determine whether the descriptions of achievement in civics appeared to be both useful and reasonable statements of what students should know and be able to do. The judgement of "reasonableness" was with respect to the NAGB policy definitions of achievement and the Civics NAEP framework.

The general recommendations from the civics focus groups were as follows:

- Basic was too high for all grades. The content described in Basic was too difficult. Basic denoted more than partial mastery.
- Advanced was truly advanced for 4^{th} and 8^{th} grades. The description of Advanced at grade 12 should include more problem-solving and evaluating skills.
- The descriptions should show a clearer progression of development in civics skills across levels and grades.
- The descriptions should be stated in simpler, clearer language. The language of the preliminary ALDs was "inflated," and most statements were too complex.
- The preliminary ALDs for 4^{th} grade were too abstract to be developmentally appropriate.
- The level of detail reflected in the preliminary descriptions was inconsistent. In particular, the descriptions for 12^{th} grade were very broad, whereas for 8^{th} grade they were very specific. There seemed to be a gap between the performance described for 8^{th} grade and 12^{th} grade.

EXPERT REVIEW PANEL

Civics content experts reviewed the work of the focus groups and modified the preliminary ALDs according to the recommendations and their own expertise regarding the Civics NAEP. The preliminary ALDs for civics went through a considerable transformation during this revision process. Descriptors were shifted, deleted, written and rewritten. Although the preliminary civics ALDs were changed substantially, the content of the revised descriptions remained true to the Civics Framework. All of the revised ALDs were traced to the content areas outlined in the framework. The level of difficulty represented by the ALDs was lowered during the finalizing process, as recommended by the focus groups.

The preliminary ALDs were in bulleted statements. The finalized ALDs were paragraphs of complete sentences. The change in format tended to make the narrative descriptions less formidable and more attainable by students than the bulleted statements. Also, a serious attempt was made to delete professional jargon in favor of words more commonly used in daily language. Although some jargon terms still existed in the revised descriptions, those words were debated and carefully chosen by the Expert Review Panel members.

In the preliminary ALDs many different verbs had been used to describe the type of cognitive demand required across levels. While verbs like *identify* and *describe* have typically been used in ALDs for Basic level performance in other content areas, these verbs were used at all levels of performance in civics. Similarly, *evaluate* and *explain* have typically been used in descriptions of Proficient and Advanced level performances, but in civics these verbs were also used at the Basic level. This mix of verbs describing cognitive demand was changed during the finalizing process. The revised ALDs tended to cluster the verbs describing more demanding cognitive skills at the higher performance levels.

COLLECTING INFORMED OPINIONS REGARDING THE RECOMMENDED ALDS

Once the preliminary ALDs had been revised, ACT requested comments on the recommended version of the ALDs from the original focus group members and key people who had been involved in the development of the Civics NAEP. ACT conducted a telephone survey of focus group members. Results of the survey indicated that the revised ALDs were generally well received. Almost all respondents (94%, n=37) judged the ALDs to be reasonable. Two respondents indicated that the Basic descriptions did not reflect "partial mastery" and one indicated that the Advanced description did not reflect "superior performance."

Members of the NAEP Civics Standing Committee, Item Development Committee, and Framework Committee were also asked to comment on the revised ALDs. Of the 26 members who responded to the request, 19 recommended no changes to the revised ALDs and 7 recommended changes. The Expert Review Panel evaluated those recommended changes and determined that none suggested significant, substantive changes in the recommended civics ALDs.

ADOPTING THE REVISED ALDS

The final version of the civics ALDs that emerged from this process was approved by NAGB on August 8, 1998 for use in the pilot study and ALS process. The ALDs were officially adopted by NAGB, as part of the 1998 Civics NAEP Achievement Levels, in May 1999. Please see Loomis & Hanick (2000b) for additional details.

ALS PANELIST SELECTION PROCESS

The following summary highlights the main features of each step in the process of selecting panelists to set achievement levels. Please see Appendix B for additional details.

SELECTION OF SCHOOL DISTRICTS

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

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

The samples of districts were drawn to include at least 15% with enrollments of 25,000 or more students, and 15% with at least 25% of the population below the poverty level. A total of 289 public districts and 58 private schools were sampled. Please see Table 1 for the distribution of districts sampled. In addition, 15 colleges and universities were sampled from the Higher Education Directory (Rodenhouse & Torregrosa, 1998). Persons in specific positions were identified as nominators in those two- and four-year institutions, both public and private. The total number of districts selected and the proportion in each nominator type were based on previous experience with response rates from nominators in other subjects. Details of the process and the projected number of nominators in each category are provided in the *Design Document* (ACT, 1997b).

Table 1
Distribution of Districts Sampled

Nominator Type	Public Districts	Private Districts	Total
Teacher	159	52	211
Nonteacher Educator	18	6	24
General Public	112	-	112
Total	289 (83%)	58 (17%)	347

NOMINATORS

ALS nominators were identified by drawing three separate samples of districts without replacement.[8] One sample of public school districts was drawn from which nominators of teacher panelists were identified, a second for nominators of nonteacher educators, and a third sample for nominators of general public representatives. Nominators of private school teachers were identified from a sample of private schools drawn separately. A total of 843 nominators were contacted. Please see Table 2 for the distribution of nominators. Nominators were persons holding a specific title or position, such as the following.

Table 2
Distribution of Nominators Contacted

Nominator Type	Public Districts	Private Districts	State	College/ Universities	Employers	Total
Teacher	268	48	43	-	-	359
Nonteacher	17	6	14	25	-	62
General Public	285	-	-	-	137	422
Total	570 (68%)	54 (6%)	57 (7%)	25 (3%)	137 (16%)	843

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

Nominators of teachers were:

- district superintendents
- leaders of teacher organizations
- state curriculum directors
- principals or heads of private schools

Nominators of nonteacher educators were:

- non-classroom educators (e.g., principals, district social studies curriculum coordinators)
- state assessment directors
- deans of colleges and universities (two-year and four-year; public and private)

Nominators of members of the general public were:

- education committee chairpersons of the local Chambers of Commerce
- mayors
- school board presidents
- employers of persons in a civics-related position or with a civics-related background

POOL OF NOMINEES

Nominees represented a specific grade perspective (4^{th}, 8^{th}, or 12^{th}) and filled a specific role (teacher, nonteacher educator, or member of the general public). Nominators could submit up to four candidates for each grade whom they judged to be "outstanding" in their civics-related field. A total of 807 persons were identified as nominators, and they nominated a total of 426 candidates. Please see Appendix B for the distribution of the nominee pool.

CHOOSING PANELISTS

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

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

Ninety panelists were required for the panels, 30 for each of the three grade groups. Approximately 45 persons were selected from the nominee pool for each grade and contacted about serving as an ALS panelist. Some of the persons who were selected were unable to serve at the scheduled time. A total of 87 panelists participated in the ALS study representing 22 states and Guam. A list of the panelists who participated in the ALS is presented in Appendix B.

THE ACHIEVEMENT LEVELS-SETTING PROCESS

OVERVIEW

The purpose of the Civics ALS was to produce a set of recommendations for NAGB to consider in establishing achievement levels on the 1998 NAEP in civics. The recommendations would include a set of cutscores on the Civics NAEP to report student performance classified as Basic, Proficient, and Advanced achievement. Further, the recommendations would include a set of exemplar items from the Civics NAEP to illustrate student performance at Basic, Proficient, and Advanced levels of achievement relative to the recommended cutscores. The third component of achievement levels, i.e., the descriptions, had already been finalized and NAGB had given them provisional approval for use in the process of setting achievement levels.

The Civics ALS lasted five days, November 12-16, 1998 (Thursday-Monday). It was conducted at the Ritz-Carlton Hotel in St. Louis. Sessions generally started at 8:30 AM and lasted until 5:00 PM, or later. The study employed three grade panels, one for each grade assessed by NAEP (4^{th}, 8^{th}, and 12^{th}). The NAEP ALS Project Director served as the primary facilitator for the five-day study. Three content facilitators and three grade group facilitators (one for each grade) assisted the Project Director during the meeting. All facilitators had participated in the civics pilot study and were experienced in the procedures.[9]

SESSION FORMATS AND FACILITATION

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

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

[9] A list of ALS staff and observers has been presented in Appendix A.

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

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

ITEM RATING GROUPS AND TABLE DISCUSSION GROUPS

Within each grade group, panelists were divided into two different item rating groups of about 15 persons: group A and group B. These groups provided a means of monitoring the ALS process by evaluating the similarity of ratings of both groups at different stages of the process. Each rating group was further divided into three discussion groups of 4 or 5 persons per table for each grade group. The demographic attributes of panelists were considered when assigning members to the item rating groups and to table groups; otherwise, the assignments were random. The goal was to have groups as equal as possible with respect to panelist type, gender, region, and race/ethnicity. The demographic profiles for the item rating groups and the table discussion groups have been included in Appendix B.

ITEM RATING POOLS

The 1998 NAEP Civics data were used for the Civics ALS meeting. Two item rating pools for each grade were constructed so that they were as nearly equal as possible with respect to item difficulty, item content area, and item type. Detailed information about the item pools has been presented in Appendix D. The design included two item rating groups and two item rating pools which provided the opportunity to examine ratings from each item rating group as a replication of the other item rating group for each grade. Table 3 presents a summary of information describing items in the rating pool for each rating group.

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

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

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

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

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

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

STEP 1: BRIEFING MATERIALS

Before the ALS meeting, all panelists were mailed materials that contained important background information on setting achievement levels. (See Appendix E.) The first advance packet was mailed August 4, 1998 and contained materials that panelists were required to study. The second mailing was October 15, 1998 and contained detailed instructions related to travel arrangements and accommodations. The briefing materials and information included:

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

STEP 2: GENERAL ORIENTATION AND TRAINING EXERCISES

In the opening session, panelists were given an orientation to the achievement levels-setting process and a complete overview of the procedures planned for the ALS study. During the orientation session, a member of the NAGB staff presented a history of NAEP, a general overview of the NAEP program, a description of the method used to develop the 1998 Civics NAEP Framework, and other such general information about NAEP and NAGB. At the start of the second full day, panelists reviewed this information to help with understanding the overall process.

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

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

Facilitators are given an outline to follow, and the outline is shared with panel members in each grade so that they can refer to the steps in the outline while performing exercises and tasks.

These procedures, along with the *Briefing Booklet* for panelists, make it relatively easy for panelists to identify each element in the process and to understand how each one fits into the overall ALS process.

TAKING A FORM OF THE NAEP

Following the general orientation session on the first day, panelists went to their assigned grade groups where they took a form of the NAEP developed for their grade group. After completing the assessments, they reviewed their own responses relative to the scoring guides. Two forms of the assessment were administered to the panelists for each grade. Item blocks in the form administered to rating group A were excluded from their item rating pool, and the same was true for the blocks in the form administered to rating group B.[10]

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS

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

In grade-level sessions, content facilitators guided the panelists through an extensive training session focused specifically on the achievement levels descriptions for their grade. Panelists were led in an evaluation of the ALDs to compare performance across levels in their grade and to

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

compare performance across each grade within each level. Panelists discussed the ALDs and participated in several training exercises to help them understand the descriptions. In one exercise, panelists used their understanding of the ALDs to determine the level of achievement that would be required of students to answer correctly and completely all items in one item block. This exercise was designed to help panelists become familiar with items of different types (i.e., multiple choice and constructed response) and to understand how the ALDs relate to all types of items. The exercise also helped panelists to become familiar with items that would not be included in their rating pools. Finally, it helped them to become familiar with the structure of NAEP item blocks and with scoring rubrics.

In another exercise, panelists applied their understanding of the ALDs more holistically (i.e., one block of items vs. one test booklet with four blocks of items). A sample of ten student assessment booklets was given to judges to review and discuss with respect to their understanding of the ALDs. Panelists were asked to determine if the performance exhibited in each booklet should be classified as Basic, Proficient or Advanced. After classifying each booklet independently, panelists discussed their classifications with each other. This exercise helped panelists to gain a better understanding of the ALDs and to become familiar with additional NAEP items and scoring rubrics. Discussions of the performances in booklets, relative to the ALDs, helped panelists to become more conversant with the ALDs and to internalize their meaning more completely.

UNDERSTANDING BORDERLINE PERFORMANCE

After working with the ALDs throughout the morning and early afternoon, panelists were trained in the concept of borderline performance[11] and instructed in rating at the borderline. As part of the training in borderline performance, the general session included a demonstration of how items would be rated. This demonstration was designed to help panelists understand the importance both of forming a clear understanding of borderline performance and of having the same understanding shared among panelists.

In grade groups, panelists continued to develop their concept of borderline performance by writing descriptions of student performance at the borderline of each achievement level. Developing descriptors of borderline performance assisted panelists in forming a common understanding of the ALDs as well as a common understanding of borderline performance. Each grade group had drafted a set of borderline descriptions by the close of Day 2. Content facilitators evaluated those sets across all grade levels to make certain that they represented "reasonable" descriptions of borderline performance—not too low and not too high.

The borderline descriptions were distributed to panelists at the start of Day 3 for review and modification. They were aware that the first round of item ratings would begin later that day (Day 3), and the goal was to make certain that they had a set of descriptions that would be useful for the rating process. A review of borderline descriptions was scheduled just prior to the training session for Round 1 ratings. As a means of keeping panelists focused on the ALDs, they

[11] Borderline performance refers to the level of performance that is minimally acceptable for each achievement level.

evaluated borderline descriptions throughout the process. Borderline descriptions were evaluated and modified, as needed, until panelists were ready to begin the final round of item-by-item ratings on the last day.

PAPER SELECTION EXERCISE

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

Panelists were instructed to chose one student paper that represented performance at the borderline of each achievement level. To do this, panelists first selected papers to represent performance at the borderline of each achievement level for each constructed response (CR) item in a block. As a group, panelists reviewed and discussed paper selections for the CR items in one block together before they independently reviewed the CR items in the remaining blocks. After selecting a paper to represent borderline performance at each achievement level for all the CR items in one block, panelists could refer to a sheet where scores for each paper were recorded. The basis for selection, however, was to be their understanding of the ALDs and borderline performance, not paper scores. If no paper was found to represent borderline performance, then no paper should be selected for that particular item at that particular level. The purposes of the exercise would be accomplished by having panelists evaluate papers and determine whether any could be found to represent borderline performance.

The training activity was designed to accomplish the following purposes:

- to provide a reality check on how students responded to open-ended questions;
- to promote a clear conceptualization of performance at the borderline;
- to familiarize panelists with the scoring rubrics for constructed response items.

STEP 3: THE ITEM RATING PROCESS AND FEEDBACK

The general procedure followed for the item rating process included instruction in a general session involving all panelists to assure that they were given the same information. Process facilitators reviewed the instructions and answered questions from panelists in the grade-level sessions. The rating tasks were performed by panelists in grade-level sessions. Similarly, feedback information was first presented in a general session where panelists learned what it was and how to use it. All feedback for the first two rounds was distributed to panelists for review and discussion in their grade groups.

ROUND 1 RATINGS

Following the Paper Selection training exercise on Day 3, all panelists participated in a general session that involved instruction in the item-by-item rating process. The Mean Estimation method

(ME), which is a form of the modified-Angoff rating method, was used. The rating method had been described in the orientation sessions of the first two days, and a demonstration had been given in Day 2. The procedure was reviewed in detail, and panelists were instructed in marking their rating forms. For multiple choice items, panelists estimated the probability that a student performing at the borderline of each achievement level would respond correctly. Panelists were instructed to think of this task as one of estimating the number of students who would give a correct response. They were told to think of a class of 100 students whose performance just met that of the ALD for a particular level, and estimate the number of correct responses from 100 borderline students. For constructed response items, judges estimated the mean or average score (e.g., 2.4 on a scale of 1-3) of students performing at the borderline of each level. They could again think of a classroom with 100 borderline students for each achievement level and estimate the average score for those students on each constructed response item. Once trained, the panelists were ready for Round 1 rating.

Panelists were told to read each item carefully, compose a mental response to the item, and refer to the scoring rubric. This procedure would help panelists form a clear concept of what was required of students. For each item in their item rating pool, panelists marked their estimate of borderline performance at each of the three achievement levels. Panelists were not allowed to discuss item ratings with each other. They were encouraged to refer to the achievement levels descriptions and descriptions of borderline performance. A copy of a rating form has been included in Appendix C. Most panelists completed the first round of ratings within three hours.

FEEDBACK AFTER ROUND 1

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

Cutpoints

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

Standard Deviation

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

Whole Booklet Feedback

Whole booklet feedback is produced for the set of items in the NAEP exam booklets that are administered to panelists as part of the orientation process on Day 1. Each rating group (A and B) was administered a different assessment booklet form. The whole booklet feedback reports the percent of total possible points that a student needs to earn in an assessment booklet in order to meet the minimal requirements for performance at each achievement level (i.e., at the cutpoint). For example, the whole booklet feedback report might state: "Based on the cutscore for your grade, students performing at the borderline Basic level are expected to get 49% of the total possible score points for this booklet." A similar statement is given for each achievement level. This feedback is based on the cutpoints the grade group had set during the first round of ratings, and it is updated after subsequent rounds of ratings. Panelists are informed of the reasons that would cause the percentages to differ for the two booklet forms used by Groups A and B, i.e., different item combinations resulting in different student performance and total points possible.

Whole Booklet Exercise

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

Panelists are given up to 4 booklets to review as representative of borderline performance at each achievement level. One hundred booklets for each of the six NAEP forms (one for each rating group) are randomly selected for this exercise. Student responses are photocopied for review by panelists, but no student background data are shared. Because the booklets are randomly selected, there is a fairly high probability that none will be available to represent performance at some cutscore(s). In cases for which no booklets are available that had been scored within 2% of the total possible points associated with the cutscore, then no booklets are presented to panelists for that achievement level. Panelists are given a complete explanation of the source of booklets and the reason for which no booklet is available.

Rater Location Feedback Charts

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

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

Student Performance Data

Panelists receive information about overall student performance on each item within each block. The proportion of students who gave the correct answer is listed as the actual "p-value" for each dichotomous item. The mean (average) score is reported for each polytomous item, along with the percentage of student responses scored at each rubric score point. The data also report various categories of "no response" for each item. Student performance data serve as a "reality check" because they show how students actually perform on each item. The data indicate how easy or difficult the items are for all students who took the 1998 Civics NAEP. They do not indicate how easy or difficult the items are for students at different achievement levels.

Reckase Charts

For each block of items in the item rating pool, panelists are given a Reckase Chart that indicates expected performance for students scoring at each score point on the ACT NAEP-like scale. Each column represents the range of IRT-based performance estimates for one assessment item. Each row represents IRT-based performance estimates for all items in a block for students scoring at a specific point on the ACT NAEP-like scale. The ACT NAEP-like scale scores range from the score associated with the value at the lower asymptote for any item in the grade-level item pool to the value associated with the upper asymptote. The score range for grade 4 is 39 to 273; for grade 8, the range is 13 to 301; and for grade 12, the range is 27 to 303. For dichotomous items, the probability of correct response (p-values) at each scale score point is reported for each item. For polytomous items, the expected score (mean) is reported for each item at each scale score. The expected performance across scale score points can be observed for each item, as can the

expected performance across items for students scoring at a particular scale score. A sample Reckase Chart has been included in Appendix G. Please note that only data for odd-numbered scale scores are reported on the charts in order to save space and fit the necessary data on the 11"x17" charts.

A special presentation was developed to train panelists in the use of Reckase Charts. The training presentation is specifically designed to reduce "number shock" experienced by panelists in earlier studies when introduced to the charts. Initially, only one column of data from a Reckase Chart is displayed and explained, then only one row. After panelists understand what the charts contain and what the numbers represent, they are able to view the entire chart without being overwhelmed by the large quantity of numbers on the charts.

Panelists mark their charts with both the grade-level cutscore and their own cutscore for each achievement level. Grade 4 panelists marked four charts and grade 8 and 12 panelists each marked five charts. Panelists' individual item ratings are electronically marked on the Reckase Charts. Panelists are instructed to connect their item ratings on the charts to help them evaluate the consistency of their ratings. Panelists draw a line to connect one item rating to the next for all ratings at each achievement level. They use three colored markers to distinguish the three achievement levels. Each block of items is printed on one chart page.

By examining the charts, the panelists are able to consider the relationship between their estimates of student performance for each item and the IRT-based expected student performance at the cutscores. Further, panelists can consider any observable patterns in their ratings, such as differences in the of ratings for multiple choice and constructed response items relative to performance level, or varying levels of consistency in item ratings with respect to a specific achievement level. They can also look for indicators of rater fatigue, such as less consistent ratings for items in the last block of the rating pool. Panelists are informed that if their judgments of students performing at the borderline of each achievement level exactly fit the estimates generated by a statistical model based on actual student performance, all of their ratings would fall along a single row. In other words, if panelists' ratings are on a single row, their ratings perfectly match IRT-based estimates of student performance.

ROUND 2 RATINGS

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

FEEDBACK AFTER ROUND 2

Day 5, the last day, was a busy day. Both cutscore and consequences data were presented in the general session, so all panelists at all grades were informed about these data for each grade.

Panelists were instructed in the use of consequences data, which were presented as graphs reporting the percentages of students scoring at or above each achievement level based on Round 2 cutscores.[12]

Feedback information was distributed to panelists in grade groups. Most of the same types of feedback were presented to panelists after Round 2 as were distributed after Round 1. Grade-level consequences data were added and the whole booklet exercise was omitted. (Please see Appendix F for feedback information based on the second round of ratings.) Panelists had time to review the feedback data, ask questions, and discuss concerns before beginning the third round of ratings. They also had the opportunity to review and modify the ALDs and borderline descriptions prior to the Round 3 ratings.

ROUND 3 RATINGS

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

FEEDBACK AFTER ROUND 3

Round 3 item ratings were again entered into data files for computations and analyses. Feedback data were produced for panelists, based on Round 3 ratings. Reckase Charts were not produced for Round 3 ratings, and individual-level consequences data were added to inform the panelists about their own cutscores.

Feedback data from Round 3 ratings were distributed to panelists in a general session. Panelists were seated in grade groups according to their panelist identification numbers so that materials could be distributed easily.

Consequences data were presented in three different formats. First, there was an update of the grade-level consequences data using the same format as that used for Round 2 consequences feedback. Second, rater location charts were modified to display also the percentages of students who scored at or above score points, reported in increments of 5 points on the ACT NAEP-like scale. Third, individual consequences data were listed for each panelist in each grade group. The list contained panelists' secret ID codes, their cutscores on the ACT NAEP-like scale for each achievement level, and the percentages of students performing at or above the individual panelists' cutscores. Together, these different ways of presenting consequences data provided panelists with a large amount of rather specific information they could use to make recommendations for their final cutpoints.

[12] In past ALS meetings, consequences data were provided for the first time after the final round of item ratings, when panelists could no longer adjust the cutscores. NAGB's Achievement Levels Committee approved the recommendation by the Technical Advisory Committee on Standard Setting (TACSS) to provide grade-level consequences data as part of the feedback following Round 2 ratings.

STEP 4: MAKE RECOMMENDATIONS FOR FINAL CUTPOINTS

Panelists were given a few minutes to review the consequences data before they received a consequences data questionnaire. The questionnaire items asked whether panelists would want to make changes to any of the cutscores after learning the consequences of their cutscores. The relationship between cutscores and consequences data was made clear, i.e., raising cutscores lowered percentages of students performing at or above the cutscores. Panelists could recommend a different cutscore to represent each achievement level for any or all three cutscores. The individual Round 3 cutscores were used to compute the final grade-level cutscores for panel members who recommended no changes to their cutpoints. Panelists were fully informed that ACT intended to recommend these final cutpoints to NAGB, unless there were compelling reasons to do otherwise. They were also informed that the new, final cutpoints would be used for selecting exemplar items.

STEP 5: SELECTION OF EXEMPLAR ITEMS

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

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

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

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

Items that met the statistical criteria for difficulty *and* discrimination constituted the primary list of exemplars, whereas items that met the difficulty criterion *only* constituted the secondary list. Each achievement level had a primary and secondary list of items for consideration as exemplars.

Panelists determined whether or not each item on the lists would serve as an appropriate illustration of performance required at the specific achievement level, based on the achievement levels descriptions. From the list of items that satisfied the statistical criteria, panelists identified items and student response scores that matched the descriptions of student performance at each achievement level. They "approved" or "vetoed" each item and student response. The number of exemplars selected for each achievement level ranged from 7-16 items. The primary and secondary lists of exemplar items for each grade have been included in Appendix H. Also displayed with the lists are the average conditional probabilities.

STEP 6: EVALUATIONS THROUGHOUT THE PROCESS

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

FINAL CIVICS ALS WRAP-UP

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

Feedback after Recommendations for Final Cutpoints

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

After reviewing the final cutscores and grade-level consequences data, each panelist was again asked to respond to a questionnaire regarding the consequences data and the final cutscores he/she would recommend to NAGB. Panelists were aware that their responses were only recommendations and that no changes would be made in cutscores on the basis of those recommendations. The stated purpose of collecting their recommendations was to share with NAGB the opinions of panelists regarding the final cutpoints and the consequences associated with them. If NAGB found their responses compelling, then NAGB could make adjustments.[13] When the panelists completed the final questionnaire, they were thanked for their work and the meeting was adjourned.

[13] ACT shared these recommendations with TACSS for review and evaluation. If TACSS had found reason to recommend changes in the final cutscores after reviewing these responses, ACT would have notified NAGB.

OUTCOMES OF THE CIVICS ACHIEVEMENT LEVELS-SETTING STUDY

The Civics ALS was designed and implemented to produce a set of recommendations for NAGB to consider in establishing achievement levels on the 1998 Civics NAEP. The recommendations included a set of cutscores on the Civics NAEP to represent minimal levels of student performance classified as Basic, Proficient, and Advanced achievement. Further, the recommendations included a set of exemplar items from the Civics NAEP that would illustrate student performance at Basic, Proficient, and Advanced levels of achievement in accordance with the recommended cutscores. And finally, the recommendations included achievement levels descriptions that had been developed and provisionally approved by NAGB for use in the ALS process.

The process ACT designed and implemented to produce the cutscores and exemplar items has been well documented in the first section of this report. Additional details and results of the ALS process are presented in this section. This information provides further evidence of the merit of ACT's recommendations to NAGB. The following ALS process outcomes have been evaluated:

- Cutpoints and their standard deviations
  - Comparison of Cutpoints by Item Type
- Intrajudge consistency
  - Consistency across rounds (changes in ratings from round to round)
  - Reckase Chart analyses
- Consequences questionnaire data
  - Individual consequences data
  - Grade-level consequences data
- Process evaluation questionnaire data
- Selection of exemplar items

The ALS process was evaluated on the basis of whether the following outcomes were achieved:

- reasonable cutscores;
- relatively low standard deviations of those cutscores;
- reasonable levels of judges' item rating consistency, between and within rounds;
- high level of positive responses to the process evaluation questionnaires;
- adequate number of exemplar items selected;
- patterns of results consistent with previous studies; and
- absence of extreme reactions to consequences data.

EVALUATION OF CUTPOINTS AND THEIR STANDARD DEVIATIONS

The cutscores[14] and their standard deviations have been included in Table 4. In nearly every case, the cutscores *increased* from round to round, while the standard deviations *decreased* for each round of ratings. It has been common for the standard deviation to decrease from round to round, so this positive outcome was expected. The cutscores for the civics pilot study showed an

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

uncommon pattern of increasing across each round at each level for each grade. Although the cutpoints for most levels and most grades increased across rounds in the civics ALS, the increases were very slight.

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

Grade	Achievement Level	Data	Round 1	Round 2	Round 3	Final
4	Basic	Cutpoint	147.4	148.6	149.7	150.2
		SD	10.0	5.7	5.4	4.9
		%≥	**76.2**	**73.1**	70.5	69.5
	Proficient	Cutpoint	163.8	164.1	164.6	164.7
		SD	4.9	3.5	3.4	3.2
		%≥	**25.8**	**24.8**	23.7	22.8
	Advanced	Cutpoint	175.5	177.0	177.8	177.8
		SD	5.9	4.4	4.5	4.0
		%≥	**3.2**	**2.2**	1.7	1.7
8	Basic	Cutpoint	148.0	149.3	149.7	149.2
		SD	9.6	6.0	5.6	5.3
		%≥	**73.3**	**70.8**	69.0	70.8
	Proficient	Cutpoint	165.1	165.2	165.4	165.4
		SD	3.5	3.0	2.9	2.8
		%≥	**23.3**	**23.3**	22.2	22.2
	Advanced	Cutpoint	177.1	177.1	177.1	177.9
		SD	4.4	3.8	3.8	3.0
		%≥	**2.3**	**2.3**	2.3	1.9
12	Basic	Cutpoint	150.6	150.3	150.9	*151.2*
		SD	7.1	5.2	5.1	*3.9*
		%≥	**66.9**	**67.8**	65.9	*64.9*
	Proficient	Cutpoint	163.6	163.9	164.1	*164.1*
		SD	4.0	3.6	3.6	*3.0*
		%≥	**28.3**	**27.3**	26.2	*26.2*
	Advanced	Cutpoint	174.2	174.8	175.2	*175.2*
		SD	5.8	3.6	3.6	*3.4*
		%≥	**5.4**	**4.6**	4.2	*4.2*

Bold font represents data that were not presented to panelists.
Italics: Data Corrected after ALS Panels Adjourned.

For each grade, the variance associated with the Basic cutscores for Round 1 ratings was higher than that for the other cutscores. This is typically the case for NAEP. Panelists seem to experience relatively more difficulty in forming a clear concept of borderline Basic performance. This is evidenced both by relatively higher standard deviations of the Basic cutscores for all rounds of ratings and by panelists' responses to questions regarding their concept of borderline performance at each achievement level. Perhaps this difficulty stems from the fact that there is no definition of performance below the Basic level, so forming a concept of borderline Basic performance is relatively more difficult. By Round 3, however, the standard deviation was very low for *all* levels and grades. No patterns of statistically significant differences appeared when comparing cutscores by panelist type, grade, round of ratings, gender, region, or race/ethnicity. When comparing

cutscores within grade by rating groups (groups A and B) and table groups, no major differences were noted. A few of the tests for differences by group were statistically significant, as would be anticipated when conducting multiple post hoc analyses. However, the few, random, significant differences did not suggest any unusual patterns. For a detailed report of the test results for group differences, please refer to Appendix I.

COMPARISON OF CUTPOINTS BY ITEM TYPE

The difference in cutscores by item types, i.e., multiple-choice (dichotomous) items and constructed response (polytomous) items, has been an on-going interest in the NAEP ALS process. In general, the cutscores set for polytomous items are higher than those set for dichotomous items. As anticipated, panelists for the Civics ALS set statistically significantly higher cutpoints for polytomous items than dichotomous items across achievement levels and across grades for all rounds of ratings. Detailed analyses of the cutpoints by item type have been presented in Appendix J.

As the rounds of ratings progressed, changes in polytomous and dichotomous item ratings resulted in smaller differences between cutscores based on the two item types. Figure 1 shows cutpoints for each achievement level, across rating rounds, for grade 4. Panelists tended to raise their dichotomous item ratings at the borderline Basic level between Rounds 1 and 2 and to lower their polytomous item ratings at the borderline Proficient and Advanced levels. While the cutscores for the two item types became much closer after the initial round of ratings, polytomous item ratings still resulted in higher cutscores. Data in Figure 2 show that grade 8 cutscores for dichotomous and polytomous item ratings changed very little. Differences by item format were very small, even at Round 1.

Round 1 ratings by grade 12 panelists resulted in a dichotomous cutscore that was much lower than that computed from their polytomous item ratings for the Basic level. Between Rounds 1 and 2, grade 12 panelists tended to raise their dichotomous item ratings at the borderline Basic level and to slightly lower their polytomous item ratings at all levels. For polytomous items, the change from Round 1 to Round 2 was most noticeable at the borderline Advanced level. Figures 1-18 in Appendix J provide additional evidence of this pattern by graphically showing how individual panelists' ratings for dichotomous and polytomous items moved closer together over rounds of ratings.

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

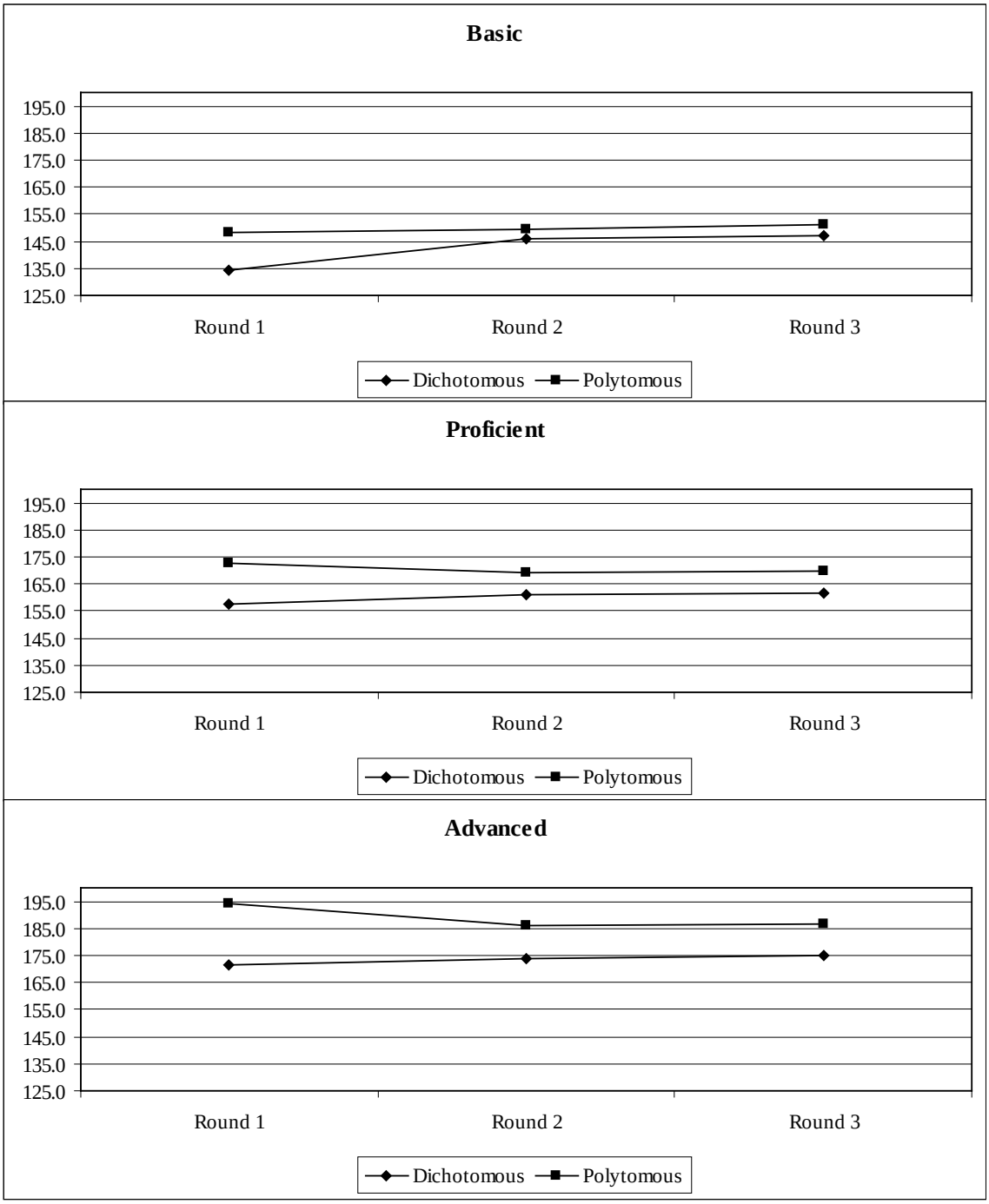


Figure 2
Grade 8 Cutpoints Averaged Across Panelists, by Item Type

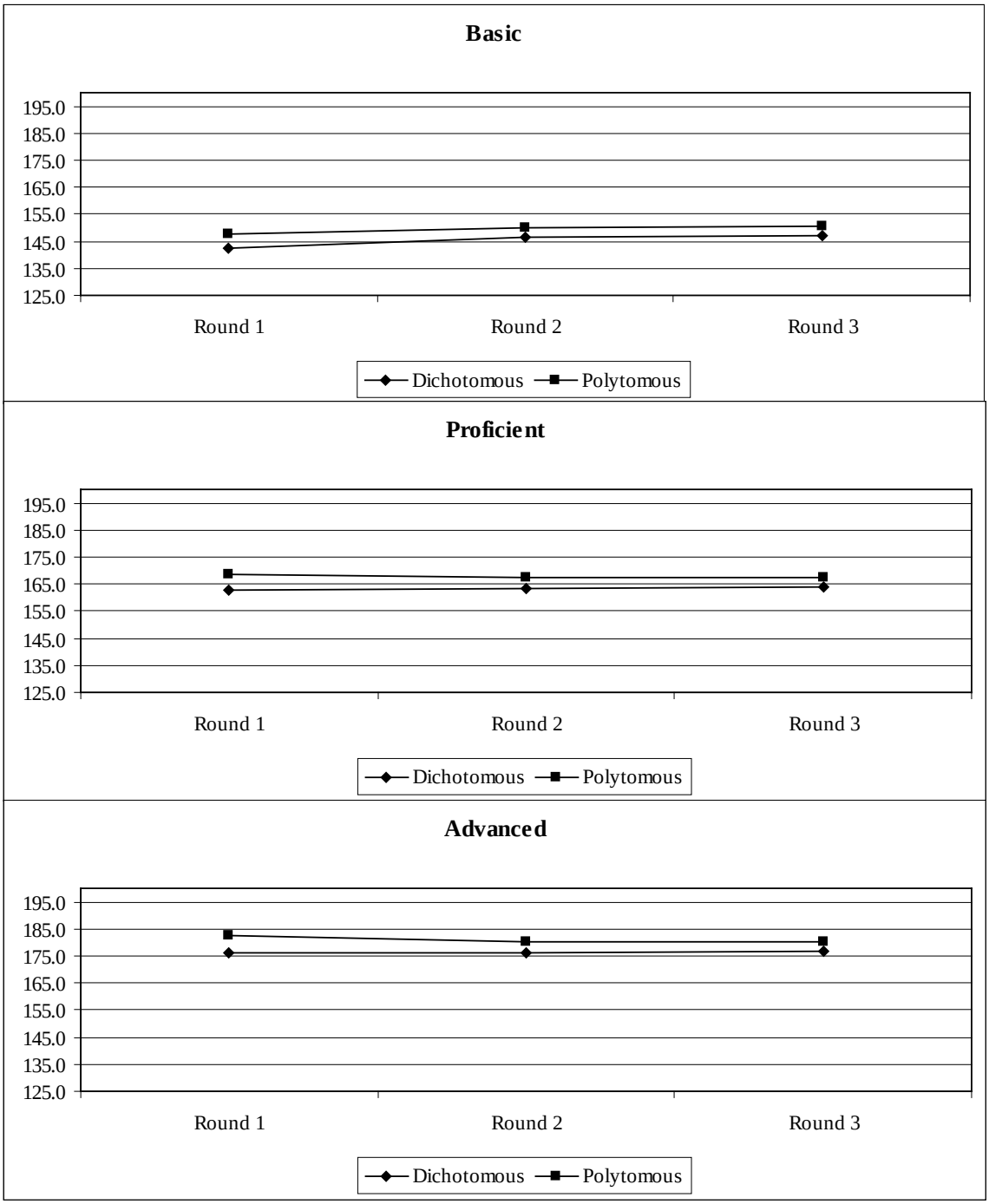
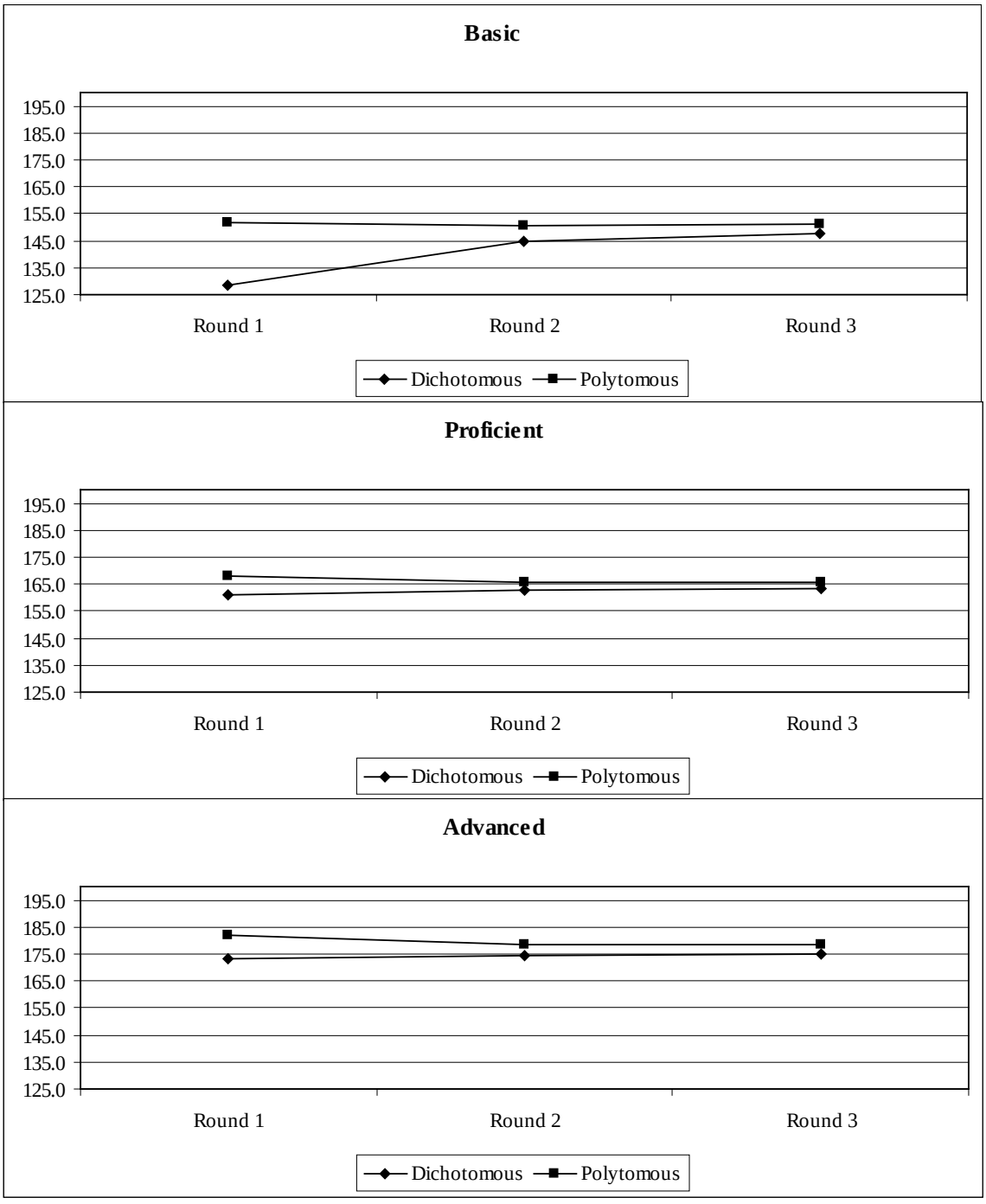


Figure 3
Grade 12 Cutpoints Averaged Across Panelists, by Item Type



EVALUATION OF INTRAJUDGE CONSISTENCY

Intrajudge consistency, both within rounds and across rounds, is generally regarded to be a reasonable criterion by which to judge a standard setting process. Indicators of intrajudge consistency include both the magnitude of change in item ratings from round to round and the number of item ratings changed from round to round. ACT examines these indicators as part of the data analyses *after* an ALS process has been completed. These comparisons of rating changes are "across rounds" measures of intrajudge consistency.

ACT has examined "within rounds" forms of intrajudge consistency data as well. ACT has provided intrajudge consistency feedback to panelists during the ALS process to inform them about the consistency of their ratings for specific items, relative to their overall item ratings. The difference between panelists' individual item ratings and the overall estimate of student performance at the borderline or cutscore provides a "within rounds" indicator of intrajudge consistency. Efforts to provide this intrajudge consistency data as feedback were not considered successful in ALS processes administered before 1998. Reckase Charts provided a means of providing this type of consistency information to panelists, along with several other consistency indicators.

INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The consistency of a judge's ratings across rounds can be examined by evaluating the percentages of items for which the ratings were changed from round to round and the magnitude of change in ratings from round to round. After reviewing the feedback presented following Round 1, Civics ALS panelists were given the opportunity to change their ratings for Round 2. These changes have been reported as percentages of item rating changes in Table 5. The same procedure was followed after Round 2 when panelists could change their ratings for Round 3. These percentages of item rating changes have been displayed in Table 6. The bar graphs in Figure 4 show the percentages of items for which ratings were raised, lowered, and unchanged from one round to the next. The bar graphs in Figure 5 show the magnitude of average rating changes from one round to the next for different types of items.

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

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	48	42	37	11	13	16
	B	43	36	34	14	17	21
8	A	30	17	17	12	13	15
	B	30	24	24	16	15	19
12	A	49	31	29	20	26	29
	B	51	40	40	19	18	20

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

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	18	17	17	5	5	7
	B	16	13	15	3	3	3
8	A	8	5	6	5	4	5
	B	11	9	11	5	4	5
12	A	13	7	7	5	5	3
	B	19	13	14	4	2	2

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

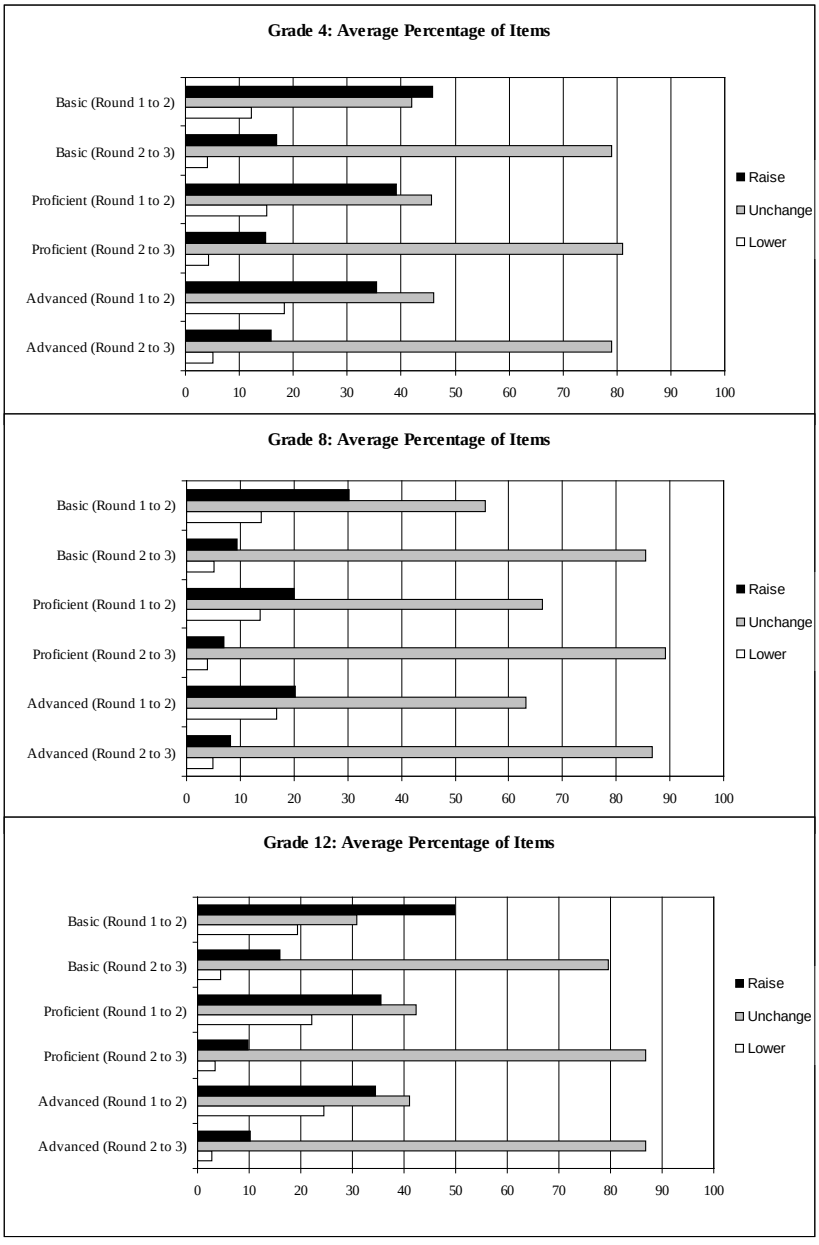
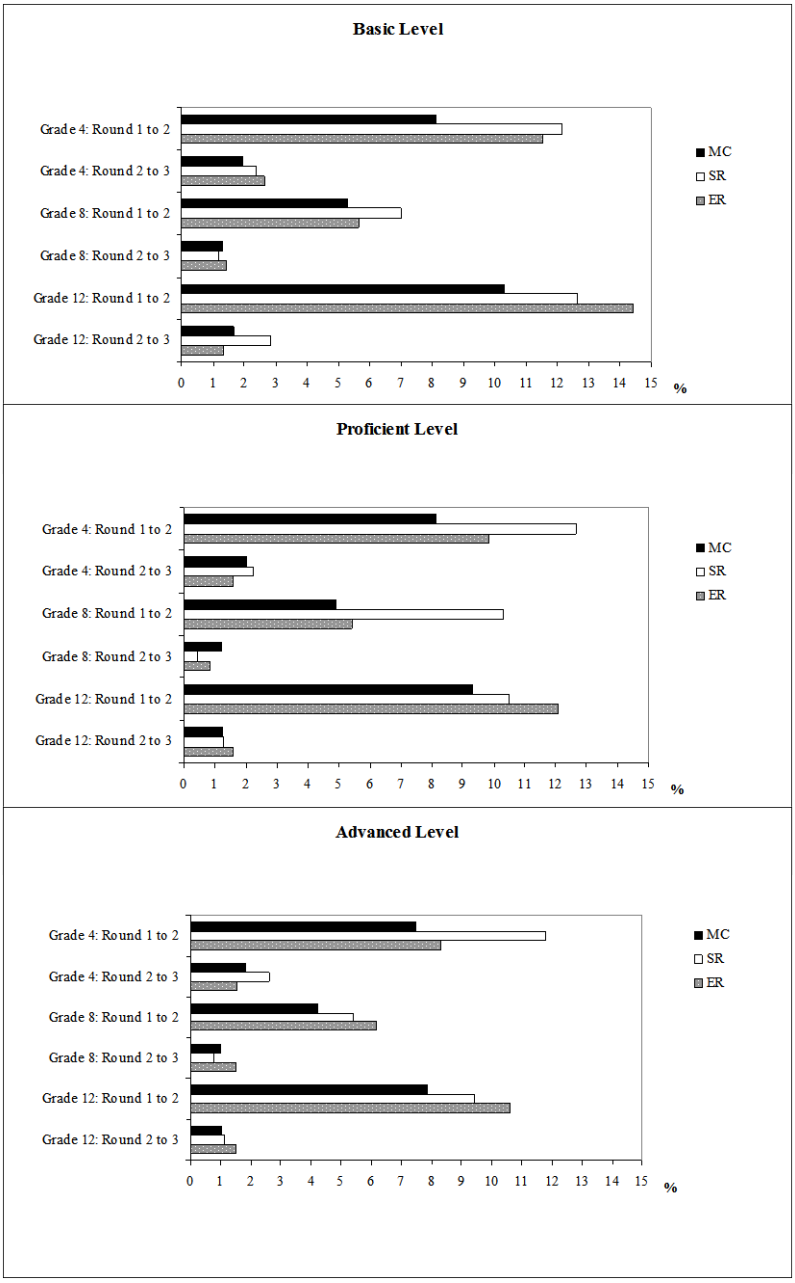


Figure 5
Magnitude of Average Rating Change



Findings from previous studies show that across all grades and all achievement levels, panelists usually change their ratings on fewer items from Round 2 to Round 3 than from Round 1 to Round 2. Grades 4 and 12 civics panelists changed approximately 60% of their item ratings from Round 1 to Round 2, and the grade 8 panelists changed only about 40% of their ratings. From Round 2 to Round 3, however, only about 20% of item ratings were changed. Further, across all grades and achievement levels, panelists tended to raise their ratings for more items than they lowered them. No noteworthy differences appeared when comparing changes in ratings by rating groups (group A and B) or by table groups.

For all grades and all levels, the magnitude of average rating changes was greater between Rounds 1 and 2 than between Rounds 2 and 3. Further, the magnitude of average rating changes tended to be greater for constructed response items than for multiple choice items for all grades and levels. Detailed analyses of rating changes are included in Appendix K.

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

INTRAJUDGE CONSISTENCY WITHIN ROUNDS (RECKASE CHARTS ANALYSES)

The Reckase Charts were introduced to the 1998 ALS Civics panelists as part of the feedback following Round 1. Panelists marked their individual cutpoints and grade-level cutpoints directly on the Reckase Charts. They analyzed their ratings to discern patterns of consistency and inconsistency with respect to item ratings of any particular item type or category. Panelists repeated this analysis after Round 2.

To study the possible "impact" of the Reckase Charts on panelists' item ratings, each judge's Round 3 observed ratings were compared with their "expected" Round 3 ratings. The "expected" ratings were derived from the panelists' Round 2 cutpoints. Specifically, for each achievement level, an individual panelist's Round 2 cutpoints were used to define a set of "expected" item ratings. These expected item ratings corresponded to the model-based estimates of performances for students who scored at the judge's Round 2 cutpoints.

The direction and magnitude of the differences between panelists' Round 3 actual ratings and expected ratings were analyzed by grade and by achievement level for each type of item (multiple choice, short constructed response, and extended constructed response). Differences between rating groups were studied, as well as differences for individual panelists. Please refer to Appendix L for more details about these analyses.

Results of the analyses of the Reckase Charts did not reveal a clear pattern. For example, ACT calculated the percentages of items for which observed ratings were higher than expected ratings. They repeated the calculations for observed ratings that were lower than expected ratings, and for those that were equal. Differences between observed and expected ratings were found for different item types, but no clear patterns were observed. Further, the average magnitude of the differences for short response items were greater than for extended response items or multiple choice items. There was little information from these analyses that could be interpreted as the "impact" of the Reckase Charts on item ratings. Although extensive analyses were conducted in an attempt to quantify the impact of the Reckase Charts, it was difficult to interpret the results in a meaningful way. A method of analysis that would quantify the impact of the Reckase Charts has yet to be developed.

EVALUATION OF CONSEQUENCES DATA

Prior to the 1998 ALS process, consequences data had been introduced before collecting the final round of ratings in research studies only. In the 1998 Civics NAEP ALS process, panelists were told the percentage of students at each grade performing at or above each achievement level as feedback from Round 2 ratings. Panelists were given consequences data again after the third round of item ratings, when they had the opportunity to adjust their cutpoints in response to those data. Comments were collected from panelists regarding their reactions to and opinions about the consequences of their cutscores.

INDIVIDUAL CONSEQUENCES DATA

Following Round 3 ratings, panelists were given both grade-level and individual-level consequences data. In general, the effects of giving panelists consequences data appeared to be consistent with ACT research. That is, the data had little impact on the cutscores (Loomis, Hanick, Bay & Crouse, 2000a and 2000b). Most panelists made no changes in their cutscores after receiving consequences data, even though they had the opportunity. When asked if these percentages of students scoring at or above his/her cutscores reflected the panelist's expectations, 53 of the 87 panelists (61%) answered "yes" and 29 panelists (33%) said "no." Five judges did not respond. Responses are reported, by grade, in Table 7. Although 29 panelists had expected a higher or lower proportion of students relative to their cutscores, they did not want to change the cutscores to be closer to their expectations. A total of 40 changes (raise or lower) to cutscores were recommended by 25 panelists. Approximately 15% of the 261cutscores were changed, and the changes were a mix of raising and lowering the cutscores for different achievement levels.

Table 7
Number of Panelists, by Grade, Recommending Changes to Individual Cutscores in Response to the Consequences Data

	Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)
Data Reflects Your Expectations?			
Yes	18	21	14
No	12	7	10
No Response	1	1	3
(If no) Change one or more cutscore(s)?			
Yes	11	6	11
No	14	15	10
No Response	6	8	6
Recommend Changes to Cutscores			
Basic			
Raise	5	0	5
Lower	0	2	3
Proficient			
Raise	1	1	2
Lower	1	1	3
Advanced			
Raise	3	3	2
Lower	2	1	5

Data in Table 8 report responses by panelist, according to type. As noted above, most panelists recommended no changes. Of those who did recommend changes, educators tended to recommend raising the Basic cutscore more frequently than general public panelists who tended to be generally satisfied with the Basic cutscore and the consequences associated with it. At the Proficient level, general public panelists were most likely to suggest raising the cutscores. At the Advanced level, fewer general public panelists recommended changes, and they were all for raising the cutscore even higher. One-third of the nonteacher educators recommended that the Advanced cutscore be raised. Teachers who recommended a change were almost evenly split between raising and lowering the Advanced cutpoint. The net effect of changes in cutscores was to slightly raise the cutscores for grade 4 Basic and Proficient, slightly lower the grade 8 cutscore for Basic and raise it for Advanced, and slightly raise the grade 12 Basic cutscore. No panelist seemed to have tried to change the grade-level cutscores by drastically changing his/her own cutscores. Data for individual panelists are reported in Table 1 in Appendix M. Please note that individual consequences data were reported as whole numbers. Responses to questions about the individual-level consequences data have been presented in Table 7.

Table 8
Percentage of Panelists, by Type, Recommending Changes to Individual Cutscores in Response to Consequences Data

Panelist Type	No Change	Raise	Lower
Basic			
Teacher (n=32)	68.8%	21.9%	9.4%
Nonteacher (n=9)	66.7	22.2	11.1
General Public (n=15)	86.7	6.7	6.7
Proficient			
Teacher (n=32)	87.5	9.4	3.1
Nonteacher (n=9)	77.8	11.1	11.1
General Public (n=15)	80.0	20.0	0.0
Advanced			
Teacher (n=33)	72.7	15.2	12.1
Nonteacher (n=9)	55.6	33.3	11.1
General Public (n=15)	80.0	20.0	0.0

GRADE-LEVEL CONSEQUENCES DATA

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

In the wrap-up, when panelists were asked if the final percentages reflected panelists' expectations for the proportions of students scoring at or above the grade-level cutpoints, 72 of the 87 panelists (84%) answered "yes" and 10 panelists (11%) said "no." Five judges did not respond. Results of the recommendations have been presented in Table 9. As those data show, 4 of the 10 recommendations were to raise the cutscores for grade 12 Proficient and Advanced. Otherwise, the final recommendations were to lower the cutscores for grade 12 Basic, and for Advanced at both grades 4 and 8. The data in Table 10 report these recommendations by type of

panelist. The most frequent recommendation for change was make by nonteachers and it was to lower the Advanced cutscore. Those responses were collected to document panelists' evaluations of the final cutscores. No adjustments were made to the cutscores.

Table 9
Number of Panelists, by Grade, Recommending Changes to Cutscores in Response to the Consequences Data

	Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)
Data Reflects Your Expectations?			
Yes	26	25	21
No	4	2	4
No Response	1	2	2
(If no) Change one or more cutscore?			
Yes	2	3	5
No	15	10	11
No Response	14	16	11
Recommend Changes to Cutscores			
Basic			
Raise	0	0	0
Lower	0	0	1
Proficient			
Raise	0	0	1
Lower	0	0	0
Advanced			
Raise	0	0	3
Lower	1	2	0
Recommend to NAGB			
Grade Cutscores as Set	18	22	21
Grade Cutscores Changed	0	2	1
Uninterpretable/No Response	13	5	5

Table 10
Percentage of Panelists, by Type, Recommending Changes to Grade Level Cutscores in Response to Consequences Data

Panelist Type	No Change	Raise	Lower
Basic			
Teacher (n=31)	96.8%	0.0%	3.2%
Nonteacher (n=6)	100.0	0.0	0.0
General Public (n=15)	100.0	0.0	0.0
Proficient			
Teacher (n=31)	96.8	3.2	0.0
Nonteacher (n=6)	100.0	0.0	0.0
General Public (n=15)	100.0	0.0	0.0
Advanced			
Teacher (n=31)	87.1	6.5	6.5
Nonteacher (n=6)	83.3	0.0	16.7
General Public (n=15)	93.3	6.7	0.0

EVALUATION OF PANELISTS' COMMENTS AND PROCESS EVALUATION QUESTIONNAIRES DATA

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

In general, Civics ALS panelists were positive in their evaluation of the ALS process and their experience as participants. The responses of panelists to the process evaluation questionnaires have been presented by grade and by panelist type in Appendix N.

UNDERSTANDING THE RATING PROCESS AND CONFIDENCE IN RATINGS

Data reported in Table 11 show the average responses (5=most positive and 1=most negative) to questions about the rating sessions round by round. As expected, panelists' responses generally reflected an increase in understanding and confidence as the rounds of ratings progressed. In particular, the responses to questions about the clarity of instructions and the level of confidence in ratings increased considerably between Rounds 1 and 3.

Table 11
Civics ALS Evaluation Questionnaires
Summary of Responses to Questions Related to Ratings

Questions	Round	Civics ALS		
		Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)
1. The instructions on what I was to do during the 1st/2nd/3rd rating session were: *(5=Absolutely Clear; 1=Not at all Clear)*	1	4.16	4.03	4.00
	2	4.90	4.41	4.52
	3	4.96	4.72	4.93
2. My level of understanding of the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Totally Adequate; 1=Totally Inadequate)*	1	4.40	4.07	3.70
	2	4.71	4.55	4.48
	3	4.87	4.72	4.93
3. The amount of time I had to complete the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Far too Long; 3=About Right;1=Far too Short)*	1	3.00	3.07	3.07
	2	3.55	3.66	3.41
	3	3.71	3.62	3.78
4. The most accurate description of my level of confidence in the ratings I provided to represent the three achievement levels during the 1st/2nd/3rd rating session is that I was: *(5=Totally Confident; 1=Not at all Confident)*	1	2.90	3.28	2.81
	2	4.00	4.10	3.85
	3	4.58	4.41	4.59

Another point of interest was the response to the question related to the amount of time panelists had to complete the rating tasks (question 3, Table 11). After Round 1, most panelists indicated that the amount of time was about right to complete the task (5= far too long, 3 = about right, and 1= far too short). For each successive round, panels responded that they had more time than they actually needed to do their work. It would seem that participants had plenty of time to complete their rating tasks and were not rushed through the rating sessions.

Data in Figures 6-11 show how panelists' perceptions of the rating methods changed over rounds. These items were added to the process evaluation questionnaires in 1994 to provide more information related to the question of why ratings for multiple choice and constructed response items result in cutscores that are significantly different. The patterns shown here were very

Figure 6: The method for rating conceptually
(5=Totally Agree; 1=Totally
Grade)

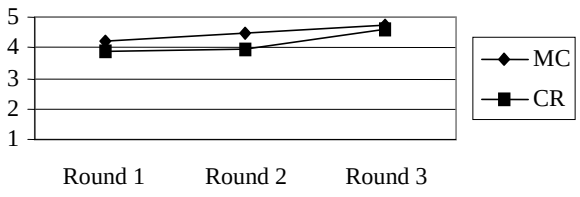


Figure 9: The method for rating items was apply:
(5=Totally Agree; 1=Totally
Grade 4)

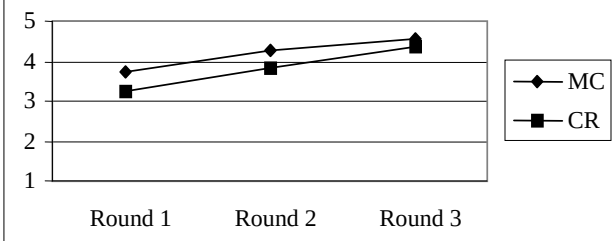


Figure 7: The method for rating conceptually
(5=Totally Agree; 1=Totally
Grade)

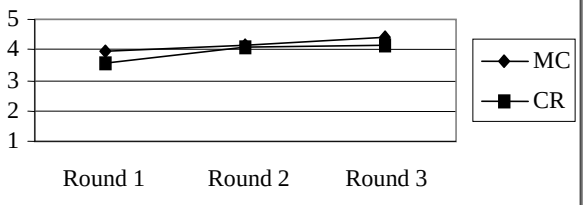


Figure 10: The method for rating items was apply:
(5=Totally Agree; 1=Totally
Grade 8)

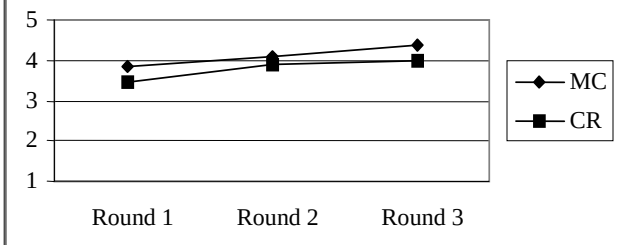


Figure 8: The method for rating conceptually
(5=Totally Agree; 1=Totally
Grade)

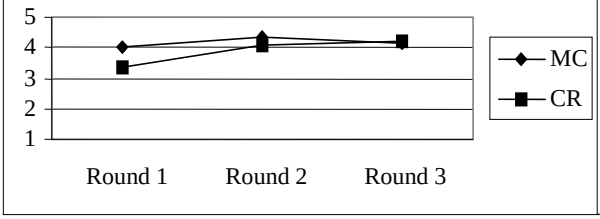
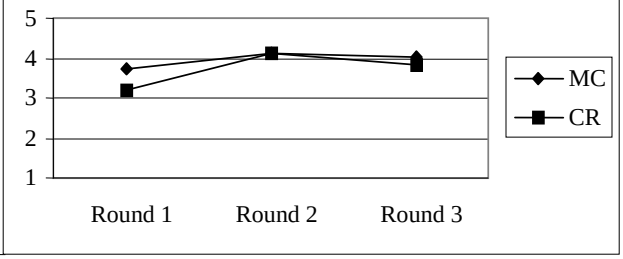


Figure 11: The method for rating items was apply:
(5=Totally Agree; 1=Totally
Di)
Grade 12



similar to those found for other studies. In general, panelists tended to find the rating method easier to apply and conceptually clearer for multiple choice items than for constructed response items in Round 1. By Round 3, their perceptions of both rating methods were about the same. Further, panelists generally found the rating method for both item types to be easier to apply and conceptually clearer across the rounds of ratings. The sharpest increases in positive response were typically observed between Rounds 1 and 2.

UNDERSTANDING OF THE ACHIEVEMENT LEVELS DESCRIPTIONS AND BORDERLINE PERFORMANCE

The typical response pattern that has emerged from past ALS meetings was present for the Civics ALS: Achievement levels descriptions were generally better understood than the borderline descriptions. Panelists' understanding of both categories of performance increased over rounds so that the difference between the two diminished by Round 3. All three grade-level panels indicated highly positive responses when asked about their understanding of the definitions of achievement level performance and borderline performance after Round 3. Panelists' understanding of student performance across the achievement levels approached *absolutely clear* by Round 3, with the lowest average responses reported for Basic level performance. The mean of their responses ranged from 4.4 to 4.6 by Round 3. Their conception of borderline performance approached *very well formed* for all achievement levels by Round 3, with the means ranging from 4.3 to 4.6. Table 12 shows that panelists' overall understanding of student performance at the borderline and at the three achievement levels increased with each round of ratings for each grade.

Having panelists write borderline descriptions did not seem to have an obvious, significant impact on their ability to form a clear concept of borderline performance. The civics panelists wrote borderline descriptions and modified them throughout the process. In contrast, the geography panelists discussed their concept of borderline performance with other panelists and used that in training exercises prior to rating items. Table 12 includes the responses of both geography and civics panelists to questions about their understanding of achievement levels descriptions and their concepts of borderline performance. Geography panelists generally reported a higher level of *clarity of concept* regarding the borderline descriptions than the civics panelists for Round 1, and they also reported a higher level of *understanding for the ALDs* at Round 1. Relative to the grade 4 civics panelists, the geography grade 4 panelists consistently reported higher levels of *understanding the ALDs* and *clarity of concept* formation for borderline performance at all rounds and at all levels of achievement. In contrast, the civics panelists for grades 8 and 12 reported higher levels of *understanding the ALDs* and *clarity of concept* formation for borderline performance for both Rounds 2 and 3 at all levels of achievement.

Table 13 shows a comparison of the relative differences between responses by geography and civics ALS panelists regarding their understanding of the achievement levels descriptions and their concept of borderline performance. This comparison was simply of the difference between the mean responses to questions about understanding the definitions of performance at a level and the extent to which conceptualizations of borderline performance were well formed. In general, the differences between the two were greater for Civics ALS panelists than for Geography ALS panelists. Geography panelists tended to understand both performances equally well. Particularly at grades 8 and 12, the differences at rounds 2 and 3 were greater for civics panelists than for the geography. The differences for grade 4 panelists were nearly the same for both subjects. Another point of interest is the fact that geography panelists at grade 12 gave equal ratings to their understanding of the ALD and their concept of borderline performance across all achievement levels. By Round 3, grade 4 panelists gave higher marks to their concept of borderline

performance than to their understanding of the definition of performance at the level for all three levels of achievement.

Table 12
Civics ALS Evaluation Questionnaires
Summary of Responses to Questions Related to Achievement Levels Descriptions

Questions	Round	Civics ALS			Geography ALS		
		Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)	Grade 4 (n=30)	Grade 8 (n=28)	Grade 12 (n=31)
1. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Basic level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.65 4.10 4.48	3.62 4.17 4.62	3.52 4.37 4.37	3.79 4.29 4.52	3.71 4.11 4.18	3.83 4.21 4.38
2. At the time I provided the 1st/2nd/3rd set of ratings my conception of Borderline Basic performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.35 4.03 4.42	3.28 4.03 4.45	3.11 4.07 4.33	3.50 4.21 4.56	3.25 4.11 4.18	3.59 4.17 4.38
3. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Proficient level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.71 4.13 4.58	3.59 4.21 4.59	3.37 4.37 4.56	3.86 4.33 4.56	3.68 4.11 4.29	3.83 4.21 4.38
4. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Proficient performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.29 4.06 4.52	3.28 4.03 4.38	3.15 4.19 4.44	3.46 4.22 4.63	3.21 4.14 4.18	3.55 4.21 4.38
5. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Advanced level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1 2 3	3.77 4.13 4.55	3.66 4.31 4.59	3.63 4.37 4.63	3.93 4.44 4.56	3.79 4.11 4.32	3.97 4.17 4.38
6. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Advanced performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1 2 3	3.35 4.13 4.55	3.28 4.03 4.41	3.22 4.19 4.56	3.54 4.22 4.59	3.36 4.07 4.18	3.79 4.17 4.38

Table 13
Comparison of the Relative Differences Between Understanding of ALDs and Concept of Borderline Performance for Geography and Civics ALS Panelists

	Understand Basic ALD vs. Concept Borderline Basic Performance		Understand Proficient ALD vs. Concept Borderline Proficient Performance		Understand Advanced ALD vs. Concept Borderline Advanced Performance	
	Civics (wrote borderline)	Geography (did not write borderline)	Civics (wrote borderline)	Geography (did not write borderline)	Civics (wrote borderline)	Geography (did not write borderline)
Grade 4						
Round 1	.30	.29	.42	.40	.42	.39
Round 2	.07	.08	.07	.11	.00	.22
Round 3	.06	-.04	.06	-.07	.00	-.03
Grade 8						
Round 1	.34	.46	.31	.47	.38	.43
Round 2	.14	.00	.18	-.03	.28	.04
Round 3	.17	.00	.21	.11	.18	.14
Grade 12						
Round 1	.41	.24	.22	.28	.41	.18
Round 2	.30	.04	.18	.00	.18	.00
Round 3	.04	.00	.12	.00	.07	.00

*difference between mean responses to process evaluation questions

In conclusion, it seems that by Round 3 the civics panelists, who wrote borderline descriptions, attained a somewhat better understanding of the meaning of the ALDs and greater clarity of the concept of borderline performance than was the case for geography panelists, who did not write borderline descriptions. Contrary to what one might expect, however, the difference between their understanding of the meaning of performance at the levels and their concept of borderline performance was greater than that for geography panelists. Civics panelists wrote descriptions of borderline performance and spent a great deal of time discussing this, whereas geography panelists only discussed their mental conceptualization of borderline performances.

Because some panels might be more inclined to respond positively to elements of their ALS experience than others, it is not possible to ascertain whether there is a real, qualitative difference between the two groups. Responses by civics ALS panelists were, overall, the least positive of any subject area in which NAEP achievement levels have been set, however. Panelists who participated in writing borderline achievement levels descriptions in the civics ALS did not report better formed conceptualizations of borderline performance than panelists in the geography ALS who did not write borderline descriptions.

EVALUATIONS OF FEEDBACK

Many different types of feedback information were given to the panelists during the ALS process. Table 1 in Appendix N (questions 1-5) presents data that indicate most panelists planned to consider *all* of the feedback information when adjusting their ratings, not just one type of feedback. These data suggest that when panelists were modifying their ratings, they were not overly influenced by one type of feedback to the exclusion of all others.

Responses to question 6 in the same table provide further evidence that panelists used different types of feedback when forming their judgments. When asked to choose one, and only one type of information to use during the rating process, each type of information was selected by at least one Civics ALS participant. There was considerable variation across grade groups in response to this question. Grade 4 panelists preferred the Reckase Charts, while the grade 8 panelists favored student performance data and rater location data. The 12th grade panel was evenly split among those three choices: Reckase Charts, rater location charts, and student performance data. Interestingly, the consequences data were not frequently identified by the panelists as the single feedback data of choice.

Panelists were asked to rank order the different types of feedback information, from most helpful to least helpful; and those data are reported in questions 7-11 in the same table in Appendix N. All three grade panels for the Civics ALS ranked the Reckase Charts first and the student performance data second, based on the combined frequencies of first and second choices. The whole booklet feedback was most frequently ranked as least helpful. The Civics ALS grade groups varied noticeably in how they ranked consequences data. The 12th grade panel ranked consequences data as being from moderately helpful to most helpful, whereas the 4th and 8th grade panels ranked consequences data as being from moderately helpful to least helpful. However, consequences data were named least frequently by all grade groups as the most helpful type of feedback, based on the combined frequencies of first and second choices.

RATINGS FOR MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEMS

In 1992 ACT found that the cutscores that would be set for multiple choice items were different than those for constructed response items scored for partial credit. Over the years, panelists have been asked for their input with regard to some possible reasons for this difference. Table 14 displays the data from the Civics ALS panels related to this issue.

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

As has been the case in previous studies, panelists showed some changes in their responses to questions about dichotomous and polytomous items across rounds. As rounds of ratings progressed and more feedback data were evaluated, more panelists responded that differences in ratings for the two types of items were likely to result from student behavior and student performance on the items (question 1). Panelists for other NAEP ALS subjects have tended to change from their initial opinion that the differences were likely because constructed response items assess different dimensions of knowledge and skills than multiple choice items (question 3). That explanation became even more tenable for grade 4 civics panelists, however. Only grade 12 panelists revealed a systematically decreasing acceptance of this explanation across the rounds.

Table 14
Summary of Responses to Questions Related to Multiple Choice/Constructed Response

Question	Round	Civics ALS		
		Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)
1. If the ratings of student performance on multiple-choice items and constructed-response items are very different, this is most likely caused by different student behavior and performance on the items. 5=Totally Agree 1=Totally Disagree	1	3.29	3.11	3.30
	2	3.48	3.10	3.58
	3	3.58	3.38	3.74
2. If the ratings of student performance on multiple-choice items and constructed-response items are very different, this is most likely caused by the different rating methods. 5=Totally Agree 1=Totally Disagree	1	3.10	3.18	3.38
	2	3.03	2.52	3.37
	3	3.19	2.86	3.19
3. I think constructed-response items assess dimensions of knowledge and skills that are significantly different from those assessed by multiple-choice items. 5=Totally Agree 1=Totally Disagree	1	3.97	3.86	4.11
	2	4.13	4.10	3.96
	3	4.13	3.90	3.78

INDIVIDUAL PANELISTS' COMMENTS

Most panelists indicated that the amount of information given to them during the rating process was enough to inform their judgment without causing confusion. They generally agreed that student performance at each cutpoint was about what they expected. Many felt that the right time to get consequences information was after Round 2, while several others suggested that they would have preferred to receive consequences data after Round 1. When asked about the impact of consequences data on the cutscores they recommended to NAGB, most panelists chose to comment on the state of civics education in the U.S.—implying that it accounted for the low performance of students—rather than to answer the question directly. The vast majority of panelists indicated that they felt confident and positive about having the results of the ALS process reported in the *Nation's Report Card for Civics*.

THE OVERALL ALS PROCESS

Data reported in Table 15 show the average responses to questions from the final questionnaire about the overall ALS process. Once again the responses indicate a generally positive reaction to the ALS process. It is possible to compare how panelists described their level of confidence in their overall ratings (question 1, Table 15) relative to their descriptions following each round of ratings (process evaluations 4, 5, and 6; question 4, Table 11). Panelists reported the greatest confidence in their Round 3 ratings. Although panelists at all three grade levels reported a higher

level of confidence in their overall ratings than they reported for Round 1 and Round 2 ratings, their level of confidence for Round 3 ratings was higher than their overall evaluation of their confidence. Recall that after Round 3, panelists were given the opportunity to recommend new cutpoints that would raise or lower the percentages of students performing at or above each level without adjusting the ratings for individual items. Judges commented that they felt making changes to the cutscores was rather arbitrary after the third round of ratings. Perhaps this lessened their confidence and caused the lower rating in their retrospective evaluation. Their overall evaluation of their level of confidence in their ratings does reflect an average of sorts over the levels of confidence reported at the different times during the process, however.

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

Question	Civics ALS		
	Grade 4 (n=31)	Grade 8 (n=29)	Grade 12 (n=27)
1. The most accurate description of my level of confidence in the achievement levels ratings I provided was: *(5=Totally Confident; 1=Not at all Confident)*	4.29	4.21	4.04
2. I would describe the effectiveness of this achievement levels-setting process as: *(5=Highly Effective; 1=Not at all Effective)*	3.97	3.86	3.59
3. I feel that this NAEP ALS process provided me an opportunity to use my best judgment in rating items to set achievement levels for the NAEP Civics Assessment: *(5=To a Great Extent; 1=Not at All)*	4.23	4.07	4.11
4. I feel that this NAEP ALS process produced achievement levels that are defensible: *(5=To a Great Extent; 1=Not at All)*	4.23	4.21	3.96
5. I feel that this NAEP ALS process produced achievement levels that will generally be considered reasonable: *(5=To a Great Extent; 1=Not at All)*	4.39	4.17	4.00

Unfortunately there is some confusion with respect to the responses to the question in Table 16. Because the questionnaire format was revised for scanning purposes for the Civics ALS, panelists marked a separate answer sheet rather than marking directly on the questionnaire. The answer sheet contained five response ovals for each question. This particular question, however, had only four responses. Several panelists darkened oval 5, even though there was no response coded 5. It seems clear that panelists who marked ovals 4 and 5 intended to respond positively to the statement, while panelists who marked ovals 1 and 2 intended to respond negatively. The intention of panelists who marked oval 3 remains unclear. While the intended responses to the question were not entirely clear, it seems apparent that a majority of panelists in each grade would sign their name to a statement recommending the use of the achievement levels.

Table 16
Final Civics ALS Evaluation Questionnaire
Frequency of Responses to Question About Willingness to Sign a Statement Recommending the Achievement Levels

I would be willing to sign a statement (after reading it, of course) recommending the use of the achievement levels resulting from this achievement levels-setting procedure: *(4=Definitely, 3=Probably, 2=Probably not, 1=Definitely not) NR = no response*	Response	n	n	n
	5	2	4	7
	4	23	11	8
	3	6	9	7
	2	0	3	4
	1	0	0	0
	None	0	2	1

EVALUATION OF THE SELECTION OF EXEMPLAR ITEMS

One of the primary outcomes of the ALS meeting is the identification of assessment items that illustrated the knowledge and skills associated with each achievement level to use in reporting NAEP results. Appendix H includes the items selected by panelists to serve as exemplar items for reporting the 1998 Civics NAEP achievement levels.

Panelists reviewed and discussed the items and responses that qualified statistically for consideration as exemplars. (Please see Table 17.) Panelists were instructed to "veto" items or response scores that met statistical criteria but that did not meet substantive, content-related criteria. They were trained in the statistical criteria that had been used for selecting the items (described earlier in this report). In addition, panelists were instructed to use their knowledge of the achievement levels descriptions to evaluate each item on the list in terms of its quality as an illustrative or exemplar item. Panelists were to select items from the secondary list only if fewer than about three of the items on the primary list were acceptable.

Table 17
Number of Assessment Items that Qualified and Were Selected as Exemplars for Civics ALS

	Number Primary List*	Number Secondary List**	Number Selected
Grade 4			
Basic	6	5	10
Proficient	5	6	9
Advanced	4	2	6
Grade 8			
Basic	9	7	15
Proficient	6	6	12
Advanced	7	5	10
Grade 12			
Basic	11	7	9
Proficient	6	6	7
Advanced	5	4	7

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

The item blocks marked for release were identified in advance and used as common blocks for rating pools in each grade. All panelists rated these items, which were part of the grade-group

discussions to train panelists in the paper selection exercise that was implemented prior to the first round of item ratings. Recall that in the paper selection exercise, panelists were asked to select papers to represent the performance of students at the borderline of each achievement level. Thus, panelists were all familiar with all the items in the lists, and they had reviewed and discussed student responses to constructed response items.

The process was based on general agreement among panelists regarding whether each item should be used as an exemplar of performance at a specific achievement level. Facilitators allowed the group, as a whole, to determine whether items would be considered or vetoed. In some instances, an item might be selected although a few people voted to veto the item. In other instances, an item may be vetoed because of intense opposition to the item—even by only a minority of panel members. The general will of the group was the guide.

As data in Table 17 show, panel members vetoed few items. Grade 4 panelists rejected one item on the primary list for consideration at the Basic level, and 2 at the Proficient level. They approved all of the items on the list for consideration as Advanced level exemplars.

Grade 8 panelists vetoed one item on the primary list for Basic exemplars, although a "dissenting minority" was recorded for two additional items. At the Proficient level, none of the items on the primary list were vetoed, but a "dissenting minority" was again recorded for two items. At the Advanced level, two items on the primary list were vetoed by about half of the panelists, so those two were eliminated from the list of items selected to illustrate Advanced level performance at grade 8.

Two of the items on the primary list for grade 12 Basic exemplars were vetoed. Since there were 8 items selected from the primary list, grade 12 panelists did not select any from the secondary list of Basic level items. All 6 of the items on the primary list for Proficient exemplars were approved by the grade 12 panelists, and one additional item from the secondary list was approved. At the Advanced level, all 5 of the items on the primary list were approved, and two items from the secondary list were approved.

In general, there was a mix of both multiple choice and constructed response items for panelists' consideration in the exemplar item selection process. Exceptions to this were at the grade 4 Basic level and grade 12 Advanced level. At grade 4, only one constructed response item qualified for consideration as a Basic level exemplar. At grade 12, only one constructed response item qualified for consideration on the primary list of Advanced exemplars.

ACT chose a set of three exemplars from those recommended by the panelists. The set of items and student papers illustrated each of the three achievement levels for each of the three grades. Both TACSS and the NAGB Achievement Levels Committee reviewed the items. Each committee found a few items or responses that did not seem to be "best choices." In some cases, the content of two items seemed too similar to achieve the goal of providing a comprehensive picture of student performance at one or more level(s) of achievement. In other cases, the content of a specific response included information or language that was deemed inappropriate for general distribution. ACT provided substitutes for those particular items or papers from those

recommended by panelists. The final set of items and papers was approved by NAGB to be used in *The Nation's Report Card for Civics*.

NOTE**:** During the process of analyzing data from the study, an error in the recommended cutscores of one panelist was discovered. The cutscores entered for this Grade 12 panelist were incorrect. There had been too little time at this stage of the process for complete verification of data. The incorrect data had been used to compute the final cutscores used in selecting exemplar items and in determining the consequences data reported to panelists. The magnitude of the error was sufficient to cause a difference in the cutscore for the grade and to change the percentage of students scoring at or above the Proficient and Advanced levels. TACSS was informed of this error and asked to evaluate the alternatives. Their recommendation was followed.

The corrected data were reported to NAGB for Grade 12 and used to compute the cutpoints and consequences data for Grade 12. Although the set of assessment items that would have qualified as exemplars was *slightly* different from those that that were actually considered, the exemplar selection was deemed to be satisfactory. Items were eliminated that did not qualify once the data were corrected. The resulting list was sufficient for reporting grade 12 student performance relative to the achievement levels cutpoints.

CONCLUSIONS DRAWN FROM THE CIVICS ALS STUDY

The procedures designed and implemented by ACT to set 1998 Civics NAEP achievement levels proved to be a highly effective process, as determined by the evaluation criteria. Analyzing the outcomes of the process from start to finish revealed remarkable consistency, agreement, and overall satisfaction of panelists at every stage. Given that the NAEP standard setting process is based on judgments by broadly representative panels of individuals, such consistency is an impressive accomplishment.

THE ISSUE OF IMPROVING AND REFINING THE STANDARD SETTING PROCESS

The process implemented in the Civics ALS was designed to be compatible with the psychometric attributes of NAEP, to meet NAGB's policies and guidelines for setting achievement levels, and to be consistent with best procedures and practices for standard setting known to ACT and TACSS. The result was a highly effective and successful standard setting process. Panelists were able to carry out the process without observed or self-reported difficulty, and their reaction to the procedures was very positive.

It is important to emphasize the refinement of the ALS process designed by ACT to set the 1998 NAEP Achievement Levels in Civics. While few modifications were made to the ALS process implemented for the 1998 NAEP, the modifications represented important changes to the ALS process. The most significant changes were:

- finalizing the achievement levels descriptions before the ALS panels were convened;
- introducing consequences data during the rating process, rather than after the cutscores were set; and
- providing the Reckase Charts as a means of informing panelists about item-level student performance and intrarater consistency.

FINALIZING THE ALDS BEFORE CONVENING THE ALS PANELS

Developing the achievement level descriptions has been an important part of the standard setting process for NAEP. A strong logical connection links NAGB's policy definitions of achievement in general to the operational definitions of achievement in civics. These operational definitions of achievement are the basis of training panelists, and they guide the item rating process. Useful and reasonable outcomes of the ALS process depend upon useful and reasonable achievement levels descriptions.

Prior to convening the 1998 ALS panels, the achievement level descriptions had been carefully crafted and thoroughly reviewed in a well-documented process. The revised achievement levels descriptions were compared not only to the Civics Framework and to the policy definitions, but they were also compared to the item pools for each grade level. The procedure for evaluating and modifying the ALDs prior to the operational ALS studies was judged to be a considerable improvement over previous practices. TACSS had expressed some concern regarding the willingness of panelists to accept the achievement levels descriptions and their ability to internalize the ALDs without a more direct role in shaping their content. In fact, none of the panelists for the 1998 Civics ALS expressed a desire to revise the ALDs they were given to use in the ALS process.

PROVIDING CONSEQUENCES DATA DURING THE PROCESS

Determining when and how much information to provide to panelists has been a continuing concern for the design of the ALS process. Of considerable debate has been the provision of consequences data to judges. The goal has been to provide the best balance of information to panelists so that their judgments will be both realistic and based on the ALDs. To assure judgments were criterion-referenced, panelists in past ALS studies received consequences data only after their final round of ratings. Panelists' reactions to this information were collected and shared with NAGB for consideration when setting the achievement levels.

For the 1998 ALS study, however, NAGB agreed to allow panelists to review consequences data during the process of setting cutscores. Accordingly, panelists first reviewed consequences data after their second round of item-by-item ratings. They were provided consequences data again after the third round of ratings, and they made recommendations for final cutscores based on their evaluation of those data. The rationale for this change was to inform panelists as fully as possible about the many aspects associated with their ratings, including the proportion of students scoring at or above each level, given the cutscores they set in the most recent round of ratings. ACT research collected in previous ALS processes and during field trials and pilot studies for the 1998 ALS indicated that the outcomes would not be significantly impacted by introducing consequences data during the process. Interestingly, the consequences data were regarded by most panelists as just one among many sources of information for their consideration. The concern that consequences data would dominate panelists' judgments was unfounded. Informing panelists of the consequences of the cutscores they set increased confidence in the credibility of the outcomes of the process. The finding that panelists' responses to the consequences did not lead to significant modifications in cutscores increased confidence in the process, in general.

USING THE RECKASE CHARTS AS FEEDBACK

Years of refinements have led to the current process, which has been considerably enhanced by the most recent addition of the Reckase Charts. The charts were created specifically for use in setting NAEP standards, although they could be used easily in other standard-setting contexts. Incorporating the charts into the ALS process helped to overcome difficult technical challenges to setting achievement levels for NAEP. The Reckase Charts proved to be a powerful tool that enabled laypersons to work with item measurement data that otherwise would have been too technical to comprehend. Panelists used the Reckase Charts to evaluate their ratings for each item along several, important dimensions. For example, Reckase Charts showed panelists that the likelihood of students correctly answering an individual item increases as the overall performance of students increases. The Reckase Charts also showed panelists that this was not always the case because some items have little or no discrimination in some ranges of the score scale. All three grade panels for the Civics ALS ranked the Reckase Charts as the most helpful feedback given to them.

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

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS

One persistent challenge to improving the ALS process has been to find a way to provide panelists with information about the relationship between their individual item ratings and student performance. This sounds relatively simple, but the issue is how to identify the relevant level of student performance. Individual item ratings can be used to compute a cutscore for each panelist. That cutscore then becomes the representation of a panelist's concept of borderline performance for a level of achievement. The panelist's ratings across all items are associated with an overall performance score (cutscore). If the item ratings are all for the same performance score, then the panelist has managed to estimate student performance for each item to be perfectly consistent with the IRT model used to estimate student performance on NAEP. No panelist achieved that level of perfection in the 1998 process. Most panelists judged some items to be much harder or much easier than others, relative to their overall cutscore. Intrarater consistency is a measure of the extent to which individual item ratings are consistent with the overall cutscore estimated from the individual item ratings, given student performance on the items. Although this information has been given to panelists as feedback in previous ALS meetings, there was little indication that

panelists either understood the information, or found it useful when modifying their judgments about student performance. Reckase Charts made that information easy to assess.

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

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

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

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

THE ISSUE OF DIFFERENCES BETWEEN MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEMS

The difference between the cutscores that would result from ratings of polytomous and dichotomous assessment items has been another persistent challenge to ACT's effort to refine the standard setting process for NAEP. As has been the case in past ALS meetings, panelists for the Civics ALS 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 when reviewing Reckase Charts. After studying the feedback, including the Reckase Charts, panelists adjusted their ratings for the subsequent round (Round 2 and Round 3). In general, the differences between multiple choice and constructed response ratings were reduced for all grades and all levels for the subsequent rounds. The cutscores set for polytomous items were still higher than those set for dichotomous items. As has been the case for previous investigations in other NAEP ALS studies, differences

in *cutscores* computed for items of the two types are greater than differences in *ratings* would suggest (ACT, 1997a).

To understand this phenomenon better, panelists were asked about some possible reasons for this difference in ratings. Most agreed that the source of any differences was due to the fact that constructed response items assess dimensions of knowledge and skills that are significantly different from those assessed by multiple choice items. Panelists in other NAEP ALS procedures typically favor this "explanation" more at Round 1 than they do by Round 3. Civics panelists showed some fluctuation in their responses across the three rounds as well. By Round 3, panelists agreed least with the statement that the source of any differences in ratings can be attributed to the rating method. Clearly this is not an oversight by the panelists. Although more research is needed to determine how judges perceive polytomous items relative to dichotomous items, the Reckase Charts appear to have been effective in helping to make panelists aware of differences and to modify their judgments of student performance. The rating data and the process evaluation response data, indicate that panelists' ratings were based on the achievement levels descriptions. Student performance simply fell short of panelists' informed judgments of what was reasonably expected of the students who met the standard of the NAEP achievement levels.

THE ISSUE OF PROVIDING CONSEQUENCES DATA DURING THE RATING PROCESS

The impact of consequences data on outcomes has been a topic of considerable interest to NAEP standard setting. No compelling differences were found in cutscores produced by the Civics ALS judges who received consequences data for the first time after Round 2 and judges from other ALS studies who received consequences data after ratings were completed and cutscores set. Judges, in general, found the consequences data informative and useful, but they did not appear to be greatly influenced by the data. When given the opportunity to change their own cutscores after learning of the consequences, few panelists chose to make changes. Those who did tended to adjust their cutscores by only a few points.

THE ISSUE OF COGNITIVE COMPLEXITY

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

PUBLIC COMMENTARY ABOUT THE CIVICS ACHIEVEMENT LEVELS

Achievement levels are necessarily judgmental, and as such there is no unqualified "right" set of performance standards. In an effort to inform NAGB fully of the usefulness and reasonableness of the Civics NAEP achievement levels, ACT collected public opinion and comments regarding

the evaluation of the levels that were produced by the ALS process. The collection of public input is required to inform NAGB regarding their decision to set the cutscores.

Because NAGB sets the cutscores, and because cutscores and performance data cannot be released prior to the announcement of NAGB's decision by the Commissioner of Education Statistics, there is little information about the achievement levels available for the public to review and evaluate. Nonetheless, ACT created a special NAEP achievement levels website to present information and collect public comments for this purpose. The achievement levels descriptions and exemplar items and student papers were posted on the website. A set of questions was developed to direct comments for the opinion survey. Please see Appendix O for more detailed information about the NAEP achievement levels website.

An announcement explaining the purpose of the NAEP survey was sent to various stakeholder organizations, and they invited their membership to respond to the survey. The following persons, groups, and organizations were contacted at the end of March, 1999 and encouraged to participate:

- Persons who served on the Civics NAEP Framework panels
- Members of the EIAC listserv
- Members of the Centers for Civic Education listserv

THE NAEP ACHIEVEMENT LEVELS WEBSITE

Respondents to the NAEP Achievement Levels Opinion Survey were asked to review the general background information about NAEP and the recommended achievement levels summarized for the website. This background information consisted of three sections:

- What are NAEP and the *Nation's Report Card*?
- How are the achievement levels set for NAEP?
- Definitions of Basic, Proficient, and Advanced achievement levels.

THE NAEP ACHIEVEMENT LEVELS OPINION SURVEY

After reviewing the background information, respondents selected the civics opinion survey and public comment form. This gave them access to the following information:

- The achievement levels descriptions for civics,
- Examples of actual student responses to the Civics NAEP,
- Scoring guides for each exemplar item, and
- The actual opinion survey respondents were to complete.

In addition to the achievement levels descriptions, participants reviewed items representing performance at each level. Three items were used to illustrate student performance on the Civics NAEP at each achievement level for each grade. Two items were multiple choice and one was constructed response, either short (maximum score 3) or extended (maximum score 4). At each grade one constructed response item was selected to illustrate performance at two different achievement levels. In this way differences in student performance on the same exercise at different score points could be observed and related to the different levels of achievement associated with each.

RESULTS OF THE NAEP ACHIEVEMENT LEVELS OPINION SURVEY

Sixteen persons replied to the civics survey. The small number of respondents was disappointing. A second request to recruit informed individuals was sent to the original professional organizations in mid-April, 1999 and resulted in a few additional responses.

Respondents (16) classified themselves as classroom teachers (6), nonteacher educators (8), or members of the general public (2). Respondents were asked to review the achievement level descriptions for all grades, giving special attention to the grade(s) for which they felt most confident in making judgments. Many individuals felt confident to evaluate more than one grade level. Eight people gave evaluations for grade 4, 9 for grade 8, and 16 for grade 12. Overall, the respondents tended to agree that the achievement levels descriptions were clear and easy to understand. All respondents thought the descriptions were *at least somewhat clear and easy to understand.* The majority of respondents indicated that they thought it was useful to have student performances reported in terms of achievement levels. Only one participant responded that it was *less than somewhat useful* to have student performance on the NAEP reported in terms of achievement levels. The majority of respondents indicated that they thought that the achievement levels generally reflected what students should know and be able to do; only two respondents did not agree. Further, when asked whether the examples of items used in reporting achievement levels performances represent the kinds of things students should know and be able to do according to the achievement levels descriptions, only one respondent said "no."

The data in Table 18 report the majority of civics respondents indicated that they thought the achievement levels were "about right." Only one or two participants thought the levels for 4th and 8th grades were either "too high" or "too low." Two respondents thought that 12th grade levels were "too low" for Proficient and three thought they were "too low" for Basic and Advanced. One respondent thought that 12th grade levels were "too high" for Basic and Proficient and two thought they were "too high" for Advanced.

Table 18
Respondents' Judgments about the Reasonableness of the NAEP Achievement Levels for Each Grade

	Basic	Proficient	Advanced
Grade 4 (n=8)			
About right	5	4	4
Too high	1	1	1
Too low	2	1	1
Grade 8 (n=9)			
About right	6	6	6
Too high	2	1	1
Too low	1	1	1
Grade 12 (n=15)			
About right	11	13	8
Too high	1	1	2
Too low	3	2	3

NAGB Approval of Achievement Levels-Setting Process Outcomes

The ACT Project Director made presentations to NAGB's Achievement Levels Committee throughout the process of design, research, and implementation of the 1998 Civics NAEP ALS process. In addition, the technical advisory committees closely monitored the process and outcomes at each stage. After careful review of all data analyses regarding the entire process, TACSS recommended adoption of the outcomes. On May 1, 1999, the NAGB Achievement Levels Committee approved the achievement levels descriptions and cutscores recommended by ACT. Although they requested some paper substitutions for the exemplar items, they gave general approval to the exemplar item selection process and the items selected by panelists. During their regularly scheduled meeting in May 1999, NAGB approved the 1998 Civics NAEP Achievement Levels, as recommended.

Summary

Achievement levels have become an important component of the National Assessment of Educational Progress. ACT has conducted a rigorous, long-term research program to study the NAEP Achievement Levels-Setting process. Years of work have led to the current process, which has been considerably enhanced by the most recent refinements. The most notable were the finalization of the ALDs prior to convening the ALS panels, the introduction of the Reckase Charts, and the provision of consequences data during the rating process. This comprehensive evaluation of the outcomes of the process revealed remarkable consistency, agreement, and overall satisfaction at every stage. Given that the NAEP standard setting process is based on judgments by broadly representative panels of individuals, such consistency is an impressive accomplishment.

REFERENCES

ACT (1997a). *Setting achievement levels on the 1996 NAEP in science: Final report, Volume IV: Validity evidence special studies*. Iowa City, IA: Author.

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

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

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

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. & Hanick, P.L. (2000a). *Developing achievement levels on the 1998 National Assessment of Educational Progress in civics: Pilot study final report*. Iowa City, IA: ACT.

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

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

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

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

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

National Assessment Governing Board (1998). *Civics Framework for the 1998 National Assessment of Educational Progress*. Washington, DC: Author.

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

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

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

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