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

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

December 2000

Developing Achievement Levels for the 1998 NAEP in Writing: 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 Writing 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 Writing 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 Writing NAEP achievement levels and provides evidence for the procedural validity of the writing ALS process. In addition, it provides an overview of field trials and the pilot study conducted to determine and refine the design for the 1998 writing 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 758 nominators were contacted to participate in the process of identifying Writing ALS panelists, and 422 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 88 panelists participated in the ALS, and they represented 34 states and Guam.

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

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 Writing 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
- 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 performances selected;
- patterns of results consistent with previous studies; and
- absence of extreme reactions to consequences data.

The process implemented in the Writing 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 writing. 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 Writing 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. Those concerns appear to have been unfounded, however. 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 Writing 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.

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

- ✓ *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.*

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.

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. 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.

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

- ✓ *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 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. 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. Responses to the process evaluation questionnaires supported this interpretation.

- ✓ *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.*

- ✓ *None of the judges adjusted ratings to be identical to IRT-based performance estimates.*

- ✓ *Panelists formed judgments that were not exactly the same as the IRT-based estimates of student performance, and this lends credibility to the outcomes of the process.*

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. Neither of these extremes was found for the Writing ALS.

- ✓ *The writing 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 RATINGS FOR DIFFERENT TYPES OF WRITING PROMPTS

ACT was very interested in determining whether there was any evidence that the ratings differed significantly for different types of prompts. Panelists were instructed to examine their Reckase Charts for evidence of such practices. Items were grouped on the Reckase Charts according to the type of writing. Panelists could easily examine the charts to determine whether there were patterns of differences in their ratings for items of different types. They could determine whether ratings for narrative prompts, for example, were typically higher than their own cutscores while ratings for informative or persuasive prompts were typically below. While individual panelists may have discerned such patterns, no significant differences were revealed at the grade level.

- ✓ *More research is needed to determine how judges perceive differences in prompts for different types of writing. 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.*

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 Writing 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 writing 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 borderline performances 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 Writing 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 Writing: 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

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

openly and frankly through extensive research conducted prior to the 1998 Writing NAEP 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. 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 (Reckase, 2000). 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 Writing achievement levels. The report also describes the process used by ACT to collect public opinion and comments evaluating the reasonableness and usefulness of the Writing NAEP achievement levels that resulted from the achievement levels-setting process.

RESEARCH CONDUCTED PRIOR TO THE WRITING NAEP ALS

ACT carried out two field trials and one pilot study each for the 1998 Writing and Civics NAEP. All six of those studies and the operational Civics ALS process were completed and reviewed by ACT's technical advisory committees prior to convening the panels for the 1998 Writing 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 & Yang, 2000; Loomis, Hanick, Bay & Crouse, 2000).

In addition to these studies that were conducted for both civics and writing, additional studies were designed for writing to help ACT better understand the 1998 Writing NAEP and design an ALS process that would be successful. A Performance Profiles study was conducted by ACT in January 1998 to gain more information about how panelists evaluate and form judgments about writing performance (Bay, 2000[4]). ACT wanted to know whether panelists were more likely to use a compensatory or conjunctive strategy (Jaeger, 1995) to make judgments about performances involving the three types of writing assessed in NAEP. The results of that study suggested that panelists were unlikely to use a compensatory model for judging student performance in an achievement levels-setting process. This finding was consistent with findings by ACT from studies involving holistic evaluations of student performances (ACT, 1995; 1997b).

In addition, ACT studied the writing frameworks of states to compare expectations in those documents to those in the NAEP preliminary ALDs for writing[5]. The 1998 Writing NAEP was

[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.

[4] A slightly modified version of this report appears in Loomis (Ed.) (2000).

[5] See Bolton, Hanick, Cook, Welch, & Loomis (2000) for a complete report on this study.

administered as a State assessment at grade 8, and that made it particularly important to ascertain the correspondence between NAEP requirements and those in states. The results of the study suggested that the NAEP writing requirements at the Proficient level were similar to those of state documents examined in the study. Given that finding, it seemed feasible to proceed with plans to finalize the preliminary ALDs for writing.

WRITING FIELD TRIAL #1

The purpose of the first writing field trial was to evaluate the Item Score String Estimation (ISSE) rating method relative to the Mean Estimation (ME) rating method used by ACT in the 1994 and 1996 NAEP ALS procedures. To set cutscores on NAEP, ACT has always used an item-by-item rating method requiring judges to estimate the performance of students at the borderline of each achievement level. ACT proposed to study the ISSE method as 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.

Results of the first field trial in writing indicated that panelists were able to use the method without difficulty. The ISSE cutpoints and their standard deviations appeared to be relatively reasonable when compared with those produced by ratings of the same items with the Mean Estimation method. The panelists expressed satisfaction with and confidence in the ISSE method and the outcomes of the process. The ISSE procedures were implemented with ease. Compared to the cutscores computed from ratings of the same items with the Mean Estimation method, however, the ISSE method resulted in higher cutscores for the Proficient and Advanced levels. The cutscores for the Basic level were lower for the ISSE method. The ISSE cutscores resulted in lower percentages of students performing at or above the Proficient and Advanced levels. Further research showed the ISSE method to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true" scores or "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 Writing ALS.[6]

WRITING FIELD TRIAL #2

The purpose of the second field trial was to identify the procedures that would be used for the 1998 pilot study and ALS process. ACT's goal was to complete the research phase prior to the

[6] For more detailed information about the writing field trials, please refer to Loomis, Hanick, Bay & Crouse. (2000) and Loomis, Bay, Yang & Hanick (1999).

.

pilot study, and the second field trial was the final opportunity to conduct research with panelists before the pilot study. The ISSE method had been eliminated from further consideration, and no final decision had been made regarding the rating method to use in the 1998 ALS process.

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

Results of the second field trial indicated that panelists had little difficulty with either the Booklet Classification method or the Reckase method. There was no statistically significant difference between cutscores set by panelists using the Booklet Classification method who were informed of the consequences of their classifications after the first round and cutscores set by panelists who were not informed until after the last round of classifications. On the other hand, panelists using the Reckase method and receiving consequences data throughout the process, generally set higher cutscores. Because the cutpoints for the Booklet Classification method were considered to be unstable, the Reckase method was judged to be more promising for use in the Writing Pilot Study and ALS.

WRITING PILOT STUDY

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

The pilot studies for the 1998 ALS process had been planned as a "dry run" for the operational ALS to determine whether modifications to training, instructions, timing, and so forth were needed. The writing pilot study (PS) was an opportunity to continue studying and refining the procedures for the ALS. The achievement levels-setting process implemented for the Writing PS incorporated the Reckase Charts as a type of feedback information. ACT was particularly interested in evaluating writing panelists' reactions to the ALS process that included the Reckase Charts. The pilot study was implemented to provide a final check of whether further modifications could be identified to make the operational Writing NAEP ALS more successful

Throughout the pilot study, ACT collected useful 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. Civics pilot study panelists had strongly recommended that ACT develop a method for electronically marking each panelist's ratings on his/her Reckase Charts. A fast and efficient method for producing these individualized charts was developed and implemented in the writing pilot study. Relatively minor but important adjustments were also made to increase the amount of feedback provided to panelists 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 only been observed in data generated from the civics pilot study. As had been the case in the second field trial for writing, 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 with respect to 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.[7]

DEVELOPING FINAL VERSIONS OF THE WRITING ACHIEVEMENT LEVELS DESCRIPTIONS

Preliminary achievement levels descriptions were developed as part of the process for developing 1998 Writing NAEP Framework (NAGB, 1998). The preliminary achievement levels were reviewed extensively, revised and finalized prior to the convening the ALS panels for writing.[8] The process of transforming the preliminary ALDs into final ALDs involved four steps:

- Revising the preliminary ALDs to reflect developmental changes across levels within each grade and across grades within each level;[9]
- Convening focus groups to review the preliminary ALDs and make recommendations to improve them;
- 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. A focus group was convened in each of four geographic regions and participants were selected to represent the three categories of persons (i.e., teachers, nonteacher educators, and general public) who serve on achievement levels-setting panels for each of the three grades assessed in NAEP. The purpose of the review was to determine whether the descriptions of achievement in writing 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 Writing NAEP framework.

- The ALD statements in general described achievement that was too high for first draft writing.
- The description of Basic achievement in particular was too high for 4^{th} and 8^{th} grades, but was about right for 12^{th} grade.
- The description of Advanced achievement for 12^{th} grade should be higher.

---

[7] For more detailed information about the writing pilot study, see Loomis, Hanick & Yang (2000).
[8] For a complete account of the process used to finalize the 1998 NAEP ALDs, please refer to Loomis & Hanick (2000).
[9] The Technical Advisory Committee on Standard Setting recommended that this change be implemented prior to distribution of descriptions to focus group panels.

- The statements should describe a clearer progression of writing skill development across levels and grades.
- The descriptions should be stated in simpler, clearer language.
- The descriptions should be more specific and use examples.

EXPERT REVIEW PANEL

Writing content experts reviewed the work of the focus groups and modified the preliminary ALDs according to the recommendations and their own expertise regarding the Writing NAEP. All of the recommendations from the focus groups were reviewed and discussed thoroughly before the panel reached any decision. The preliminary ALDs for writing went through a considerable transformation during this revision process. Descriptors were shifted, deleted, written and rewritten. Although the preliminary ALDs were changed substantially, the revised descriptions remained true to the *Writing Framework*. While the expert panel initially felt that the ALDs would require separate descriptive statements for each type of writing (narrative, informative, and persuasive) they determined during the review process that this would not be necessary. One important decision reached by this group was that no description of *voice* would be included in the revised ALDs.

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 write descriptions that used simple, concise language and to incorporate examples into the statements. Great care was taken to produce statements that described a clear progression of writing skill development across levels and grades. Overall, the level of difficulty represented by the ALDs was lowered during the finalizing process. However, the level of difficulty for 12th grade Advanced was raised.

An additional concern had been raised regarding the use of the same terms in both the scoring rubrics and ALDs to describe expected performance. It was important to assure that assigning a score to a student response was not also denoting a level of achievement. Content staff at ETS, the operations contractors for NAEP item development, modified the scoring rubrics to avoid this problematic situation.

COLLECTING INFORMED OPINIONS REGARDING THE RECOMMENDED ALDS FOR WRITING

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 Writing NAEP. ACT conducted a telephone survey of focus group members. Results indicated that the revised ALDs were generally well received. Thirty-three of 39 respondents (84.6%) judged the ALDs to be reasonable. Five respondents indicated that the Basic descriptions did not reflect "partial mastery," 2 indicated that the Proficient descriptions did not reflect "solid academic performance," and two indicated that the Advanced descriptions did not reflect "superior performance." Focus group members also were asked to evaluate the descriptions for 8th grade Basic and Proficient, relative to one another. Specifically, they were asked whether or not there appeared to be a "gap" in the progression of skill

development described by the two levels. Of the 29 respondents to this question, 10 (34.5%) said that there was a "gap." Basic seemed too low and Proficient too high for grade eight.

Members of the NAEP Writing Standing Committee, Item Development Committee, and Framework Committee were also asked to comment on the revised ALDs. Of the 7 members who responded to the request, 5 recommended adoption without further substantive changes. One or two reviewers indicated that the descriptions were too high for all levels except Basic for 4th and 8th grades. One member indicated that the description for Basic was too low for 8th grade. The Expert Review Panel evaluated those recommended changes and determined none seemed suggestive of significant, substantive changes in the recommended writing ALDs.

ADOPTING THE REVISED ALDS FOR THE 1998 WRITING NAEP

The final version of the writing 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 Writing NAEP Achievement Levels, in May 1999.

ALS PANELIST SELECTION PROCESS

The following summary highlights the main features of each step in the panelist selection process. 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 258 public districts and 40 private schools were sampled. Please see Table 1 for the distribution of the samples by type of panelist to be nominated. In addition, 21 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 was 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, 1997a).

Table 1
Distribution of School Districts Sampled for Nominating Panelists to the Writing NAEP ALS

Nominator Type	Public Districts	Private Districts	Total
Teacher	128	34	162
Nonteacher Educator	15	6	21
General Public	115	-	115
Total	258 (87%)	40 (13%)	298

NOMINATORS OF CANDIDATES FOR ALS PANELS

ALS nominators were identified by drawing three separate samples of districts without replacement.[10] 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 758 nominators were contacted. Please see Table 2 for the distribution of nominators. Nominators were persons holding a specific title or position, such as the following.

Nominators of teachers were:

- district superintendents
- leaders of teacher organizations
- state curriculum directors
- 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 writing-related position or with a writing-related background

Table 2
Distribution of Nominators Contacted for the Writing NAEP ALS

Nominator Type	Public Districts	Private Districts	State	College/ Universities	Employers	Total
Teacher	247	33	9	-	-	289
Nonteacher	15	5	13	27	-	60
General Public	315	-	-	-	94	409
Total	577 (76%)	38 (5%)	22 (3%)	27 (4%)	94 (12%)	758

[10] The districts were sampled from a data file produced by Market Data Retrieval for 1997. More details of the sampling procedure are available in Chen & Loomis (2000).

POOL OF PANELIST 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). All nominees had been judged by nominators to be "outstanding" in their writing-related field. Each nominator could nominate up to four candidates for each grade. Of the 758 nominators identified, 422 candidates were nominated. Please see Appendix B for the distribution of the nominee pool.

CHOOSING ALS 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. Although an ample number of candidates were nominated, there were not enough different nominators to draw panels according to plans. Only one candidate per grade level is selected from the list submitted by any one district-level nominator. There was a shortage of teacher nominees to select for the Grade 12 panel without having more than one panelist from the same district. Further, there were very few males nominated to the panels, particularly for Grade 4. Despite the low response rate from nominators and the rather uneven representation of nominees in the various targeted categories, the panels were reasonably representative in the overall counts. A total of 88 panelists (grade 4 = 29; grade 8 = 30; grade 12 = 29) participated in the ALS study representing 34 states and Guam. A list of the panelists who participated in the ALS is presented in Appendix B.

THE ACHIEVEMENT LEVELS-SETTING PROCESS FOR THE 1998 WRITING NAEP

OVERVIEW

The purpose of the Writing ALS was to produce a set of recommendations for NAGB to consider in establishing achievement levels on the 1998 NAEP in writing. The recommendations would include a set of cutscores on the Writing NAEP to report student performance classified as Basic,

Proficient, and Advanced achievement. Further, the recommendations would include a set of exemplar items from the Writing NAEP to illustrate student performance at Basic, Proficient, and Advanced levels of achievement in accordance with 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 writing ALS lasted five days, December 9-13, 1998 (Wednesday-Sunday). 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 writing pilot study and were experienced in the process.[11]

ALS SESSION FORMATS AND GROUP 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.

Each grade-level panel was led by two facilitators: one process facilitator and one content facilitator. 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. In addition, they reviewed scorer training materials and observed one day of scorer training at the NAEP scoring contractor's (NCS) Iowa City facility. Content facilitators led the discussions of the *1998 Writing NAEP Framework* and achievement levels descriptions, and answered "content" questions. All content facilitators had participated in developing the Writing NAEP 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.

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.

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

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

WRITING 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.

WRITING ITEM RATING POOLS

The 1998 NAEP Writing data were used for the Writing 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 and item type. Detailed information about the item pools has been presented in Appendix D. Table 3 presents a summary of information describing items in the rating pool for each rating group. The design, including two item rating groups and two item rating pools, 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
Description of Items in Each Item Rating Pool for Each Item Rating Group for Writing NAEP ALS

Grade Group	Percent at Each Rubric Score Point *						# Prompts of Each Type			Summary Statistics	
	%1	%2	%3	%4	%5	%6	Narr	Infor	Persu	Mean	SD
4A	2.66	9.12	30.38	37.02	11.26	2.39	3.63	3.46	3.59	3.56	0.21
4B	3.13	9.70	31.10	35.12	11.81	2.86	3.62	3.45	3.56	3.55	0.20
8A	2.59	10.25	29.18	40.27	12.35	3.17	3.68	3.60	3.53	3.60	0.15
8B	3.06	9.97	29.17	39.22	13.28	3.00	3.70	3.58	3.53	3.60	0.15
12A	3.03	7.59	19.03	41.31	22.53	3.63	3.98	3.96	3.71	3.86	0.23
12B	2.83	7.98	19.98	40.39	21.31	4.79	3.97	3.97	3.71	3.86	0.24

* %n = percentage of student responses scored 1, 2, 3, etc.

The 1998 Writing NAEP consisted of 20 different writing prompts for each grade. Each prompt represented a block or section in the assessment booklets: one block or section contained one prompt. Each student was assigned two prompts, and responses to each prompt had to be completed within 25 minutes.

Each grade 4 rating group rated 12 blocks of items (12 prompts): 5 narrative prompts, 4 informative and 3 persuasive. Eight prompts in each rating pool were unique to each rating group, and four prompts were in common with the rating pool of the other rating group for grade 4. Group A rated prompts from 12 of the 20 blocks; group B rated items from 12 of the 20 blocks; and both groups rated the same prompts for 4 of the 12 blocks.

Each grade 8 rating group rated 12 blocks of items (12 prompts): 4 narrative prompts, 4 informative and 4 persuasive. Eight prompts in each rating pool were unique to each rating group, and four prompts were in common with the rating pool of the other grade 8 rating group. Group A rated prompts from 12 of the 20 blocks; group B rated items from 12 of the 20 blocks; and both groups rated the same prompts for 4 of the 12 blocks.

Each grade 12 rating group rated 12 blocks of items (12 prompts): 3 narrative prompts, 4 informative and 5 persuasive. Eight prompts in each rating pool were unique to each rating group, and four prompts were in common with the rating pool of the other grade 12 rating group. Group A rated prompts from 12 of the 20 blocks; group B rated items from 12 of the 20 blocks; and both groups rated the same prompts for 4 of the 12 blocks.

STEP 1: BRIEFING MATERIALS FOR ALS PANELISTS

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 November 10, 1998 and contained materials that panelists were required to study. The second mailing was November 25, 1998 and contained detailed instructions related to travel arrangements and accommodations. The briefing materials and information included:

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

STEP 2: GENERAL ORIENTATION AND TRAINING EXERCISES FOR ALS PANELISTS

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

start of the second full day, an overview of the process outcomes and how to produce them was presented to help panelists understand 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 Booklets* 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 projected on screens for 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 Booklets for panelists, make it relatively easy for panelists to identify each procedure, to follow instructions for each task, and to understand how each one fits into the overall ALS process.

TAKING A FORM OF THE WRITING 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.[12] As usual, most panelists found this to be a very informative and useful exercise.

UNDERSTANDING THE WRITING ACHIEVEMENT LEVELS DESCRIPTIONS

During a general session the morning of Day 2, the three content facilitators presented an overview of the *NAEP Writing Framework* and the ALDs as a general session. 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 understanding of the content and organization of the *Writing Framework* and a clear understanding of how the ALDs were related to both the framework and to the NAGB policy definitions.

[12] 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."

The content facilitators spent about an hour discussing the framework and development of achievement levels descriptions (ALDs). One content facilitator discussed the components of the framework. The second discussed item development, rubrics, and scoring procedures. The third content facilitator discussed the development of the achievement levels descriptions and explained how the focus groups had informed the modifications. This content facilitator provided a clear rationale for why the ALDs are different from the scoring rubrics, why the levels are not set by the scoring rubrics, and why all element of the ALS process are necessary.

In grade-level sessions, content facilitators guided the panelists through an extensive training session focused specifically on the achievement levels descriptions for their grade. Panelists were led in an evaluation of the ALDs to compare performance across levels in their grade and to compare performance across each grade within each level. Panelists discussed the ALDs and participated in several training exercises to help them understand the descriptions. They were lead through an explanation of the scoring guide and the description for each score point in the generic rubric for each type of prompt: narrative, informative, and persuasive. This exercise was designed to help panelists become familiar with prompts of different types and to understand how the ALDs relate to all types of prompts. The exercise also helped panelists to become familiar with prompts 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 prompt vs. one test booklet containing two prompts). A sample of ten student papers was given to judges to review and discuss with respect to their understanding of the ALDs. The particular writing exercise was the first one in the NAEP booklet form that was administered to the panelists on the first day of the ALS session. Panelists were asked to determine if the performance exhibited in the paper should be classified as Basic, Proficient Advanced, or below Basic. After classifying each paper independently, panelists discussed their classifications with each other.

Next panelists looked at whole booklets produced by the same students who had written the first paper. Panelists classified the booklet performance into the same four categories based on the ALDs. Their paper classifications from the previous exercise had been collected and were unavailable when they classified the complete student booklets. Following this task, panelists received their earlier classifications of single papers to compare with their classifications of performance on both writing exercises. They discussed the differences between performance on a single writing task and performance on a whole booklet that required writing for two prompts. This exercise helped panelists gain a better understanding of the ALDs and become familiar with additional NAEP items and scoring rubrics. It also caused panelists to confront the fact that a student's performance on each prompt could differ sharply.

UNDERSTANDING BORDERLINE PERFORMANCE

After working with the ALDs throughout the morning and early afternoon, panelists were again convened in general session for training in the concept of borderline performance[13] and instruction 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 that 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. Panelists were given copies of the ALDs where each statement was printed in a separate cell. This format highlighted the descriptions of the various attributes of writing so that panelists could examine the information across the achievement levels for each grade. Each grade group had drafted a set of borderline descriptions by the close of Day 2. Content facilitators evaluated those lists across all grade levels to make certain that they appeared to be appropriately calibrated descriptions of borderline performance—not too low and not too high, relative to the ALDs.

Those lists were distributed to panelists at the start of Day 3 for review and modification. Panelists were informed that there would be opportunities for further review and modification of the borderline descriptions. They were aware that the first round of item ratings would begin later that day (Day 3), and the goal was to make certain that they had a useful set of descriptions to use by that time. A review of borderline descriptions was scheduled just prior to the training session for Round 1 ratings. As a means of keeping panelists focused on the ALDs, they evaluated borderline descriptions throughout the process. Borderline descriptions were evaluated and modified, as needed, until panelists were ready to begin the final round of item-by-item ratings on the last day.

PAPER CLASSIFICATION EXERCISE

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

Panelists were instructed to first sort papers into four categories: below Basic, Basic, Proficient, and Advanced. They then were to look at the papers within each category and determine whether any of those represented performance at the borderline of the level. If no paper represented

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

borderline performance, then no paper was classified as borderline performance. For each of the three prompts, panelists classified each paper, discussed their classifications as a group, and modified their classifications if they chose to do so. Panelists classified and discussed 54 student papers and 7 achievement categories[14]. After selecting papers to represent borderline performance at each achievement level, panelists could refer to a sheet where the score for each paper was recorded. The basis for selection, however, was to be their understanding of the ALDs and borderline performance and not the paper score. Their classifications were recorded on a form and tabulated for use in the process of selecting exemplar performances on Day 5.

This training activity was designed to accomplish the following purposes:

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

STEP 3: THE ITEM RATING PROCESS AND FEEDBACK

The general procedure followed for the item rating process included instruction in a general session involving all panelists to assure that they were given the same information. Process facilitators reviewed the instructions and answered questions from panelists in the grade-level sessions. 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) was used. The rating method had been described in the orientation sessions of the first two days, and a demonstration had been given on Day 2. The procedure was reviewed in detail, and panelists were instructed in marking their rating forms. For rating the prompts, judges estimated the mean or average score (e.g., 2.4 on a scale of 1-6) of students performing at the borderline of each level.[15] They were told to think of a class with 100 borderline students for each achievement level and estimate the average score for those students on each prompt. Once trained, the panelists were ready for Round 1 ratings. During the rating process, panelists could refer to student papers scored at each rubric point. Each panelist was given a packet of three papers scored at each rubric score point for all prompts in the rating pool. Scores appeared on the papers for prompts that were not included in the paper classification exercise.

Panelists were told to read each item carefully, compose a mental response to the item, and refer to the scoring rubric. This procedure would help panelists form a clear concept of what was required of students. For each item in their item rating pool, panelists marked their estimate of

[14] The seven categories are: Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient, Borderline Advanced and Advanced.

[15] Details on computing averages were included in the instructions as recommended by panelists in the Writing Pilot Study De-briefing Session.

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. Each panelist rated 12 prompts.

FEEDBACK AFTER ROUND 1

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

Cutpoints

The cutpoints are the combined ratings over 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 were presented on the ACT NAEP-like scale which is a linear transformation of the NAEP score scale. Data were reported on the ACT NAEP-like scale for security reasons and to decrease the potential impact that achievement level data from other NAEP subjects could have on panelists in the Writing ALS.

Standard Deviation

The standard deviation is the indicator of the level of variability around each cutscore for a grade. The standard deviation reported to panelists was computed on the basic of individual raters' cutscores for each achievement level.

Whole Booklet Feedback

Whole booklet feedback was produced for the set of items in the NAEP exam booklet that had been administered to panelists as part of the orientation process on Day 1. Each rating group (A and B) had a different assessment form. The whole booklet feedback reported the percent of total possible points that a student needed to earn in an assessment booklet in order to meet the minimal requirements for performance at each achievement level. That is, it is the percentage of total possible points at a cutscore. For example, the whole booklet feedback report might state: "Based on cutscore for your grade, students performing at the borderline Advanced level are expected to get 98% of the total possible score points for this booklet." A similar statement was given for each achievement level. This feedback was based on the cutpoints the grade group had set during the first round of ratings, and was updated after subsequent rounds of ratings. Panelists

were informed of the reasons that would cause the percentages to differ for the two booklet forms, i.e., different item combinations resulting in different performance and total points possible.

Whole Booklet Exercise

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

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

Panelists were given up to 4 booklets to review as representative of borderline performance at each achievement level. One hundred randomly selected booklets for each of the forms used in the exercise were available for use as feedback. Only the assessment sections were copied for use in the exercise. There was a relatively high probability that no booklets would be available on site for feedback at an achievement level for a rating group. If no booklets were available that had a score within 4% of the total possible points associated with the cutscore, then no booklets were presented to panelists for that achievement level. Panelists were given a complete explanation of the source of booklets and the reason for which no booklet was available.

Rater Location Feedback Charts

The rater location feedback charts are histograms. The horizontal axis represents scores on the ACT NAEP-like scale, and the vertical axis represents the number of raters. Letter codes that identified individual raters are positioned along the ACT NAEP-like scale at the point where each panelist set his/her cutscores based on his/her individual ratings. Letter codes are used so the cutscores for each panelist could remain confidential. (In fact, most panelists openly and freely discussed their rater location data.) The graphs indicate the cutscores that resulted from the item ratings by each panelist for Basic, Proficient, and Advanced levels, and the relationship of the

panelists' ratings to each other (interjudge consistency). There was one chart for each achievement level within each grade.

Facilitators examined the charts to identify panelists who were "outliers" or panelists whose patterns of cutscores across levels indicated potential problems with the item rating process. For example, rater location charts were examined to determine whether a panelist tended to set very high or very low cutscores for all levels relative to other panelists in the grade group and whether a panelist set cutscores that were very close together or very far apart. Facilitators made a specific point of discussing such findings with the panelists to make certain they understood the implications of such patterns and how to change them through subsequent ratings, if the panelist so desired.

Student Performance Data

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

Reckase Charts

The meeting director introduced the Reckase Chart to panelists as a type of feedback information. She explained the features of the chart with the aid of a special computer graphics presentation. The computerized show displayed the entire chart and had the capability to "zoom in" during instruction to highlight its features. The demonstration used colored markings to indicate individual cutscores, grade-level cutscores, and item ratings on the chart. The presentation aided panelists when they evaluated their own charts and interpreted the information displayed on the charts.

Panelists receive a paper copy of the Reckase chart that presents information for each prompt in their 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 for each prompt in the item rating pool. Each column represents the range of IRT-based performance estimates for one assessment item. Each row represents IRT-based performance estimates for the prompts for students scoring a specific point on the ACT NAEP-like scale. The ACT NAEP-like scale scores range from the score associated with the lowest asymptote value for any prompt in the grade-level item pool to the value associated with the highest asymptote. On the ACT NAEP-like scale score, the score range for grade 4 is 88 to 220; for grade 8, the range is 93 to 206; and for grade 12, the range is 93 to 208. The expected score (mean) is reported for each prompt at each scale score. The expected performance across scale score points can be observed for each prompt, as can the expected performance across prompts for students scoring at a particular scale score. A sample Reckase Chart and instructions have been included in Appendix G. Please note that only

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

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

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

ROUND 2 RATINGS

Panelists studied and discussed the feedback information from Round 1. To prepare for Round 2, they spent about an hour to review the ALDs and modify 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 datafiles 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—new feedback added after Round 2—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.[16]

Feedback information was distributed to panelists in grade groups. Panelists evaluated the Reckase Charts a second time and marked their charts with both the grade level cutscore and their own cutscore for each achievement level. Grade-level consequences data were added and the whole booklet exercise was omitted. Otherwise, the same types of feedback as were distributed

[16] 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.

after Round 1 were presented to panelists after Round 2. (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

After examining and evaluating their round 2 ratings relative to consequences data, Reckase Charts and the other forms of feedback, 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. Panelists were allowed to discuss ratings for specific items with other panelist in their table group. Round 3 ratings were completed in a very short amount of time.

FEEDBACK AFTER ROUND 3

Round 3 item ratings were again entered into datafiles for computations and analyses. Feedback data were produced for panelists, based on Round 3 ratings. Reckase Charts were not generated for Round 3 feedback, but other feedback was updated and distributed. Grade-level and individual-level consequences data were presented to inform the panelists about their own cutscores.

Feedback data from Round 3 ratings were distributed to panelists in 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 for 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. Please see Appendix F for an example of the consequences feedback data.

STEP 4: MAKE RECOMMENDATIONS FOR FINAL CUTPOINTS

Panelists were given a few minutes to review the consequences data before they received a consequences data questionnaire. A sample questionnaire has been included in Appendix H. 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. They were asked to consider the data and not to discuss it

with others. Panelists could recommend a different cutscore to represent each achievement level for any or all three cutscores. The individual Round 3 cutscores were used to compute the final grade-level cutscores for panel members who recommended no changes to their cutpoints. Panelists were fully informed that these would be the final cutpoints, and they would be used as the standard for selecting exemplar items.

STEP 5: SELECTING EXEMPLAR WRITING PERFORMANCES

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

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

The average conditional probability of a specific response score served as the indicator of item difficulty. To be on the list of exemplars, a response score had at least a 50% average probability across the score interval of an achievement level. Each rubric score point was evaluated as if it were an item, so prompts could appear on the list five times (once for each credited response). Response scores were “assigned” to the list for the lowest achievement level for which this criterion was met. If the criterion were met below the Basic level, that score was eliminated from consideration. Please see Chen & Loomis (2000) for more information on the procedures used to produce the item lists for panelists to use in selecting exemplar performances.

Panelists determined whether or not each response score that qualified as an exemplar would serve as a good illustration of performance required at the specific achievement level, based on the achievement levels descriptions. They identified papers that matched the descriptions of student performance at each achievement level and satisfied the statistical criteria that qualified the paper score as an exemplar. They “approved” or “vetoed” each paper. The number of exemplars selected for each achievement level ranged from 3-4 papers. For some types of writing, no exemplar was selected for a particular achievement level. The lists of exemplar performances for each grade have been included in Appendix I. 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 WRITING 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 inform NAGB of panelists' opinions regarding the final cutpoints and the consequences associated with them.[17] When the panelists completed the final questionnaire, they were thanked for their work and the meeting was adjourned.

OUTCOMES OF THE WRITING NAEP ACHIEVEMENT LEVELS-SETTING STUDY

The Writing ALS was designed and implemented to produce a set of recommendation for NAGB to consider in establishing achievement levels on the 1998 Writing NAEP. The recommendations included a set of cutscores on the Writing 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 Writing 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 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:

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

- Cutpoints and their standard deviations
- 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 and their standard deviations have been reported in Table 4. Please refer to Chen & Loomis (2000) for a description of computational procedures used to produce feedback data for the ALS process. Some cutscores were raised, some were lowered, and some remained unchanged from one round to the next. Thirteen cutpoints were raised, 8 were lowered, and 6 remained unchanged from one round to the next in the writing ALS. From Round 1 to the final levels, cutscores were raised at 6 levels and lowered at the remaining 3. The average net change in the 6 cutscores that were raised was 1.20, and the average net change in the 3 cutscores that were lowered was 1.23.

The standard deviations *decreased* from round to round for all levels. Typically, the variability in the Basic cutscore is highest. 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 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. In the writing ALS, however, this pattern was found for grade 8 only. For both grades 4 and 12, the standard deviation associated with the Advanced cutscore was highest for each round of ratings.

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 J.

Table 4
1998 Writing 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 n=29	Basic	Cutpoint	137.6	138.7	139.2	139.5
		SD	5.4	4.2	3.8	3.4
		%≥	**87.8**	**86.0**	85.3	84.6
	Proficient	Cutpoint	163.1	164.9	164.9	164.9
		SD	5.2	3.9	3.4	3.2
		%≥	**27.7**	**23.1**	23.1	23.1
	Advanced	Cutpoint	185.6	186.8	185.6	184.8
		SD	5.4	4.6	4.4	4.0
		%≥	**1.2**	**1.0**	1.2	1.4
8 n=30	Basic	Cutpoint	138.5	139.7	139.7	139.7
		SD	6.5	3.6	3.2	3.0
		%≥	**86.2**	**84.4**	84.4	84.4
	Proficient	Cutpoint	163.6	164.0	163.8	163.7
		SD	5.7	2.5	2.2	2.1
		%≥	**27.5**	**26.6**	27.5	27.5
	Advanced	Cutpoint	185.3	185.2	184.9	184.9
		SD	4.2	2.3	2.2	2.2
		%≥	**1.0**	**1.1**	1.1	1.1
12 n=29	Basic	Cutpoint	141.8	142.6	142.8	143.1
		SD	6.4	3.5	3.4	3.3
		%≥	**81.5**	**80.1**	80.1	79.2
	Proficient	Cutpoint	164.9	165.6	165.8	165.8
		SD	6.5	3.2	2.7	2.4
		%≥	**24.3**	**22.5**	22.5	22.5
	Advanced	Cutpoint	189.3	189.7	187.7	186.8
		SD	8.1	4.5	4.7	4.1
		%≥	**0.6**	**0.4**	0.8	1.0

Bold font represents data that were not presented to panelists.

THE ISSUE OF DIFFERENCES BETWEEN RATINGS FOR DIFFERENT TYPES OF WRITING PROMPTS

ACT conducted an ALS process for the 1992 Writing NAEP, and NAGB decided not to set achievement levels until some additional adjustments had been made to the frameworks, test specifications, item pool and scoring rubrics. Analyses of the 1992 results showed that the cutscores for different types of writing would have been significantly different.

The decision was made to scale the three types of writing on a single, unidimensional scale. ACT was very interested in determining whether there was any evidence that the ratings differed significantly for different types of prompts. Panelists were instructed to examine their Reckase Charts for evidence of such practices. Items were grouped on the Reckase Charts according to the type of writing. Panelists could easily examine the charts to determine whether there were patterns of differences in their ratings for items of different types. They could determine whether ratings for narrative prompts, for example, were typically higher than their own cutscores while ratings for informative or persuasive prompts were typically lower. While individual panelists may have discerned such patterns, no significant differences were revealed at the grade level.

More research is needed to determine how judges perceive differences in prompts for different types of writing. 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.

EVALUATION OF INTRAJUDGE CONSISTENCY

Intrajudge consistency, either within rounds or across rounds, is generally regarded to be a reasonable criterion by which to judge a standard setting process. Indicators of intrajudge consistency include, for example, 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 a panelist's rating for an individual item and the overall estimate of student performance at the panelist's cutscore provides a "within rounds" indicator of intrajudge consistency. Previous efforts to provide this intrajudge consistency data as feedback were not considered successful. Reckase Charts provided a means of providing this type of consistency information to panelists, along with several other consistency indicators.

INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The consistency of a judge's ratings across rounds can be examined by evaluating the percentages of items for which the ratings were changed from round to round and the magnitude of change in ratings from round to round. These data can be found in Tables 5-8. After reviewing the feedback presented following Round 1, panelists were given the opportunity to change their ratings for Round 2. These changes have been reported as percentages of item rating changes in Table 5. The same procedure was followed after Round 2 when panelists could change their ratings for Round 3. These percentages of item rating changes have been displayed in Table 6. Tables 7 and 8 display the magnitude of average rating changes following the same procedures.

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

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	26	26	17	23	29	31
	B	46	54	54	14	11	9
8	A	42	27	27	16	38	24
	B	28	34	26	31	25	30
12	A	31	24	26	41	35	41
	B	45	37	35	16	19	23

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

Grade	Group	Raise			Lower		
		Basic	Proficient	Advanced	Basic	Proficient	Advanced
4	A	18	16	5	5	12	27
	B	13	12	3	6	12	34
8	A	16	12	9	12	28	20
	B	4	9	5	11	6	12
12	A	13	7	11	17	17	45
	B	19	23	9	7	9	29

Table 7
Magnitude of Average Rating Changes (in Percentage) for Rating Group
Round 1 to Round 2

Grade	Group	Achievement Level		
		Basic	Proficient	Advanced
4	A	3.6	4.2	3.5
	B	4.7	5.0	4.4
8	A	4.9	4.8	3.2
	B	3.4	3.9	3.3
12	A	5.1	4.0	4.4
	B	5.7	3.7	3.3

Note: The percentages reported here are based on a maximum change of 5. Thus, an average change of 3.6% in ratings would result in a 0.18 average change in performance estimates.

Table 8
Magnitude of Average Rating Changes (in Percentage) for Rating Group
Round 2 to Round 3

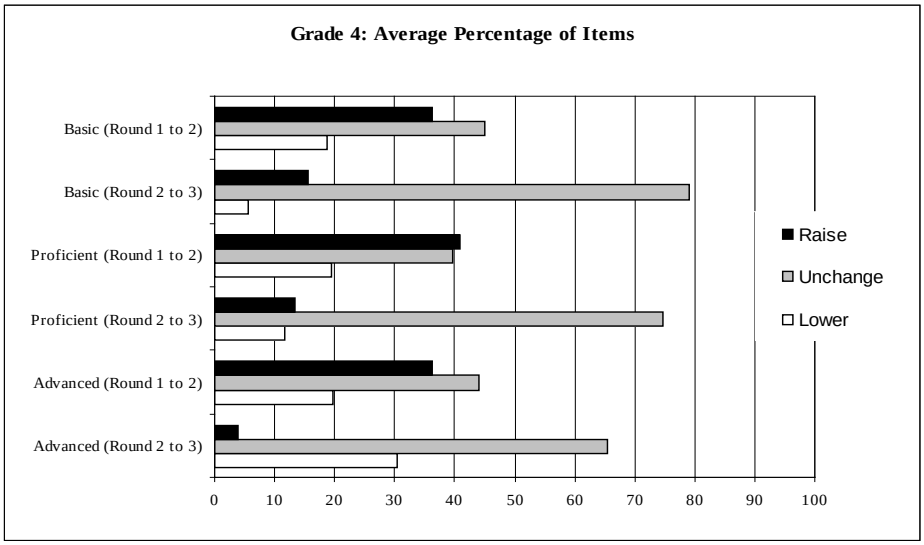
Grade	Group	Achievement Level		
		Basic	Proficient	Advanced
4	A	1.2	1.0	1.6
	B	0.8	1.0	1.6
8	A	1.0	1.5	1.1
	B	0.5	0.6	0.5
12	A	1.0	0.8	2.6
	B	1.3	1.2	1.5

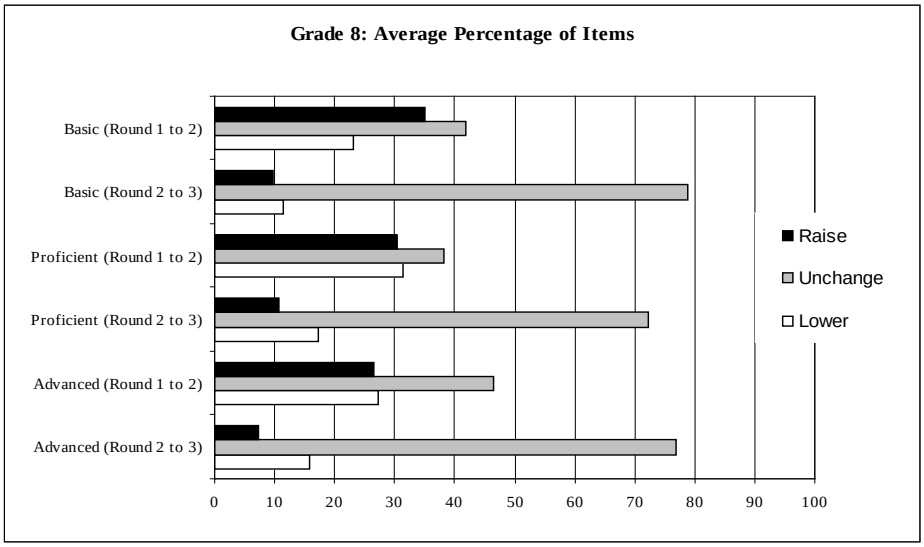
Note: The percentages reported here are based on a maximum change of 5. Thus, an average change of 3.6% in ratings would result in a 0.18 average change in performance estimates.

These data are displayed as bar graphs. Figure 1 shows the percentages of items for each grade for which ratings were raised, lowered, and unchanged from one round to the next and Figure 2 shows the magnitude of rating changes from one round to the next.[18]

[18] Please refer to Appendix K for a description of the data computations in this small study.

Figure 1
Summary of Writing Item Rating Changes





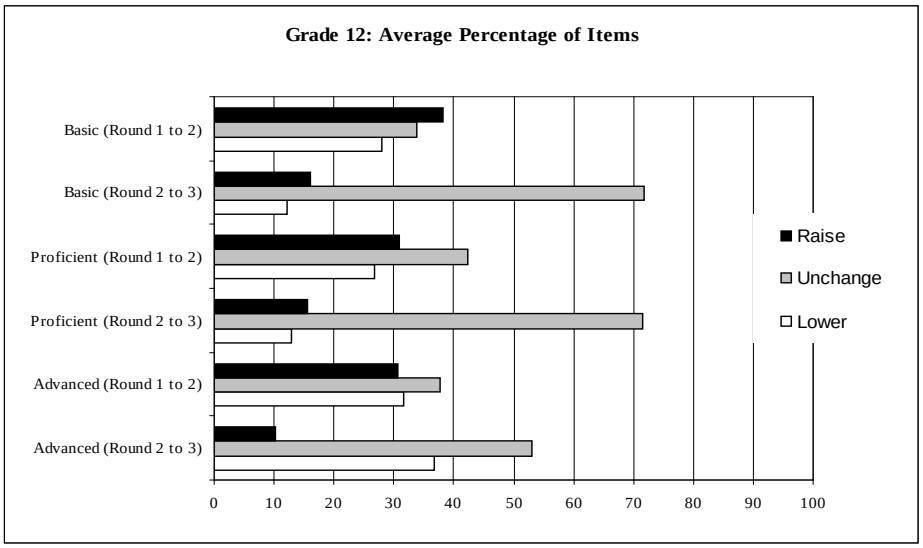
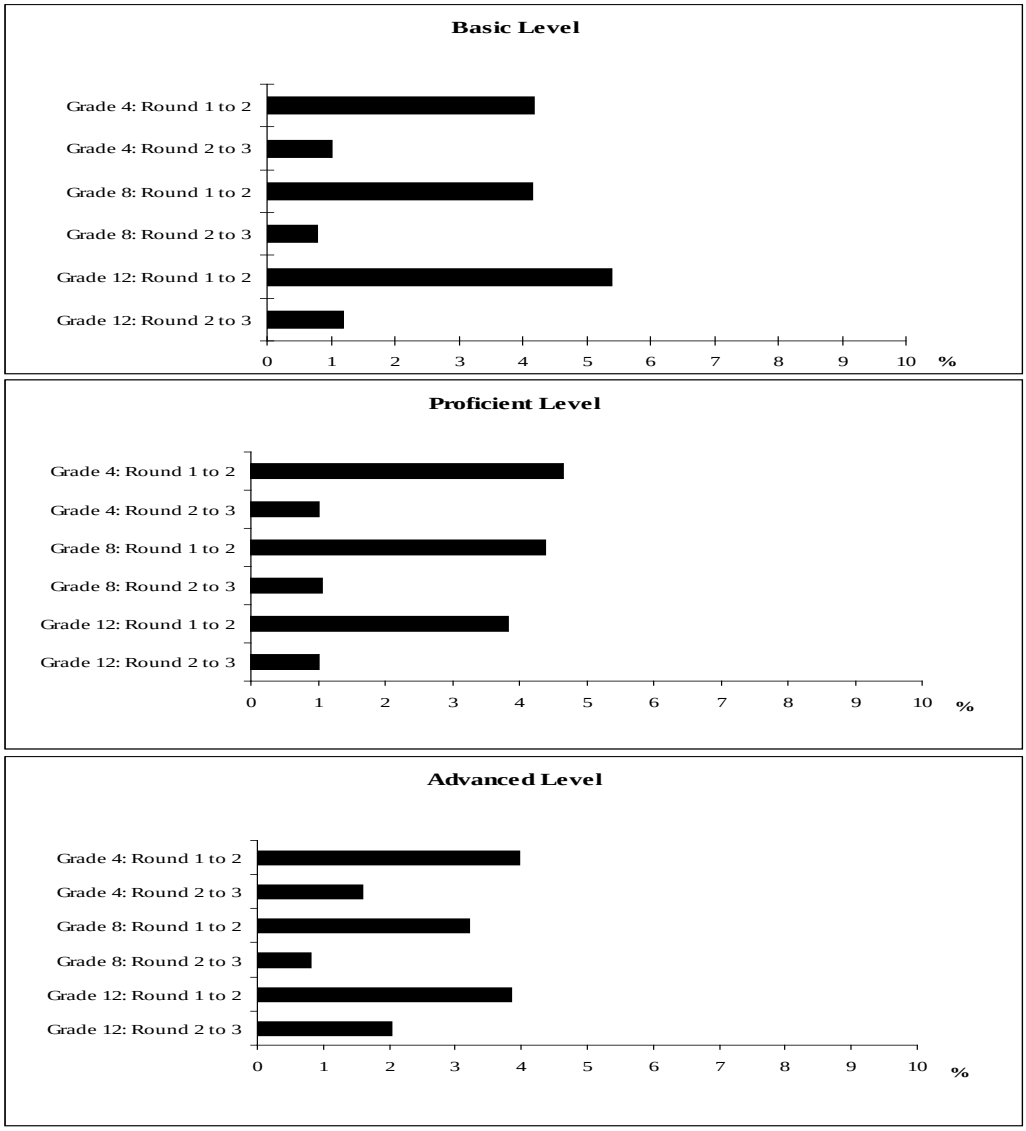


Figure 2
Magnitude of Rating Changes (RMSD) on ACT NAEP-Like Scale



Across grades and achievement levels, panelists have consistently been found to change their ratings on fewer items from round 2 to round 3 than from round 1 to round 2. Approximately 55%-65% of item ratings were changed from round 1 to round 2. From round 2 to round 3, however, only about 20%-35% of item ratings were changed. The exception was grade 12 Advanced, where over 50% of the item ratings were changed from round 2 to round 3. 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 also greater between rounds 1 and 2 than between rounds 2 and 3. 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 been found to make large adjustments to item ratings between rounds 2 and 3, this would have caused concern because it would have indicated that panelists were perhaps confused by the feedback data or item rating method.

INTRAJUDGE CONSISTENCY WITHIN ROUNDS (RECKASE CHARTS ANALYSES)

The Reckase Charts were introduced to the 1998 ALS Writing 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 inconsistency with respect to item ratings. After round 2 ratings, the panelists again worked with the charts in the same manner.

By examining the charts, the judges were able to consider the relationship between their estimates of student performance for each item and IRT model-based estimates of student performance at the cutscores. Further, judges could consider observable patterns in their ratings, such as different performance level estimates for narrative, informative, and persuasive prompts. Reckase Charts provided panelists with intrajudge consistency data for each prompt. Consistency could be evaluated with respect to grade level cutscores, panelist's cutscores, type of prompt, and so forth.

Panelists could determine whether their consistency diminished across prompts in their rating pool, for example.

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 that level.

The direction and magnitude of the differences between panelists' round 3 actual ratings and expected ratings were analyzed by grade and by achievement level. Differences between rating groups were studied, as well as differences for individual panelists. Please refer to Appendix J for more details about these analyses.

Results of the analyses did not reveal a clear pattern. For example, the percentages of items were calculated for which observed ratings were higher than, lower than, or equal to expected ratings for each grade at each achievement level. Differences were found for different grades, but no clear patterns were observed. For all achievement levels, the 4th grade panelists' ratings resulted in the largest average percentages of items for which the actual ratings differed from the expected ratings. 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 1998, consequences data had never been introduced in the operational NAEP ALS process before the final round of ratings were collected. For the 1998 Writing NAEP ALS process, consequences data were introduced before collecting the final round of ratings. Panelists were told the percentage of students at each grade performing at or above each achievement level as feedback from Round 3 ratings. They had the opportunity to adjust their cutpoints in response to those data. Comments were collected from panelists regarding their reactions to and opinions about the consequences of their cutscores.

When asked what they considered when making their cutscore recommendations, most panelists indicated that consequences data had little impact on their recommended cutscores. Panelists did indicate, however, that they considered consequences data helpful in forming their judgment about student performance. Although many panelists were quite concerned about the unexpectedly low performance of students relative to their Advanced level cutscores, they generally seemed unwilling to make substantial changes in their cutscores. They wanted to make some adjustments, but were reluctant because they felt confident in their item-by-item judgments. Several voiced concerns regarding the seemingly arbitrary nature of recommending cutscores, in contrast to the methodical collection of judgments during the item-by-item rating rounds.

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, 2000). Most panelists made no changes in their cutscores after receiving consequences data, even though they had the opportunity. The Advanced cutscore seemed least acceptable to panelists overall, particularly grade 4 and grade 12 panelists. Of the panel members who wanted to change one or more cutscores, most of them lowered the Advanced cutscore. When asked if these percentages reflected the panelist's expectations for the proportions of students scoring at or above his/her cutscores, 52 of the 88 panelists (59%) answered "yes." A total of 37 panelists recommended changes to 53 cutscores. Approximately 20% of the cutscores were changed (264 possible cutscores), and the changes were a mix of raising and lowering the cutscores for different achievement levels. Thirty of the 53 changes recommended were for the Advanced level cutscores, 13 were for Basic, and the remaining 10 were for changes at the Proficient level. Table 9 displays these data. The net effect of changes in cutscores was to slightly raise the Basic cutscores for grade 4 and grade 12, and to slightly lower the Advanced cutscores for grade 4 and grade 12. The cutscores for the grade 8 panel essentially remained the same cutscores. Individual consequences data for each panelists are reported in Appendix F.

Table 9
Number of Changes Made to Individual Cutscores in Response to the Consequences Data Questionnaire #1 Reported by Grade Groups for Writing ALS

	Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)
Data Reflects Expectations?			
Yes	14	22	15
No	15	8	14
No Response	0	0	0
(If no) Change one or more?			
Yes	15	6	16
No	8	6	2
No Response	6	18	11
Recommend Changes to Cutscores			
Basic			
Raise	4	1	5
Lower	1	2	0
Proficient			
Raise	2	0	2
Lower	2	1	3
Advanced			
Raise	0	1	1
Lower	12	3	13

GRADE LEVEL CONSEQUENCES DATA

Panelists received grade-level consequences data 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.

When asked if the final percentages reflected panelists' expectations for the proportions of students scoring at or above the grade level cutpoints, 76 of the 88 panelists (86%) answered "yes" and 12 (14%) answered "no." Results of the recommendations have been presented in Table 10. As those data show, 5 panelists indicated that they would raise the Basic cutscore. No changes were suggested for the Proficient level. Eight persons suggested changes in the Advanced cutscore: 1 recommended setting it higher, and 7 recommended setting it lower. Fifty-seven panelists (64%) recommended that NAGB report the achievement levels as set, while 11 panelists (12%) recommended changes consistent with their expectations about the proportions of students scoring at or above the cutscores. Twenty panelists did not respond. These responses were collected to document panelists' evaluations of the final cutscores. There was no plan to make adjustments to the cutscores as a result of panelists' recommendations.

Table 10
Number of Changes Made to Cutscores in Response to the Consequences Data Questionnaire #2 by Grade Groups for Writing ALS

	Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)
Data Reflects Expectations			
Yes	24	27	25
No	5	3	4
No Response	0	0	0
(If no) Change one or more?			
Yes	2	3	5
No	10	6	7
No Response	17	21	17
Recommend Changes to Cutscores			
Basic			
Raise	2	0	3
Lower	0	0	0
Proficient			
Raise	0	0	0
Lower	0	0	0
Advanced			
Raise	0	1	0
Lower	3	2	2
Recommend to NAGB			
Grade Cutscores as Set	20	19	18
Grade Cutscores Changed	3	3	5
Uninterpretable/No Response	6	8	6

EVALUATION OF PANELISTS' COMMENTS AND PROCESS EVALUATION QUESTIONNAIRES DATA

Panelists were asked to respond to seven process evaluation questionnaires. Most responses were collected on a Likert-type scale, but several open-ended questions were always included for panelists' comments regarding specific aspects of the process. Some items included on the questionnaires 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.

Some of the responses of Writing ALS panelists have been presented for comparison with those of the 1998 Writing Pilot Study. In general, Writing ALS panelists responded quite positively to the ALS process and their experience as participants. The Writing ALS panelists' responses were noticeably more positive than those for the pilot study, particularly for responses made after Round 3. The comments and responses of panelists to the process evaluation questionnaires have been presented in Appendix L. They are presented both by grade and by panelist type.

PANELISTS' UNDERSTANDING OF THE RATING PROCESS AND CONFIDENCE IN RATINGS

Data reported in Table 11 show the average responses (5=most positive and 1=most negative) to questions about the rating sessions round by round. As expected, panelists' responses generally

reflected an increase in understanding and confidence as the rounds of ratings progressed. By Round 3, the responses were very high to questions about the *clarity of instructions* and the *level of understanding* of the tasks (range 4.8 – 4.9). The *level of confidence* increased substantially from Round 1 to Round 3 as reflected by the considerable increase the degree of positive response for each grade (grade 4 increased 2.0 points, grade 8 increased 1.5 points, and grade 12 increased 1.2 points). Many judges commented that providing the first round of rating was a difficult task for them, which is reflected in the lower responses after Round 1.

Another point of interest is the response to the question related to the amount of time panelists had to complete the rating tasks. After Round 1, most panelists indicated that the amount of time was about right to complete the task (5= far too long, 3 = about right, and 1= far too short). For each progressive 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. When comparing the Writing Pilot Study results with the Writing ALS, the amount of time required to complete the rounds of ratings was closer to the amount of time allocated for the ALS panelists than for the pilot study panelists.

PANELISTS' UNDERSTANDING OF THE ACHIEVEMENT LEVELS DESCRIPTIONS AND BORDERLINE PERFORMANCE

A typical response pattern that has emerged from past ALS meetings is that panelists generally understand achievement levels descriptions across the level better understood than the borderline descriptions (see Table 12). Panelists' understanding of both categories of performance usually increases over rounds so that the difference between the two diminishes by Round 3. Results from the Writing ALS reflected this expected pattern. For the Writing ALS, understanding the definition of borderline performance was noticeably lower than understanding the definition of achievement level performance for all grades for Round 1. Grade-level panels at all grades indicated highly positive responses when asked about their understanding of the definitions of achievement level performance and borderline performance after Round 3. Panelists' understanding of student performance across the achievement levels approached *absolutely clear* by Round 3. The mean of their responses ranged from 4.4 to 4.8 by Round 3. Their conception of borderline performance approached *very well formed* for all achievement levels by Round 3. Most of the mean scores for understanding borderline descriptors ranged between 4.4 to 4.7, except grade 12, which was 3.9 for borderline Advanced.

The grade 12 panel responded less positively than the other grade panels to the question related to their conception of borderline Advanced performance. Even by Round 3, the grade 12 Writing ALS panelists responses were noticeably lower than the other grades for the Writing ALS and the Writing Pilot Study.

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

Questions	Round	Writing ALS			Writing Pilot Study		
		Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)	Grade 4 (n=20)	Grade 8 (n=18)	Grade 12 (n=18)
1. The instructions on what I was to do during the 1st/2nd/3rd rating session were: *(5=Absolutely Clear; 1=Not at all Clear)*	1	3.2	3.8	3.9	3.04	3.33	3.71
	2	4.8	4.6	4.8	4.57	4.44	4.55
	3	4.9	4.9	4.9	4.81	4.88	4.83
2. My level of understanding of the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Totally Adequate; 1=Totally Inadequate)*	1	3.1	3.7	4.0	3.00	3.38	3.71
	2	4.7	4.6	4.6	4.52	4.44	4.61
	3	4.8	4.8	4.9	4.76	4.77	4.77
3. The amount of time I had to complete the tasks I was to accomplish during the 1st/2nd/3rd rating session was: *(5=Far too Long; 3=About Right;1=Far too Short)*	1	3.1	3.1	3.1	3.09	3.55	3.18
	2	3.5	3.6	3.4	3.71	3.77	3.66
	3	3.7	4.0	3.3	4.00	4.11	3.55
4. The most accurate description of my level of confidence in the ratings I provided to represent the three achievement levels during the 1st/2nd/3rd rating session is that I was: *(5=Totally Confident; 1=Not at all Confident)*	1	2.5	3.0	3.1	2.57	2.83	3.47
	2	4.0	4.0	4.0	3.71	4.05	3.88
	3	4.5	4.5	4.3	4.19	4.55	4.38

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

Questions	Round	Writing ALS			Writing Pilot Study		
		Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)	Grade 4 (n=20)	Grade 8 (n=18)	Grade 12 (n=18)
1. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Basic level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1	4.2	4.1	4.0	3.81	3.76	3.66
	2	4.5	4.5	4.6	4.19	4.22	4.50
	3	4.8	4.7	4.8	4.33	4.55	4.55
2. At the time I provided the 1st/2nd/3rd set of ratings my conception of Borderline Basic performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.8	3.5	3.3	3.38	3.24	3.88
	2	4.4	4.2	4.4	4.04	4.05	4.44
	3	4.7	4.7	4.7	4.28	4.44	4.38
3. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Proficient level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1	4.1	4.2	4.0	3.66	3.71	3.66
	2	4.4	4.5	4.6	4.09	4.22	4.05
	3	4.6	4.7	4.8	4.33	4.44	4.44
4. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Proficient performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.7	3.5	3.5	3.38	3.24	3.77
	2	4.3	4.2	4.3	3.71	4.00	4.05
	3	4.6	4.6	4.6	4.47	4.38	4.33
5. At the time I provided the 1st/2nd/3rd set of ratings, my understanding of the definition of student performance at the Advanced level of achievement was: *(5=Absolutely Clear; 1=Not at all Clear)*	1	4.3	4.3	4.1	3.95	3.72	3.65
	2	4.5	4.6	4.4	4.04	4.33	4.22
	3	4.5	4.7	4.4	4.42	4.44	4.55
6. At the time I provided the 1st/2nd/3rd set of ratings, my conception of Borderline Advanced performance was: *(5=Very Well Formed; 1=Not Well Formed)*	1	3.7	3.6	3.4	3.52	3.16	3.82
	2	4.4	4.2	3.5	3.85	4.00	4.00
	3	4.5	4.6	3.9	4.47	4.38	4.33

PANELISTS' EVALUATIONS OF FEEDBACK

Many different types of feedback information were given to the panelists during the ALS process. When asked if they were planning to use *all* the feedback information to adjust their ratings during Round 2, most panelists responded positively (grade 4 = 4.6; grade 8 = 4.6; grade 12 = 4.7; when 5 = *totally agree* and 1 = *totally disagree*). These data suggest that when panelists were modifying their ratings, they were not overly influenced by one type of feedback to the exclusion of all others. Most panelists indicated that the Reckase Chart was the most useful type of feedback information (please see Table 13). With regard to the amount of feedback given to panelists during the rating process, most panelists remarked that they were able to manage the amount of information without confusion, but acknowledged that they were reaching their limit.

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

If you had to choose one, and only one, of the following types of information to use during the rating process, what would it be? (Reported after Round 3 ratings)	Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)
Consequences data	1 (3.6%)	3 (10%)	3 (10.3%)
Student performance data on each item	4 (14.3%)	10 (33.3%)	6 (20.7%)
Rater location feedback	8 (28.6%)	3 (10%)	4 (13.8%)
Whole booklet feedback	0	0	2 (6.9%)
Reckase Charts	15 (53.6%)	14 (46.7%)	14 (48.3%)
Blank	1 (3.6%)	0	0

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 Writing ALS participant. However, all three grade groups most frequently selected the Reckase Charts as the single feedback data of choice. The number of responses for the second most frequently selected form of feedback was fairly evenly split between rater location feedback and student performance data for each prompt. 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. Those data are reported Table 14. All three of the grade panels for the Writing ALS ranked the Reckase Charts first. Grade 4 and grade 8 panels ranked rater location feedback second, whereas the 8th grade group ranked student performance data about equally with rater location feedback for their second choice.

Table 14
Writing ALS: Rank Order of Types of Feedback

Type of Feedback	Response	All Grades		Grade 4		Grade 8		Grade 12	
		N	%	N	%	N	%	N	%
Whole Booklet	5	4	15.7	3	10.3	4	13.3	7	24.1
	4	13	14.8	5	17.2	2	6.7	6	20.7
	3	17	19.3	3	10.3	8	26.7	6	20.7
	2	20	22.7	11	37.9	3	10.0	6	20.7
	1	24	27.3	7	24.1	13	43.3	4	13.8
Rater Location	5	10	11.4	2	6.9	5	16.7	3	10.3
	4	35	39.8	18	62.1	6	20.0	11	37.9
	3	30	34.1	7	24.1	14	46.7	9	31.0
	2	7	8.0	1	3.4	4	13.3	2	6.9
	1	6	6.8	1	3.4	1	3.3	4	13.8
Student Performance	5	18	20.5	4	13.8	7	23.3	7	24.1
	4	14	15.9	2	6.9	7	23.3	5	17.2
	3	25	28.4	10	34.5	9	30.0	6	20.7
	2	19	21.6	8	27.6	5	16.7	6	20.7
	1	12	13.6	5	17.2	2	6.7	5	17.2
Reckase Charts	5	47	53.4	20	71.4	13	43.3	14	48.3
	4	20	22.7	5	17.9	9	30.0	6	20.7
	3	10	11.4	1	3.6	5	16.7	4	13.8
	2	4	4.5	1	3.6	1	3.3	2	6.9
	1	6	6.8	1	3.6	2	6.7	3	10.3
Consequences	5	21	23.9	2	7.1	1	3.3	6	20.7
	4	23	26.1	9	32.1	3	10.0	7	24.1
	3	19	21.6	5	17.9	7	23.3	7	24.1
	2	16	18.2	5	17.9	11	36.7	4	13.8
	1	8	9.1	7	25.0	8	26.7	5	17.2

5=Most Helpful; 1=Least Helpful

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. Still others indicated it would have been better after Round 3. When asked about the impact of consequences data on the cutscores they recommended to NAGB, most panelists chose to comment on the state of writing education in the U.S. 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 Writing*. Many expressed concern about how the results will be reported and interpreted, fearing that important details about NAEP and the context of the Writing NAEP will be misunderstood. They strongly urged that data reports be accompanied by information regarding the time limits and other circumstances for writing on the NAEP.

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

Questions		Writing ALS			Writing Pilot Study		
		Grade 4 (n=29)	Grade 8 (n=30)	Grade 12 (n=29)	Grade 4 (n=20)	Grade 8 (n=18)	Grade 12 (n=18)
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)*		4.3	4.3	4.1	4.10	4.11	4.11
2. I would describe the effectiveness of this achievement levels-setting process as: *(5=Highly Effective; 1=Not at all Effective)*		4.3	4.1	3.9	3.95	3.89	3.67
3. I feel that this NAEP ALS process provided me an opportunity to use my best judgment in rating items to set achievement levels for the NAEP Writing Assessment: *(5=To a Great Extent; 1=Not at All)*		4.6	4.7	4.3	4.10	4.11	4.00
4. I feel that this NAEP ALS process produced achievement levels that are defensible: *(5=To a Great Extent; 1=Not at All)*		4.2	4.5	4.1	4.10	3.83	3.94
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)*		4.4	4.4	4.0	4.10	3.27	4.00
6. I would be willing to sign a statement (after reading it, of course) recommending the use of the achievement levels resulting from this achievement levels-setting procedure: *(4=Definitely, 3=Probably, 2=Probably not, 1=Definitely not)*	#						
	4	58.6%	73.3%	61.5%	40%	44.4%	44.4%
	3	37.9%	26.7%	38.5%	45%	44.4%	44.4%
	2	3.4%	0	0	15%	11.1%	5.6%
	1	0	0	0	0	0	5.6%

PANELISTS' EVALUATION OF THE OVERALL ALS PROCESS

Data reported in Table 15 show the average responses to questions from the final questionnaire about the overall ALS process used for the Writing ALS. When asked if the achievement levels were "defensible" and "reasonable," Writing ALS panelists' responses were consistently higher compared with those for the Writing Pilot Study. The responses of the Writing ALS panelists (range 4.3 – 4.7) were particularly positive to the question related to panelists using their best judgment in rating items. When asked if they would be willing to sign a statement recommending the use of the ALS results, all but one of the 88 panelists replied positively.

EVALUATION OF RESPONSES OF "EXTREME" RATERS

ACT closely examined the relationship between panelists' ratings of prompts and their responses to key questions on the process evaluation questionnaires. Of interest was to identify "extreme" raters, based on their markings of the Reckase Chart and changes in their round three ratings. Although it was very interesting to study how panelists adjusted their ratings using the Reckase Charts, no revealing link was discovered between "extreme" raters and their responses to process evaluation questionnaires. Overall, the responses of "extreme" raters were quite similar to the mean responses of all Writing ALS panelists. Please see Appendix M for the outcomes of these analyses.

EVALUATION OF THE SELECTION OF EXEMPLAR PERFORMANCE

One of the primary outcomes of the NAEP ALS process is the identification of assessment performances to illustrate the knowledge and skills associated with each achievement level for each type of writing (narrative, informative, persuasive) to use in reporting NAEP results. Appendix I includes papers selected by the Writing ALS panelists to serve as exemplar performances for reporting the NAEP achievement levels. Recall that the paper classifications were collected from an exercise to sharpen panelists' concepts of borderline performance prior to the first round of ratings.

Panelists reviewed and discussed the student writing that qualified statistically for consideration as exemplars. (Please see Table 16.) One example of each type of writing prompt was included for consideration in the exemplar selection process. Panelists were instructed to veto performances that met statistical criteria but that did not meet the criteria described in the ALDs. Judges were trained in the statistical criteria that had been used for selecting the performances (described earlier in this report). In addition, panelists were instructed to use their knowledge of the achievement levels descriptions to evaluate each paper in terms of its quality as an illustrative or exemplar performance. Note that no scores qualified or no qualifying papers were selected for some prompt types for some achievement levels.

Table 16
Number of Papers Selected as Exemplar Performance for Writing NAEP ALS for Each Grade and Achievement Level

	# Qualified	# Selected
Grade 4		
Basic	9	6
Proficient	9	7
Advanced	9	6
Grade 8		
Basic	12	2
Proficient	9	4
Advanced	15	6
Grade 12		
Basic	9	4
Proficient	9	3
Advanced	9	5

ACT used the performances selected by panelists to compile a set of three exemplar papers for each of the three achievement levels for each of the three grades. These student papers were selected for approval by NAGB to be used in the *Nation's Report Card* for writing. Both TACSS and the NAGB Achievement Levels Committee reviewed the papers selected by panelists. Each committee found a few responses that did not seem to be best choices because the content of a specific response included information or language that was deemed inappropriate for general distribution. ACT provided substitutes for those particular papers, and a final set of papers was prepared for use in reporting results relative to the achievement levels for each type of writing.

CONCLUSIONS DRAWN FROM THE WRITING NAEP ALS STUDY

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

THE ISSUE OF IMPROVING AND REFINING THE NAEP STANDARD SETTING PROCESS

The process implemented in the Writing 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 evaluations of the procedures were very positive.

It is important to emphasize the refinement of the ALS process designed by ACT to develop the 1998 NAEP Achievement Levels in Writing. 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 THE ALS MEETING

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 writing. 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 with the *Writing Framework* but also with the policy definitions 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 Writing ALS expressed a desire to revise the ALDs they were given to use in the ALS process.

PROVIDING CONSEQUENCES DATA DURING THE ALS 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 in past ALS studies, panelists 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 prompt along several, important dimensions. For example, panelists could easily evaluate the relative consistency of their ratings for each prompt and the relative consistency of their ratings for each type of writing prompt. All three grade panels for the Writing 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 problem is to identify the relevant level of student performance. Individual item ratings can be used to compute a cutscore for each panelist. That cutscore then becomes the representation of a panelist's concept of borderline performance for a level of achievement. The panelist's ratings for each item are associated with an overall performance score (cutscore). If the item ratings are all for the same performance score, then the panelist has managed to perfectly estimate student performance for each item consistent with the IRT model used to estimate student performance on NAEP. It seems safe to assume that no panelist has ever achieved that level of perfection in estimating student performance across NAEP items. Certainly, that was not the case in the 1998 process. Most panelists judge some items to be much harder or much easier than others, relative to their overall cutscore. Intrarater consistency is a measure of the extent to which individual item ratings are consistent with the

overall cutscore estimated from the individual item ratings, given student performance on the items. Although this information has been given to panelists in previous ALS meetings, there was little indication that panelists either understood the information, or found it useful when forming their judgments about student performance. Reckase Charts made that information easy to assess.

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

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

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The ALS process designed by ACT provided panelists with extensive feedback and instructions for interpreting information when forming their judgments of student performance. Panelists were encouraged to reconsider their ratings and adjust them according to their interpretation of the many sources of information available to them. It was reasoned that if panelists understood the item rating method and the feedback produced by the method, they would adjust their ratings from round to round. If panelists did not adjust their ratings at all, it indicated that they probably did not understand the rating method or the feedback. On the other hand, if they changed all—or most—of their ratings after two rounds, it indicated that they probably did not understand the rating method or the feedback. The writing 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 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 Writing ALS judges, who received consequences data for the first time after round 2, and judges in other ALS studies, who received consequences data after ratings were completed and cutscores recommended. 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, relatively 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 item-by-item rating methods cannot produce valid cutpoints because panelists are incapable of performing the cognitively complex task of borderline performances with reasonable accuracy (NAE, 1993: Shepard, 1995: Impara and Plake, 1998). ACT has collected considerable data during the writing 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 borderline performances with reasonable accuracy.

PUBLIC COMMENTARY ABOUT THE WRITING NAEP 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 Writing 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 in advance. 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 N 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 which distributed an invitation to respond to the survey to their membership. The following persons, groups, and organizations were contacted at the end of March, 1999 and encouraged to participate:

- Persons who served on the Writing NAEP Framework panels
- Members of the EIAC listserv
- Members of the National Writing Project

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:

- 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 WRITING NAEP ACHIEVEMENT LEVELS OPINION SURVEY

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

- The achievement levels descriptions for writing,
- Examples of actual student responses to the Writing NAEP,
- Scoring guides for each exemplar paper,
- The actual opinion survey and comments form respondents were to complete.

In addition to the achievement levels descriptions, papers representing performance at each level were included for review. Although the goal was to show an example of each of the three types of writing assessed (narrative, informative, and persuasive) at each level, this was not always possible because of statistical restrictions. In some instances, the same prompt could be used to illustrate skill progression across achievement levels at each grade. In this way differences in student performance on the same prompt at different score points could be observed and related to the different levels of achievement associated with each.

RESULTS OF THE WRITING NAEP ACHIEVEMENT LEVELS OPINION SURVEY

Ten persons replied to the writing 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 classified themselves as a classroom teacher (5), nonteacher educator (5), or member of the general public (10).

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. Some individuals felt confident to evaluate more than one grade level. Four panelists provided evaluations for grade 4, five for grade 8, and 6 for grade 12.

Overall the respondents tended to agree that the achievement levels descriptions were clear and easy to understand. All of the 10 respondents indicated that the achievement levels descriptions are at least *somewhat easy to understand*. Their comments have been reported in Appendix N. This finding was of particular interest because of the extensive development and revision of the preliminary achievement level descriptions that occurred prior to the achievement levels-setting meeting.

The majority of respondents indicated that they thought it was useful to have student performance reported in terms of achievement levels. Only one participant responded negatively to this question.

The majority of respondents indicated that they thought in general the achievement levels reflected what student should know and be able to do. One participant responded negatively to this question, and 3 noted a *mixed* response.

The responses to the question about the examples of student work representing the kinds of things student should know and be able to do according to the achievement level descriptions were varied. Three respondents checked that the exemplar papers were representative of the knowledge and skills in the achievement levels descriptions; 2 respondents checked "mixed" for their response; and 1 checked "no."

The following responses referred to participants' judgments about the reasonableness of the NAEP achievement levels for each grade. *Too high* meant that the standard was too demanding to be a reasonable reflection of the level. *Too low* meant that the standard was not demanding enough to be a reasonable reflection of the level. *About right* means that the standard was a reasonable reflection of the level. Not all respondents replied to every achievement level.

Only a few persons responded to the writing survey for each grade. (See Table 17.) Based on this very small sample, it would appear that there is some support for the achievement levels being *about right* at the 4th grade level. Support for the reasonableness of the 8th grade levels seems somewhat mixed, with less support for the achievement levels being *about right* at the 12th grade level. Four respondents indicated that the standard is too low for 12th grade Basic and 3 respondents indicated that the standard is too low for 12th grade Proficient.

Table 17
Respondents' Judgments about the Reasonableness of the Writing NAEP Achievement Levels for Each Grade

	Basic	Proficient	Advanced
Grade 4 (n=4)			
About right	3	3	3
Too high	1	1	0
Too low	0	0	1
Grade 8 (n=5)			
About right	3	2	2
Too high	2	2	1
Too low	0	1	2
Grade 12 (n=7)			
About right	1	2	4
Too high	2	1	1
Too low	4	3	1

NAGB APPROVAL OF ACHIEVEMENT LEVELS-SETTING PROCESS OUTCOMES FOR 1998 WRITING NAEP

The ACT Project Director made presentations to NAGB's Achievement Levels Committee throughout the process of design, research, and implementation. In addition, the technical advisory committees were closely monitoring the process and outcomes at each stage. After careful review of all data analyses regarding the entire 1998 Writing NAEP ALS process, TACSS

recommended adoption of the outcomes. On May 1, 1999, the NAGB Achievement Levels Committee approved the descriptions and cutscores recommended by ACT. They requested some paper substitutions for the exemplar performances, but they gave general approval to the exemplar performances selection process and the papers selected by panelists. During their regularly scheduled meeting in May 1999, NAGB approved the 1998 Writing 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 (1995). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history.* Iowa City, IA: Author.

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

ACT (1997b). *Setting achievement levels on the 1996 National Assessment of Educational Progress in science: Final report, Volume IV. Validity evidence special studies*. Iowa City, IA: Author.

Bay, L. (2000). *1998 NAEP writing achievement levels-setting process performance profiles.* Paper presented at the meeting of the National Council on Measurement in Education, New Orleans.

Bolton, S., Hanick, P.L., Welch, C., & Loomis, S.C. (2000). "Expectations for Student Writing Skills: A Comparison of Requirements in States to "Solid Academic Performance" on NAEP" in Loomis, S.C. (Ed.). *Developing achievement levels on the 1998 National Assessment of Educational Progress in Writing: Research studies*. Iowa City, IA: ACT.

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

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

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

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

Jaeger, R.M. (1995). Setting performance standards through two-stage judgmental policy capturing. *Applied Measurement in Education*, 8, 15-40.

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

Loomis, S.C. (Ed.) (2000). *Developing achievement levels on the 1998 National Assessment of Educational Progress in Writing: Research studies*. Iowa City, IA: ACT.

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

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

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

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

National Assessment Governing Board (1998). *Writing 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.

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

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.