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

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

December 2000

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

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

December 2000

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

Table of Contents

EXECUTIVE SUMMARY OF THE WRITING FIELD TRIALS

Susan Cooper Loomis

OVERVIEW

Several studies, including two field trials, were conducted for the achievement levels-setting (ALS) process for the1998 Writing NAEP. ACT proposed to conduct field trials as a means of collecting research information regarding new methods and procedures designed for the 1998 ALS process. Further, ACT had conducted ALS procedures for the 1992 Writing NAEP. ACT wanted to collect additional research data to determine whether problems similar to those in the 1992 writing NAEP were likely to be encountered in the 1998 Writing ALS process. ACT proposed to conduct the research involving panelists before the pilot study so that the pilot studies could be used to test the procedures selected for the ALS. Experiences with research-laden pilot studies led ACT to recommend this additional set of studies for research purposes so that pilot studies could be used for practice in implementing and fine-tuning procedures.

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

This report documents two field trials conducted for writing. Taken together, the field trials research provided important information about various elements that constitute the ACT/NAGB standard-setting process. The opportunity to study new methods implemented with panelists led to significant improvements in the ACT/NAGB ALS Process. Findings from the field trials research greatly informed the procedures that were developed for the pilot study and implemented to set achievement levels in 1998.

WRITING FIELD TRIAL 1

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

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

- ✓ *Results of the first field trial in writing indicated that panelists were able to use the ISSE method without difficulty, and that they were satisfied with and confident in the results of the ISSE method.*

- ✓ *The ISSE cutpoints for the Basic level were lower and for the Advanced level were higher than those produced by the Mean Estimation procedure. The standards deviation of the cutpoints for the ISSE method was smaller than for the Mean Estimation method.*

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

WRITING FIELD TRIAL 2

- ✓ *The fundamental purpose of the second field trial was to identify the method that would be used for the 1998 ALS process.*

- ✓ *ACT studied the new Reckase standard setting method, the Booklet Classification method, and the timing of consequences data on the outcomes of the process.*

- ✓ *Results of Field Trial 2 indicated that panelists had no difficulty using either the Booklet Classification method or the Reckase method.*

- ✓ *The Advanced cutscore resulting from the Booklet Classification method was lower than that for the Reckase method. This finding was contrary to previous results indicating that using the Booklet Classification would consistently result in higher cutscores.*
  - *The fact that booklets in the classification study were ordered from lowest to highest performance scores was seen as the likely reason for this result.*

- ✓ *The cutscores of panelists using the Booklet Classification method did not differ according to the timing of consequences data.*

- ✓ *The cutscores of panelists using the Reckase method were higher for panelists who received consequences data throughout the process than for those who received it at the end of the process.*

- ✓ *The decision was made to use the Reckase method for the 1998 Writing ALS.*
  - *There was no reliable method for computing cutscores.*
  - *Previous research conducted by ACT had consistently pointed to higher cutscores with the Booklet Classification method than with item-by-item methods.*
  - *Previous research indicated that panelists were likely using a noncompensatory model for judging student performances in the booklets.*
  - *The logistic requirements for the Booklet Classification method are immense.*

FINDINGS REGARDING ISSUES IN NAEP STANDARD SETTING

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

THE ISSUE OF COGNITIVE COMPLEXITY

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

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

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

- ✓ *There is no evidence to indicate that panelists are unable to make these judgments.*

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

One of the considerations for evaluating the various experimental methods explored by the field trials was based in part on indicators of "reasonable" intrajudge consistency across rounds. Panelists were encouraged to reconsider their ratings and classifications and adjust them according to their interpretation of the many sources of information available to them. Judges were expected to adjust their ratings and classifications from round to round. It was reasoned that if panelists understood the method and the feedback produced by the method, they would adjust their ratings or classifications from round to round. If panelists did not adjust their ratings or classifications at all, this was an indicator that they probably did not understand the method or the feedback. On the other hand, if they changed all—or most—of their ratings or classifications after two rounds, this was also an indicator that they probably did not understand the method or the feedback.

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

- ✓ *Statistical analyses of rating changes for field trial 1 and visual inspections of the results for field trial 2 led to the overall judgment that writing field trial panelists exhibited "reasonable" intrajudge consistency across rounds.*

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS (RECKASE CHARTS)

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 relative to their overall ratings and cutscores. Although this information had been given to panelists in previous ALS meetings, there was no evidence to suggest that panelists either understood the information, or found it useful when forming their judgments about student performance.

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

- ✓ *Panelists who used the Reckase method in general adjusted their ratings for Round 2 to be more similar to the IRT-based performance estimates of students at the cutscores. This finding was consistent for all three achievement levels.*

It is important to note that for the third rating session panelists were instructed to select a single row or scale score to represent their ratings for each achievement level. There was concern that this procedure could overly influence judgments and that the final cutpoints would be based on data from the Reckase charts rather than the standards represented by the ALDs. Panelists' reactions to the charts did not support this concern, however.

- ✓ *Panelists indicated that they were influenced by the achievement levels descriptions and other forms of feedback in addition to the charts when forming their final judgment of student performance.*

THE ISSUE OF THE TIMING OF CONSEQUENCES DATA

The impact of consequences data on outcomes has been a topic of considerable interest in setting standards. The impact of receiving consequences data seemed moderate. About an equal number of comments were made by panelists regarding the positive impact of consequences data as were made about the limited impact.

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

- ✓ *Judges, in general, found the consequences data informative and useful, but they were not greatly influenced by the data.*

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

THE ISSUE OF COMPUTING CUTSCORES FOR THE BOOKLET CLASSIFICATION METHOD

Of particular interest in the second writing field trial was the computational procedure used to calculate cutscores for the Booklet Classification method. Each panelist classified a set of booklets into one of seven categories: Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient, Borderline Advanced, and Advanced.

ACT computed two sets of cutpoints for each panelist to analyze the computational procedures more carefully. One set of cutpoints was computed using just the booklets classified at the lower borderline of each achievement level, and another set was computed using all booklets classified within the interval.

- ✓ *The two sets of cutscores differed greatly, although theoretically the cutscores should have been essentially equivalent.*

- ✓ *TACSS concluded that given the considerable effort required to administer the Booklet Classification method, there should be strong support for the method to merit its implementation.*

- ✓ *Because TACSS members were concerned about the differences associated with alternative methods for computing the cutpoints for the Booklet Classification method, they recommend that it not be used for the next phase of planned research.*

- ✓ *Computing stable cutscores using the Booklet Classification method for NAEP remains an unresolved issue.*

PLANNING FOR THE WRITING PILOT STUDY

- ✓ *The Reckase charts, introduced in the second field trial for writing, were judged to be a promising addition to the ALS process designed for NAEP.*

- ✓ *The charts appeared to have improved and strengthened the ALS process.*

- ✓ *Computing cutscores for the Booklet Classification method proved to be problematic, resulting in unstable cutpoints for the Booklet Classification method.*

- ✓ *After reviewing the results of both the civics and writing field trials, it was agreed that research would continue on the Reckase charts.*

- ✓ *TACSS recommended using a modified version of the Reckase method for further research in the writing pilot study.*

- ✓ *TACSS recommended implementation of a design that would produce two sets of cutscores: one set based solely on ratings, and one modified as a result of consequences data. The latter were to be the cutscores used for selecting exemplar items and, presumably, to be recommended to NAGB.*
  - *This recommendation was made in recognition of NAGB wishes to have the NAEP achievement levels set via a criteria-referenced methodology.*

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

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

INTRODUCTION

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

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

Because the results of the first field trial did not lead to identification of the method that would be used for the 1998 ALS process, additional exploratory research was conducted in the second field trial in an effort to identify new methods for setting achievement levels. Taken together, the field trial research has provided important information about various elements that typically constitute standard-setting processes. In particular, the field trials explored topics that related to the cognitive complexity of the item-rating task, methods of computing cutscores, timing of consequences data,[1] differences between polytomous and dichotomous item ratings and cutscores, and effective formats for presenting intrajudge consistency feedback[2] to panelists. Findings from the field trials research greatly informed the procedures that were developed for the pilot study. The following report is a full account of the methodologies and findings from the methods used in the first and second writing field trials.

BACKGROUND INFORMATION FOR FIELD TRIALS RESEARCH

For the 1992 Writing NAEP, the Paper Selection method was used to set achievement levels. Since 1994, some form of item-by-item rating method has been used in the NAEP ALS process to determine the cutpoints for each achievement level. The specific method has been called "Mean

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

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

Estimation" (ME) because judges estimate the average or mean score of polytomous items for a student performing at the borderline of each achievement level. Critics expressed concern that the item-by-item rating method could not produce valid cutpoints because panelists were incapable of performing the cognitively complex task of estimating probabilities with reasonable accuracy (NAE, 1993; Shepard, 1995; Impara and Plake, 1998). ACT has conducted research and collected considerable evidence to the contrary. The Technical Advisory Committee on Standard Setting (TACSS) presented some of that evidence and refuted the claims of the critics (Hambleton, et al., 2000).

For the 1998 NAEP ALS process, ACT proposed to explore a new method for setting NAEP achievement levels that was thought to be less cognitively complex than the Mean Estimation method. This method was based on the work of Impara and Plake (1997). Their studies compared the modified-Angoff method with the "Yes/No" method.[3] Instead of estimating probabilities, the "Yes/No" method required judges to select "yes" if they thought a student performing at the borderline of each achievement level would respond to the assessment item correctly. Panelists selected "no" if they judged that a student performing at the borderline would not respond correctly. It was argued that selecting "yes/no" was a task panelists could perform easily, whereas estimating percentages was a task requiring a level of accuracy that panelists could not attain.

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

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

THE RATING METHOD

DESCRIPTION OF THE ITEM SCORE STRING ESTIMATION (ISSE) METHOD

ACT needed to identify a rating method that could be easily used to rate both dichotomous and polytomous items and that would produce reliable cutscores. The 1998 Civics NAEP included a large number of both dichotomous and polytomous items, while the 1998 Writing NAEP included only polytomous items (two writing tasks for each student). ACT created a new procedure that combined two rating methods that had been reported in the literature. It merged the "Yes/No" method for dichotomous items described by Impara and Plake (1997) with the extended-Angoff

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

method for polytomous items described by Hambleton and Plake (1995). This combination of rating methods formed the basis of the new Item Score String Estimation (ISSE) method.

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

COMPUTING CUTSCORES USING THE ISSE METHOD

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

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

KEY ASPECTS OF WRITING FIELD TRIAL 1

The purpose of the first writing field trial was to study the ISSE method relative to the Mean Estimation (ME) method. The criteria for evaluating the ISSE method were measures of the reasonableness of the cutpoints, the level of interjudge consistency,[4] the level of intrajudge

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

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

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

The first field trial for writing was conducted February 28 and March 1, 1997 (Saturday-Sunday) at the Tyler Conference Room on the ACT campus in Iowa City, Iowa. The study was conducted only for grade eight. The NAEP ALS Project Director facilitated the two-day study.

DATA, ACHIEVEMENT LEVELS DESCRIPTIONS, AND ITEM RATING POOLS

The 1992 NAEP Writing data were used for the field trials. Administration of the 1998 Writing NAEP had not been completed and consequently, 1998 writing data were not available to use for the field trial. The problems ACT experienced with the writing assessment data in 1992 were of concern, however. The writing NAEP framework document had been revised somewhat, and the test specifications had been sharpened and tightened since administration of the 1992 assessment. ACT used the achievement levels descriptions that had been developed for the 1998 Writing NAEP in the writing field trials, along with the 1992 data. The correspondence, or lack thereof, between the 1992 scoring rubrics which were specific to each prompt and the 1998 ALDs was not taken into account for the field trials. Although the 1992 NAEP writing data presented many limitations and restrictions, they were the best data available to use for the field trial.

The 1992 Writing NAEP for grade 8 consisted of 11 prompts in all. Nine of those were 25-minute prompts and 2 were 50-minute prompts. Since only 25-minute prompts were to be used in reporting the 1998 NAEP writing results, only the nine 25-minute prompts from the 1992 NAEP for grade 8 were used in the field trial.

In past ALS processes conducted by ACT, items used during training exercises were excluded from the item rating pools used for setting the achievement levels. In this way the potential for contaminating the ratings was eliminated. This was achieved by assigning panelists to one of two

rating groups in each grade. Each group rated a different half of the grade-level item pool. The sets of items used for training one group came from the set of items used for ratings in setting the achievement levels by the other group. The writing field trial, however, used only nine prompts. Since each panelist rated each prompt, it was necessary that the same prompts be used in the training exercises that were used for item ratings in setting the achievement levels. All but one prompt was used in training raters during the first field trial.

RECRUITMENT AND SELECTION OF PANELISTS

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

More members of the general public than educators participated in the writing field trial panel, which is unusual for an ALS panel. Most of these panelists were employed by a large international corporation with headquarters in a nearby community and participated in an employee-student writing program sponsored by the company. Teachers from different area schools served as panelists, as did professional writers. A university graduate student and a school administrator also participated in the study. Please see Table 1 for more details about the composition of the panels.

Table 1
Composition of Writing Field Trial 1 Panel

Panelist Type			Gender		Race		Total
TR	NT	GP	Male	Female	White	Minority	N=15
2	2	3	3	4	7	0	7
3	1	4	3	5	8	0	8
5	3	7	6	9	15	0	15

THE ALS PROCESS DESIGNED FOR WRITING FIELD TRIAL 1

STEP 1: BRIEFING MATERIALS

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

- 1994 NAEP *Writing Framework*;
- 1998 NAEP Writing Achievement Levels Descriptions;
- *Multiple Challenges*, a booklet on the 1998 NAEP;

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

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

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

TAKING A FORM OF THE NAEP

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

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS (ALDS)

Panelists were given time to work with the ALDs and to discuss them. They developed an understanding of the achievement levels descriptions and reached a common agreement on their meaning. Panelists studied the prompts and the scoring guides for each type of writing. They were not given the opportunity to revise the ALDs. Panelists applied the ALDs to different types of writing and to student test booklets requiring different types of writing. The final version of the grade 8 writing achievement levels descriptions (ALDs) was used for the field trials. Please refer to Loomis & Hanick (2000) for a description of how the final version of the ALDs was developed.

UNDERSTANDING BORDERLINE PERFORMANCE

Panelists also received brief training in the concept of borderline performance. They did not write borderline descriptions as part of their training, however.[5] Panelists discussed borderline performance and reached a common understanding of what constituted borderline performance at each achievement level.

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

PAPER SELECTION EXERCISE

After discussing the concept of borderline performance and reaching some agreement on how borderline performance would differ from performance across an achievement level, panelists were engaged in an exercise to implement this preliminary understanding. Panelists engaged in a paper selection exercise using three prompts, one for each type of writing (narrative, informative, and persuasive) assessed by both the 1992 and the 1998 NAEP. Each panelist was given a packet of 51 single student papers written in response to each of the three prompts. Three student papers were included for each score level (1-6) for informative and narrative writing tasks. All of the persuasive prompts in the 1992 Writing NAEP for grade 8 were rescaled to five score levels. Thus, only 15 papers, 3 at each score level, were included in the packet for the persuasive prompt used in the paper selection exercise. Panelists were asked to select at least one paper that represented borderline performance for each prompt. Panelists received a separate data sheet that listed the scores of the student papers in the packet. They could check the score of the papers after completing their selection of borderline representatives for each level for a writing task.

The paper selection training exercise served several purposes. It helped panelists

- to arrive at a common understanding of borderline performance at each achievement level,
- to become familiar with the scoring rubrics and how they were used, and
- to have a "reality check" of student performance relative to the ALDs.

STEP 3: THE ITEM RATING PROCESS

ROUND 1 RATINGS (END OF DAY 1)

After the paper selection exercise, the two groups of panelists were trained separately in the two different rating methods by the same facilitator. The rating task required judges to estimate the scores of students performing at the borderline of each achievement level. The ISSE group indicated the most likely score of those in the scoring rubric, and the Mean Estimation group indicated the average score using up to two decimal places. The rating forms used by the two groups were the same except for the color of the paper. For Round 1 ratings, panelists read each prompt, considered what constituted a Basic, Proficient, and Advanced response, checked the scoring rubric, and marked their rating forms. Please see Appendix C for detailed instructions to panelists on the ISSE and Mean Estimation rating methods.

It was necessary to inform the panelists that all three persuasive prompts in their rating pool had been rescaled to five score levels, rather than the usual six. Panelists in both rating groups were informed about the rescaled items even though rescaling affected only the Mean Estimation rating method.

FEEDBACK AFTER ROUND 1 (BEGINNING OF DAY 2)

At the start of Day 2 in a general group session, panelists were given feedback data resulting from their first round of ratings. They were provided with cutpoints, rater location charts, student performance data, and whole booklet feedback prior to the second round of ratings. Training in the use of each form of feedback data was provided. Although the groups received instructions

together, the feedback itself was provided to the panelists in separate groups. Copies of the different types of feedback based on the first round of ratings are included in Appendix D.

Cutpoints

The cutpoints are the combined ratings over all raters using the same rating method and all prompts, for each achievement level. Cutpoints are based on the mean (average) of the ratings provided by each panelist in the rating groups. The cutpoints are presented on the ACT NAEP-like scale and take into account statistical information about item difficulty, item discrimination, and chance probabilities.

Standard Deviation

The standard deviation of the cutpoint is the indicator of how different the cutpoint for each individual rater is with respect to the overall group cutpoint. The standard deviations of the cutpoints are computed separately for each rating group and distributed to the two panels.

Rater Location Feedback Charts

The rater location feedback charts consist of a horizontal axis, which represents scores on the ACT NAEP-like scale, and a vertical axis, which represent the number of judges. Letter codes that represent individual judges are positioned along the ACT NAEP-like scale at the point where each judge set his/her cutscores based on his/her individual ratings. The letter codes are "secret," and a panelist's individual cutscores cannot be identified by other panelists. The graphs display the cutscores that result from the ratings of each judge for Basic, Proficient, and Advanced levels, and the relationship of all the judges' ratings to each other (interjudge consistency).

Student Performance Data

These data are a list consisting of measures of overall student performance for each prompt. The mean (average) score is given for each prompt and the percent of students performing at each score point is reported, along with the percent of omits, off-task, and not reached. This information gives the panelist a "reality check" because it shows how students actually performed on each item. The means for the items indicate how easy or difficult the prompts are.

Whole Booklet Feedback

Whole booklet feedback is based on the set of two prompts in the NAEP exam booklet that was administered to judges as part of the orientation process. It is important to point out the distinction between *paper* and *booklet*. *Paper* refers to a single student response to a single writing prompt. *Booklet* refers to the two prompts that comprise the writing test form, with both responses written by the same student. The whole booklet feedback reports the percent of total possible points that a student would need to earn for that set of prompts in order to meet the minimal requirements for performance at each achievement level set by the rating group in the round of ratings just completed. For example, the whole booklet feedback report might state: "Based on your group's average ratings, students performing at the Borderline Basic level are

expected to get 49% of the total possible score points for this booklet." A similar statement is given for each achievement level. Initially, this feedback is based on the cutpoints the group set during the first round of ratings, and it is updated after subsequent rounds.

Whole Booklet Exercise

As part of Round 1 feedback, the panelists participate in a whole booklet exercise, which is an extension of the whole booklet feedback. They are shown actual student booklets with scores around the cutpoints that were computed from round one ratings. Booklets scored within two percentage points above or below the cutpoint are evaluated by panelists. As an example of Borderline Basic performance, panelists might be shown a booklet for which the score is approximately 49% of the total possible points. They examine the responses of students to both writing tasks in the booklet and determine if the responses represent student performance expected at the lower borderline of each level. If they perceive a discrepancy between the expected performance and the observed performance in the booklets scored at the cutpoint, then they discuss the achievement levels descriptions and borderline performances again with other panelists in order to understand the cause for this discrepancy. Performance higher than expected signals that they set their cutpoints too high. Performance lower than expected signals that they set their cutpoints too low.

Because there are only two prompts in each writing test form, the distribution of total points for booklets is less continuous. ACT needed to slightly vary the usual procedure for selecting the booklets for the whole booklet exercise. Each of the two responses was scored 1, 2, 3, 4, 5, or 6. Instead of selecting booklets that were scored within two percentage points from the cutpoint as has been done in the past, booklets were used that were scored within as many as four percentage points from the cutpoints. For example, booklets scored as 46% could be selected to represent a score of 6 out of 12 points (50% - 4%) or a score of 5 out of 12 points (42% + 4%). Furthermore, the combination of prompt scores was considered in selecting the booklets. For example, three booklets were selected to represent the Borderline Proficient level (66%). All three booklets had a score of 8 points (67% of 12 total 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. There were no booklets scored as high as the Advanced level in the set of 100 randomly selected booklets used in the whole booklet exercise, so no booklets were available for panelists to review at the Advanced level.

ROUND 2 RATINGS (DAY 2)

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

FEEDBACK AFTER ROUND 2 (DAY 2)

Except for the Whole Booklet Exercise, panelists received the same forms of feedback after Round 2 that were presented after Round 1. The feedback was updated with Round 2 rating data.

STEP 4: CONSEQUENCES DATA (DAY 2)

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

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

STEP 5: EVALUATIONS THROUGHOUT THE PROCESS

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

OUTCOMES OF WRITING FIELD TRIAL 1

In general, panelists' reacted positively to the procedures implemented for the writing field trial. Neither the Mean Estimation nor the ISSE method posed any particular problem for panelists. Each rating group understood the rating method and how to apply it. They did, however, have difficulty reconciling the scoring rubrics with the scores they had seen for some student papers during the paper selection process. They also had problems with the scoring rubrics relative to the achievement levels descriptions. They expressed concern about the limited amount of time (25 minutes) allowed for student responses.

EVALUATION OF CUTPOINTS, STANDARD DEVIATIONS, AND RESULTING CONSEQUENCES DATA

Table 2 contains the cutpoints and their standard deviations computed from the two rounds of ISSE and Mean Estimation ratings. Computational procedures for producing all outcome data for this field trial are described in Chen & Loomis (2000). For comparison purposes, the percent of scores at or above the final cutpoints resulting from the 1992 ALS process also are included in Table 2. More complete analyses of cutpoints by type of panelist and type of writing have been included in Appendix F.

It should be noted that the cutpoints produced by the field trial panels were not entirely comparable with those produced by the 1992 ALS panels because of significant differences in the processes. For example, the 1992 ALS process used all of the items in the grade 8 assessment, including the 50-minute prompts. The field trial did not include the 50-minute prompts. The 1992 ALS process used the paper selection method, whereas the field trial used the ISSE and the Mean Estimation methods. The 1992 ALS process used NAGB policy definitions of achievement and descriptions of achievement that were different from those used for the field trial. The computational method to calculate cutscores was different for the two studies. The 1992 ALS process lasted five days while the field trial lasted two days. Panelists for the 1992 ALS process were drawn from nationally representative samples, whereas the panelists for the field trial were recruited locally. Because many elements of the processes varied, the cutpoints for the two studies are not wholly comparable.

ROUND 1

The Round 1 Basic cutpoints were quite low for both methods relative to the cutpoints for Proficient and Advanced. The Basic level cutpoint on the ACT NAEP-like score scale for panelists using the ISSE method (134.87) was lower than the Basic level cutpoint for panelists using the Mean Estimation method (147.31). The Advanced level cutpoint for panelists using the ISSE method (229.12) was higher than the Advanced level cutpoint for panelists using the Mean Estimation method (220.83). The Proficient level cutpoint for panelists using the ISSE method (177.81) was somewhat similar to the Proficient level cutpoint for panelists using the Mean Estimation method (184.15). The standard deviation of the cutpoints computed from the Mean Estimation ratings were noticeably higher for all three achievement levels (B=14.78; P=12.17; A=12.05) than the standard deviation of the cutpoints computed from the ISSE ratings (B=7.02; P=10.18; A=6.92).

ROUND 2

The cutpoint after Round 2 for Basic was raised by panelists using the ISSE method (137.15) and lowered by panelists using the Mean Estimation method (142.63). The cutscore for Proficient was lowered by panelists using both methods (ISSE=174.24; ME=176.39) and was similar after Round 2 for panelists using both methods. The cutscore for Advanced was lowered by panelists using both methods, but was considerable higher for the ISSE method (221.92) than for the Mean Estimation method (213.05). The standard deviations of the cutpoints computed for the ISSE ratings decreased for the Basic and Proficient levels, but increased for the Advanced level (B=1.82; P=5.9; A=10.52). The standard deviations of the cutpoints computed for the Mean Estimation ratings decreased for all levels, but remained relatively high overall (B=12.95; P=10.92; A=9.47).

FINAL

Panelists were given consequences data reporting the percentages of students scoring at or above each level of achievement. Consequences data were computed for each rating group and reported to each group separately. Panelists were asked to recommend new cutscores for any or all of the three achievement levels, if the panelists felt that modifications were in order. Those

recommendations were used to compute a final cutscore. The group cutscore for Round 2 was used for computing the average reported as the final cutscore for panelists who made no different recommendations. The final cutscore for the Basic level was lower for the ISSE group, the Proficient cutpoint was about the same for both groups, and the Advanced cutpoint was considerably higher for the ISSE group. The standard deviations were lower for the ISSE method than for the Mean Estimation method. Moreover, when contrasting the percentages of students performing at the different achievement levels, the percentages were higher for ISSE ratings than the Mean Estimation ratings at Basic and Proficient. Neither rating method resulted in even .01% students performing at the Advanced level. Since the NAEP ALS cutpoints have generally been criticized as being too high, a method that resulted in even higher cutpoints would not likely be selected, other things being equal.

Table 2
Comparison of the Outcomes from the ISSE Method and Mean Estimation Method for Writing Field Trial 1 Using 1992 NAEP Writing Grade 8 Data

Round	Achievement Level	Writing FT#1 ISSE		Writing FT#1 ME		1992 ALS Paper Selection
		Cutpoint* (SD)	%≥	Cutpoint (SD)	%≥	%≥
1	Basic	134.87 (7.02)	92.3	147.31 (14.78)	71.9	
	Proficient	177.81 (10.18)	5.0	184.15 (12.17)	1.6	
	Advanced	229.12.(6.92)	0.0	220.83 (12.05)	0.0	
2	Basic	137.15 (1.82)	89.9	142.63 (12.95)	82.0	
	Proficient	174.24 (5.9)	8.7	176.39 (10.92)	6.2	
	Advanced	221.92 (10.52)	0.0	213.05 (9.47)	0.0	
Final**	Basic	137.46 (0.96)	89.4	141.11 (7.65)	83.0	85.0
	Proficient	173.68 (1.38)	9.0	169.83 (6.46)	14.9	54.4
	Advanced	217.34 (5.23)	0.0	206.03 (7.41)	.009	0.1

* Cutpoints are reported on the ACT NAEP-like score scale.
**The data for the 1992 process are from Round 3.

EVALUATION OF INTERJUDGE CONSISTENCY

Interjudge consistency is an important criterion in judging the reasonableness of the achievement levels-setting process. The standard deviations of the cutscores are suitable indicators of interjudge consistency. In general, the standard deviations decrease across rounds of ratings, which indicates that the judges' ratings become more similar and less variable. This pattern occurred for both rating methods at each achievement level except the Advanced level for the group using the ISSE method. The standard deviations of the cutscores were fairly high because of the relatively small number of panelists and the few rounds of ratings.

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

Visually inspecting the rater location charts for the first writing field trial revealed different patterns between groups using the two methods. For Round 1, the cutscores for groups using the

Mean Estimation method tended to be clustered more closely together within the intervals than the cutscores from the ISSE method. For Round 2, the pattern of cutscores for panelists using the ISSE method was nearly the same as for Round 1. For panelists using the Mean Estimation method, the Basic and Proficient cutscores clustered even more closely together within the interval. In particular, all of the Basic cutscores fell within the range of 136-138 score points on the ACT NAEP-like scale. (See Table 3.) The range for Advanced cutscores set by panelists using the Mean Estimation method increased slightly for Round 2. These patterns are contrary to the more typical patterns observed for ALS results. The level of interjudge consistency tends to be lowest for the Basic level.

Table 3
Range of Individual Cutscores Displayed on Rater Location Charts for Field Trial 1

	Range of Cutscores	
	Round 1	Round 2
ISSE Method (n=8)		
Basic	132-161	130-165
Proficient	171-201	157-189
Advanced	209-239	195-222
Mean Estimation Method (n=7)		
Basic	133-141	136-138
Proficient	160-189	169-179
Advanced	217-234	208-234

EVALUATION OF INTRAJUDGE CONSISTENCY

Intrajudge consistency, both within rounds and across rounds, is generally regarded to be a reasonable criterion by which to judge a standard setting process (e.g., Berk, 1994). Indicators of intrajudge consistency across rounds include both the magnitude of change in item ratings from round to round, and the number of item ratings changed from round to round. If panelists make extreme adjustments to their ratings it would indicate that they had little confidence in their ratings. If panelists make changes of great magnitude to item ratings, or change a large number of item ratings after Round 2 for example, this would signal that panelists were confused or that they had not understood the process. Indicators of intrajudge consistency within rounds inform panelists about the consistency of their individual item ratings relative to their overall estimate of student performance at the cutscore.

INTRAJUDGE CONSISTENCY ACROSS ROUNDS

The writing field trial panelists exhibited "reasonable" intrajudge consistency across rounds of ratings. The percentages of items for which the ratings were changed from Round 1 to Round 2 can be used as an indicator of the consistency of a judge's ratings across rounds. After reviewing the Round 1 feedback, panelists were given the opportunity to change their ratings for Round 2. The changes have been reported as percentages of item rating changes in Table 4. Panelists had only two rounds of item ratings for this field trial (as opposed to three for an operational ALS process), so the across-rounds measure is somewhat less useful as an indicator of intrajudge consistency.

Many ratings remained unchanged. Panelists who used the Mean Estimation method changed a larger proportion of items than panelists who used the ISSE method. Perhaps this was because small adjustments (less than a score level) were possible for panelists using the Mean Estimation method. Panelists using both methods tended to lower their ratings more frequently than to raise them. This was especially true for Proficient and Advanced level ratings. The percentages of changes to lower the item ratings for the Mean Estimation group were considerably greater for all three levels than the total percentages of changes, per se, made by the ISSE group.

Table 4
Percentages of Changes in Item Ratings from Round 1 to Round 2 Using the Mean Estimation and ISSE Methods for Prompts in the Grade 8 NAEP Writing Pool

Method Used by n Panelists	Basic			Proficient			Advanced		
	Raise	No Change	Lower	Raise	No Change	Lower	Raise	No Change	Lower
ME n=8	7.9	50.8	41.3	0.0	44.4	55.6	3.2	42.9	54.0
ISSE n=7	19.4	68.1	12.5	9.7	69.5	20.8	1.4	77.8	20.8

Note: Percentages are based on 72 score estimates (9 prompts x 8 raters) for the Mean Estimation method group, and 63 score estimates (9 prompts x 7 raters) for the ISSE method group.

INTRAJUDGE CONSISTENCY WITHIN ROUNDS

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

One of the major attractions of the ISSE method was that it afforded ACT the opportunity to produce intrarater consistency feedback for panelists in a format that would be informative to each panelist, regardless of the panelist's level of intrarater consistency. One version of intrajudge consistency feedback has been included in Appendix G. Although panelists were not shown these data during the field trial, the sample format illustrates one way these data could be presented.

The sample shows how ratings for each panelist follow the mapping order of the prompts on the reporting score scale. In this example, scores on the ACT NAEP-like score scale are listed sequentially from highest to lowest. The format lists ("maps") the assessment items from the most difficult (highest score on the ACT NAEP-like scale) to the least difficult (lowest score on the ACT NAEP-like scale) with a brief description of the content of the item. Data reported for each item represent individual panelists' ratings for Basic, Proficient and Advanced levels.

By studying the chart, panelists could understand the relationship of their item ratings to each other, the relative level of item difficulty, and the ACT NAEP-like scale score associated with each**.** As the degree of item difficulty increased, the item rating pattern would show patterns to

reflect this increasing difficulty. Higher expected score ratings would cluster at the higher scale scores, and score ratings would decrease as the scale scores decreased.

To produce this type of intrajudge consistency feedback, each prompt in the rating pool was mapped onto the performance scale. The response probability criterion of 65% was used for mapping items. Each prompt was mapped to the ACT NAEP-like scale where a response probability of 65% was reached. The figures in Appendix G illustrate these data as intrajudge consistency feedback charts for each panelist. Figures 3a-3b were based on the ratings that each panelist provided in Round 1, and Figures 4a-4b were based on ratings for Round 2. Although ACT and the technical advisors reviewed the intrarater consistency feedback using this format as part of data analysis, these data were not used as feedback in the field trial.

After reviewing the intrarater consistency feedback using this format, the conclusion was that the writing field trial panelists exhibited "reasonable" intrajudge consistency within rounds of ratings. These data showed expected patterns of ratings that indicated that panelists understood the rating methods and could implement them successfully. In other words, ratings at the Advanced level of performance were higher than ratings at the Proficient and Basic levels. Ratings for more difficult prompts were generally lower than ratings for easier prompts. No patterns appeared that suggested panelists were confused or misunderstood the rating tasks.

Evaluation of Panelists' Comments and Responses to Process Evaluation Questionnaires

To better understand their perceptions of the rating processes, panelists were asked to respond to three process evaluation questionnaires. The complete set of responses to the process evaluation questionnaires has been included in Appendix H.

Given that the essential elements of the five-day ALS process were condensed into an intense two-day process for the field trial, it was reasonable to expect somewhat less positive responses from the field trial panelists than the ALS panelists. However, the overall responses of the field trial panelists were positive when evaluating the processes. When the study was completed, panelists asked many questions about the purpose of the study and they were willing to stay to engage in a general discussion of the rationale behind some procedures.

After Round 1, the responses of panelists using the Mean Estimation rating method were lower (less positive) for all questions than those of panelists using the ISSE rating method. After Round 2, the responses of panelists using the Mean Estimation rating method increased, becoming more positive for all questions than those of panelists using the ISSE method. These results indicate that practice and experience using the rating method had a greater impact on the panelists using the Mean Estimation method than on those using the ISSE method.

In general, there are indications that panelists who used the ISSE method were more satisfied with their experience than those who used the Mean Estimation method. When evaluating the rating process itself (i.e., understanding the rating task, instructions for rating prompts) both groups responded equally positively, indicating that instructions and training activities were evaluated as equally effective by both groups. *Conceptual clarity* of the rating process was

greater for panelists using the Mean Estimation method, while *ease of application* was greater for panelists using the ISSE method. On a scale of 1-5 (5=most positive), the average response for conceptual clarity was 4.14 for Mean Estimation panelists and 3.88 for ISSE panelists. The average responses regarding ease of application for ISSE panelists was 3.75 compared to 3.00 for Mean Estimation panelists. Questionnaire data indicate that *confidence* and *understanding of the process and tasks* increased for Mean Estimation panelists more so than for those using the ISSE method. Indeed, the conceptual clarity of the rating method increased from Round 1 to Round 2 for panelists using the Mean Estimation method, and the Mean Estimation method became easier to apply. The responses of ISSE panelists indicated no improvement by Round 2 in the conceptual clarity of the rating method, and their evaluation of the ease of applying the method indicated that it was even less easy to apply in Round 2 than in Round 1. However, when evaluating the overall ALS process (i.e., effectiveness of process) the ISSE group responded more positively than the Mean Estimation group.

When asked if they would be willing to sign a statement recommending the use of achievement levels resulting from this ALS procedure, the groups were noticeably different in their response. Two of the 8 members of the ISSE group responded negatively to the question, whereas 5 of the 7 members of the Mean Estimation group responded negatively.

Observations by the project staff suggest that the apparent differences in panelists' level of satisfaction might be due to the personality differences of panelists belonging to different groups. The panelists in the ISSE rating group required far less time to complete all training exercises, and they spent less time carrying out the rating process. There was no evidence to suggest that they took their work less seriously, however.

COGNITIVE COMPLEXITY OF RATING TASKS

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

The issue of cognitive complexity was included as an integral part of evaluating both rating methods. Responses to the evaluation questionnaires reflected the perceptions of panelists regarding the level of cognitive complexity required to perform the rating task, among other questions. Experience gained while practicing the rating method had a greater impact on the panelists using the Mean Estimation method than on those using the ISSE method. Because judges needed more practice to learn the Mean Estimation method, it could be considered more cognitively complex than the ISSE method. When asked about their level of understanding of the tasks they were to accomplish during the second rating session, both groups of panelists were highly positive (ME = 4.57; ISSE = 4.62). Both groups perceived their level of understanding to approach *totally adequate*. Although panelists learned the ISSE method more quickly than the

Mean Estimation method, with adequate practice panelists were able to learn the Mean Estimation method equally well. There was no evidence to suggest that panelists were unable to perform the estimation tasks involved in the process. None the less, when evaluating the overall ALS process, the ISSE group responded more positively than the Mean Estimation group.

EVALUATION OF PANELISTS' RESPONSES TO CONSEQUENCES DATA QUESTIONNAIRES

Panelists engaged in a lengthy group discussion about the consequences information. They seemed generally appalled by the consequences data and were very concerned about having no students scoring at the Advanced level. They expressed interest in lowering the Advanced cutpoints to a level where there would be students. Members of the ISSE group even asked how much they would have to lower the cutpoint to have 3% of the students scoring at the Advanced level. They soon realized that the score was considerably lower than their recommended cutpoint.

When asked whether the results reflected their expectations, only one panelist in the ISSE rating group and none in the Mean Estimation group said "yes." There were many fewer changes recommended by panelists in the ISSE group in response to the consequences data for Basic and Proficient than for the Mean Estimation group. Most panelists from both groups recommended lowering the Advanced cutpoint. Please refer to Table 5 for the frequencies of changes that were recommended to the cutpoints after reviewing consequences data.

Table 5
Frequencies of Recommended Changes to Cutpoints
After Reviewing Consequences Data

Recommended Cutpoints (n=15)	Basic		Proficient		Advanced	
	ME n=7	ISSE n=8	ME n=7	ISSE n=8	ME n=7	ISSE n=8
As Set	3	7	2	7	1	2
Lower	2	0	4	1	5	6
Higher	1	1	0	0	0	0
No Response	1	0	1	0	1	0

Panelists were asked to recommend a cutscore that would seem more acceptable than the cutscores computed from Round 2 ratings. New cutscores and consequences data were computed based on the recommendations collected after the first set of consequences data. These new data were distributed to panelists who were allowed to discuss the adjusted cutpoints and updated consequences data. They were asked to name the final cutscores that they would recommend to NAGB, if they had the opportunity. During the discussion, panelists expressed a general lack of confidence in the "arbitrariness" of the cutpoints computed on the basis of recommendations. There was a general preference for the cutpoints based on their ratings. However, when asked whether or not the revised, "final" percentages reflected their expectations, five panelists in the Mean Estimation group now said "no," and seven in the ISSE group said "no."

DETECTION OF BIAS IN ISSE CUTSCORES

The Technical Advisory Committee on Standard Setting (TACSS) reviewed the outcomes of the first writing field trial and expressed some reservation about the cutscores computed from the

ISSE ratings. The ISSE method was found to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true scores" or judgments of panelists (Reckase, 1998b; Bay, 1998; Reckase & Bay, 1999). The authors explained:

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

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

DEVELOPING A NEW RATING METHOD: THE RECKASE METHOD

Having eliminated the ISSE method, several new achievement levels-setting methods were considered for further study in a second writing field trial: the Grid method, the Booklet Classification method, and a rather new methodology developed by TACSS member Mark Reckase (Reckase, 1998a). In keeping with ACT's intent to explore new procedures for collecting and summarizing judgments, the Reckase method showed promise in meeting the unique challenges inherent to setting achievement levels for NAEP. "The method is particularly designed to fit with the scaling of the NAEP assessments and using the procedures typically used by NAGB to define the scores that set the boundaries for achievement levels." (Reckase, 1998a, p. 1).

The Reckase method addressed two major issues associated with setting NAEP achievement levels. The first issue related to providing a format for intrajudge consistency feedback. The Reckase charts were designed to meet the challenge of giving panelists information about the relationship of their ratings to item characteristics in a technically accurate and readily understandable way. The method greatly advanced efforts to inform panelists of the correspondence between their ratings and expected student performance. The second issue related to differences in cutscores by type of writing, e.g. narrative, persuasive, and descriptive. The design of the Reckase chart would also reveal any pattern of differences in ratings of the three kinds of writing included in the assessment. Like other methods considered by ACT for the 1998 ALS process, the Reckase method also involved a simple way for panelists to adjust cutscores for

their final round of ratings. TACSS reviewed the proposed Reckase method and recommended that ACT investigate its merits through continued field trial research.

KEY ASPECTS OF WRITING FIELD TRIAL 2

The second writing field trial was conducted at the Radisson Hotel, Highlander Plaza in Iowa City, Iowa on Wednesday and Thursday, July 8 and 9, 1998. The fundamental purpose of Field Trial 2 was to identify the procedures that would be used for the 1998 ALS process. The ISSE method had been eliminated from further consideration and ACT was committed to researching new methods to collect item-by-item ratings. For this field trial, ACT studied the Booklet Classification (BC) method and the new Reckase standard setting method (Reckase, 1998a) as alternatives to the Mean Estimation method. ACT also studied the effect of consequences data given to panelists during the item rating process.

The two methods that would be implemented were quite different. Panelists using the Booklet Classification method would classify booklets two times and then discuss consequences data and decide a final cutscore. Panelists using the Reckase method would participate in two rounds of item-by-item ratings before deciding their cutpoint for each level on the Reckase charts. Then they would discuss consequences data and decide a final cutscore. Appendix I shows the diagrams of the various research designs proposed for Field Trial 2.

The same criteria were used to evaluate and compare both methods. The methods were evaluated on the basis of judgements of the reasonableness of the cutpoints and their standard deviations; the level of interjudge and intrajudge consistency; the ease with which the method was implemented by panelists; and panelists' satisfaction with and confidence in the process and the results of the process. Logistic concerns and the reactions of panelists were particularly important criteria for evaluating the methods during this stage of the process. If either the Booklet Classification method or the Reckase method seemed as easy or easier than the Mean Estimation for panelists to understand, and if the results generally seemed reasonable—both to panelists and to ACT and our technical advisors—then the method that seemed better would be the choice of methods to implement in the pilot study and the writing ALS process.

PANELISTS

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

RECRUITMENT

The first effort to recruit panelists for the second writing field trial around the Iowa City area began in January of 1998. The response from educators was poor, primarily due to heavy teaching responsibilities. The second field trial was originally scheduled for Saturday and Sunday, March

14-15, but was postponed until the end of the school year when educators said they would be available to participate.

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

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

School district administrators and Area Education Agencies in Iowa, Wisconsin and Illinois were asked to identify outstanding teachers, educators, and qualified members of the general public as possible panelists. Elected officials in Johnson and surrounding counties were also asked to nominate panelists. Local business leaders were asked to nominate writers and journalists to serve as panelists.

Nominees were accepted as panelists if they fit the general description of "panelist type," were familiar with the knowledge and skills of eighth graders in the area of writing, and were willing to participate in the study. The nominees were confirmed as panelists as they were recruited until the goal was reached of 41 committed panelists. Later one panelist dropped out. About 70% of the 40 panelists were teachers (TR), 7.5% were educators who were not teaching in a K-12 classroom (NT), and 17.5% were knowledgeable members of the general public (GP). All of the panelists were white; 7% were males and 83% females. Please refer to Table 7 for more details about the composition of the panel.

Table 7
Composition of Writing Field Trial 2 Panel

	Type			Gender		Race	
Group	TR	NT	GP	Female	Male	White	Minority
A	7	1	2	8	2	10	0
B	8	1	1	9	1	10	0
C	8	0	2	8	2	10	0
D	7	1	2	8	2	10	0
N=40	30	3	7	33	7	40	0

TABLE DISCUSSION GROUPS

According to the study design (see Appendix I) panelists within each method group were further divided into two different groups: those who would receive consequences data after each round of ratings and those who would not. There were two table discussion groups of about five persons per table for each of the four treatment groups. The demographic attributes of panelists were considered when assigning members to treatment groups and to the table groups within each

treatment group. The goal was to have each group as equal as possible with respect to panelist type, gender and race.

DATA, ACHEIVEMENT LEVELS DESCRIPTIONS, AND ITEM RATING POOLS

NAEP Writing data from 1992 were used for field trial 2 because scaling of the Writing NAEP had not been completed and, consequently, 1998 writing data were not available. The problems ACT experienced with the writing assessment data in 1992 were of concern, however. The writing NAEP framework document had been revised somewhat, and the test specifications had been sharpened and tightened since administration of the 1992 assessment. ACT planned to use the achievement levels descriptions that had been developed for the 1998 Writing NAEP in the writing field trials using 1992 data. The correspondence, or lack thereof, between the 1992 scoring rubrics which were specific to each prompt, and the 1998 ALDs was not taken into account for the field trials. Although the 1992 NAEP writing data presented many limitations and restrictions, they were the best data available to use for the field trial.

The 1992 Writing NAEP for grade 8 consisted of 11 prompts in all. Nine of those were 25-minute prompts and 2 were 50-minute prompts. Since only 25-minute prompts were to be used in reporting the 1998 NAEP writing results, only the nine 25-minute prompts from the 1992 NAEP for grade 8 were used in the field trial.

In past ALS processes conducted by ACT, items used during training exercises were excluded from the item rating pools used for setting the achievement levels. In this way the potential for contaminating the ratings was eliminated. However, only nine prompts could be used for Field Trial 2. Since each panelist rated each prompt, it was necessary that some prompts be used in the training exercises that were used for item ratings in setting the achievement levels. All but one prompt was used in training raters during the second writing field trial.

THE ALS PROCESS DESIGNED FOR WRITING FIELD TRIAL 2

The recommended design (see Appendix I) divided panelists into four groups (A-D) composed of ten members each. Groups A and B used the Booklet Classification method (BC). Groups C and D used the Reckase method. Groups A and C were given consequences data after each round of ratings or classifications, whereas groups B and D were first given consequences data after the final ratings or classifications. The Booklet Classification method groups classified booklets two times and then discussed consequences data, whereas the Reckase method groups completed two rounds of item-by-item ratings before deciding their cutpoint for each level and marking it on the Reckase charts.

IMPLEMENTING THE RECKASE METHOD

The Reckase method used the Mean Estimation (ME) item-by-item rating procedures for the first round of ratings. (Please see Field Trial 1 for a detailed explanation of the Mean Estimation rating procedures.) Those estimates were then transferred onto charts, now called Reckase charts. Columns on the Reckase charts displayed the IRT-based performance estimates for each assessment prompt in the rating pool, and rows displayed IRT-based performance estimates for

each score point on the ACT NAEP-like scale. The performance estimate was the expected score for each prompt at each scale score. The expected performance across scale score points could be observed, as could the expected performance across prompts for students scoring at a particular scale score. A sample Reckase chart[6] with accompanying instructions has been included as Appendix J.

As Reckase initially designed the method, panelists marked the charts by circling the expected performance score that corresponded to their item-by-item ratings for the first round of judgments. They then connected the circles to see their rating patterns graphically. Panelists also marked their own cutscores at each achievement level and the group cutscore for each level. By examining the charts, the judges were able to consider the relationship between their estimates of student performance for each prompt and expected student performance at the cutscore. Further, judges considered any observable patterns in their ratings, such as different estimates for different types of writing (narrative, persuasive, and descriptive). Panelists were instructed that if their judgments of students performing at the borderline of each achievement level exactly fit the estimates generated by the model based on actual student performance, all of their ratings would fall along a single row. In other words, if panelists' ratings were on a single row, their ratings perfectly matched estimated student performance.

Panelists evaluated their ratings with respect to different types of prompts, their own cutscores, and the cutscores set by the rating group. They considered this information along with other types of feedback when rating prompts a second time. After the second round of ratings, panelists again transferred their ratings onto the charts along with the group cutscore for each level and their own cutscores. After examining and evaluating their Round 2 ratings using the Reckase charts and other forms of feedback, panelists engaged in a third rating session.

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

IMPLEMENTING THE BOOKLET CLASSIFICATION METHOD

Each panelist using the Booklet Classification method classified 40 booklets. There were 20 booklets from each of two different forms. Each panelist evaluated forms that included at least one prompt for each type of writing. In order to provide panelists the opportunity to discuss booklet classifications after Round 1, the design provided 10 booklets in each form to be classified by two panelists who would be seated together. Thus, each panelist had 10 booklets of form A to discuss with panelists A (seated on the right) and 10 booklets of form B to discuss with panelist B (seated on the left).

[6] Because student performance on NAEP is secure information, the data displayed on the sample Reckase Chart are fictitious.

The booklets were ordered on writing performance scores from lowest to highest. Panelists were told that the ordering was to facilitate their classification task and that it represented only one of many such orderings that might be used. They were told that their classifications did not have to reflect the ordering because classifications were to be made on the basis of the achievement levels descriptions. Panelists classified booklets into one of seven categories:

- Below Basic
- Borderline Basic
- Basic
- Borderline Proficient
- Proficient
- Borderline Advanced
- Advanced

An elaborate system was developed for selecting the booklets that would be classified using the Booklet Classification (BC) method. The procedure for selecting booklets was repeated for both Booklet Classification groups. NAEP forms were assigned Booklet Classification identification numbers from 1 to 10 (see Appendix K, Table 1). The order of prompt types was balanced across the forms so that narrative, informative, and persuasive prompts appeared as the first or second prompt in the form about the same number of times. Booklet forms also were evenly distributed by rank, as determined by theta values. The sum of the rankings across each set of 20 booklets was approximately equal. (See Appendix K, Table 2).

Every panelist (except panelist #5 in groups A & B) reviewed four distinct prompts, at least one from each of the three types of writing. The fourth prompt type was evenly distributed across the three types of writing. To arrange for panelists to read duplicate booklets so that they could later discuss their classifications, 30 booklets for each form were organized into 3 groups of 10 booklets each. The three groups of booklets were marked X, Y, and Z (see Appendix K, Table 3).

To illustrate the system for assigning booklets, Panelist Three (P3) will serve as an example. Panelist Three classified 20 booklets from Form Three (F3), and 20 booklets from Form Four (F4). Within F3, Panelist Three classified booklet groups Y and Z. Within F4, Panelist Three classified booklet groups X and Y. The same procedures were applied when assigning booklets to Panelists Two (P2) and Panelist Four (P4). As a result of this assignment, P3 would have in common with P2 booklets Y from F3, and with P4 booklets Y from F4. In this way, every panelist was able to discuss 20 duplicate booklets with two other panelists, 10 booklets for each partner.

STEP 1: BRIEFING MATERIALS

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

- 1994 NAEP *Writing Framework*;
- 1998 NAEP Writing Achievement Levels Descriptions;
- *Multiple Challenges*, a booklet on the 1998 NAEP;
- NAGB brochure;

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

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

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

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

TAKING A FORM OF THE NAEP

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

UNDERSTANDING THE ACHIEVEMENT LEVELS DESCRIPTIONS (ALDS)

Panelists were given time to work with the ALDs and to discuss them, but they were not given the opportunity to revise them. The final version of the 1998 Writing NAEP achievement levels descriptions (ALDs) for grade 8 was used for the field trial. Panelists participated in exercises to help them become familiar with ALDs. They examined the scoring rubric and a writing task for each type of writing. Panelists were asked to think about the performance required for students to score at each rubric point (for the 1992 NAEP) with respect to the descriptions of writing skills associated with the ALDs (for the 1998 NAEP). Panelists were asked to decide which score was most likely for students performing at each level of achievement.

UNDERSTANDING BORDERLINE PERFORMANCE

Panelists also received brief training in the concept of borderline performance. They did not write borderline descriptions as part of their training, however.[7] Panelists discussed borderline

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

performance and reached a common understanding of what constituted borderline performance at each achievement level.

PAPER CLASSIFICATION TRAINING EXERCISE

Both groups of panelists practiced working with the ALDs and continued developing their concept of borderline performance using a paper classification training exercise. Each panelist reviewed a packet of 51 single student papers written in response to three prompts, one from each type of writing. There were 18 papers each for the narrative and informative prompts, and 15 for the persuasive prompts. Panelists then reviewed 10 student booklets, each contained writing by the same student in response to two different types of prompts. The narrative and informative prompts each were scored from 1 to 6; the persuasive prompt was scored from 1 to 5.

This exercise was modified for the writing field trial to help panelists using the Booklet Classification method to start thinking in classification terms. The exercise also helped focus panelists on the holistic aspects of writing by evaluating responses to two prompts written by the same student. Panelists were asked to classify papers and booklets at one of the three achievement levels, or at the Below Basic level, according to their understanding of the ALDs. Further, they were asked to determine which papers and booklets represented performance at the lower borderline of each achievement level.

STEP 3: THE RATING PROCESS

After orientation, panelists were separated into the two methods groups. Half of the panelists (groups A and B) were trained in the Booklet Classification (BC) method, and groups C and D were trained in the Reckase method (MR). Please refer to Appendix K for detailed instructions to panelists on the Booklet Classification method, and to Appendix J for instructions on the Reckase method.

ROUND 1 (END OF DAY 1)

Booklet Classification Method

Once trained, panelists classified 40 booklets into one of seven categories (Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient, Borderline Advanced, Advanced) based on student performance and the ALDs. After classifying all of the booklets, panelists completed a process evaluation questionnaire. That completed Day 1 activities.

Reckase Method

Panelists using the Reckase method (groups C and D) were trained in the Mean Estimation (ME) method for Round 1 ratings. Panelists were required to estimate the average scores (e.g., 2.25 on a scale of 1-6) of students performing at the borderline of each achievement level, using up to two decimal places. Panelists read each prompt, considered what constituted a response at the borderline of Basic, Proficient, and Advanced levels, checked the scoring rubrics, and marked

their rating forms. After rating all prompts in the pool, panelists completed a process evaluation questionnaire. That completed Day 1 activities.

ROUND 2 (DAY 2)

Booklet Classification Method

After reviewing and discussing the results and feedback from Round 1 classifications, groups A and B were reconvened to discuss their classifications of common booklets (10 booklets for each of two forms), and to evaluate and adjust their booklet classifications for Round 2. The panelists using the Booklet Classification method remained separate from the panelists using the Reckase method for the discussion. The panelists classified the same booklets a second time using the same methodology. They could change all, some or none of their classifications for any or all achievement levels. Booklet classifications by individual panelists and analyses of various elements of the classifications appear in Appendix L.

Reckase Method

After reviewing and discussing the results and feedback from Round 1 ratings, Groups C and D were moved to another room and were instructed in the Reckase method. Panelists were instructed to evaluate their ratings, considering different types of writing, the cutscores set by their group, and the cutscores set by their own ratings. As part of their instruction, panelists were shown how to mark their Round 1 ratings for each prompt on the Reckase charts. The Reckase charts used to review Round 1 ratings did not include scale score values. Instead, letters of the alphabet were used to label the rows. This was done because of concerns that panelists would place too much emphasis on the cutscore and not enough emphasis on the achievement level descriptions as their guide for ratings.

Panelists using the Reckase method rated the same items a second time using the same methodology. That is, panelists estimated the average score expected for students performing at the borderline of each achievement level. They could change all, some or none of their ratings for any or all achievement levels.

ROUND 3 (END OF DAY 2)

Booklet Classification Method

Groups A and B did not classify booklets a third time.

Reckase Method

Groups C and D marked their Round 2 ratings and cutscores on the Reckase charts. They selected a single row (i.e., scale score) to represent their ratings for each item. They were instructed to select a row that represented the expected performances at that cutscore for each prompt. Panelists were instructed to draw a horizontal line through the row at the cutscore. One row was selected to represent the Basic cutscore, one the Proficient, and one the Advanced. The IRT-based

expected performance estimates for each item associated with each cutscore were, in effect, their Round 3 ratings. The Reckase charts for Round 3 included the numeric scale score values, which replaced the alphametric values used during Round 2.

STEP 4: FEEDBACK AFTER EACH ROUND

FEEDBACK AFTER ROUND 1 (BEGINNING OF DAY 2)

At the start of Day 2, panelists were instructed in the various forms of feedback data based on their Round 1 ratings and classifications. Four sets of feedback data were provided—one for each method by consequences data treatment group. Panelists received this information together in a general session at the start of the day. They were provided with cutpoints and standard deviations, rater location charts, student performance data, and whole booklet feedback prior to the second round of ratings and classifications. Please refer to the definitions of terms included in the description of Field Trial 1. Panelists did not participate in the Whole Booklet Exercise.[8] Data were not shared across treatment groups. Copies of the feedback based on the first round of ratings and classifications are included as Appendix M.

Reckase Charts

Panelists using the Reckase method were given Reckase charts, as described previously. They evaluated their ratings on the charts to determine whether they could detect ratings and patterns of ratings that needed further consideration. They were specifically asked to examine the charts to determine whether one type of writing had been rated as more or less difficult than their overall borderline performance estimate. They were instructed to re-examine item ratings that differed sharply from their other ratings or from their overall cutscores.

FEEDBACK AFTER ROUND 2 (DAY 2)

After the second round of ratings and classifications, all panelists were given updated feedback data using the cutpoints set from the second round of ratings and classifications. They received this information in separate groups. Judges were provided with cutpoints and their standard deviations, rater location charts, student performance data, and whole booklet feedback.

FEEDBACK AFTER ROUND 3 (END OF DAY 2)

After Round 3, groups C and D received the standard forms of feedback that had been updated using the cutpoints marked on the Reckase charts. Cutscores were computed by averaging the cutpoints recommended by each panelist for each level.

[8] The booklets for the exercise were not shipped to the site of the field trial. By the time this was discovered, it was too late to retrieve them. The Project Director decided to cancel the Whole Booklet Exercise scheduled for Day 2.

STEP 5: CONSEQUENCES DATA (DAY 2)

Although consequences are considered a type of feedback, the timing of when panelists received consequences data was an experimental condition that defined treatment groups.

CONSEQUENCES DATA AFTER EACH ROUND

Round 1

Following the general session when panelists were instructed in the various forms of feedback, groups A and C were moved to another room where they were given information about the consequences of their ratings and classifications for Round 1. That is, panelists were told the percentages of students scoring at or above the cutpoint set for each achievement level based on the first round of ratings and classifications. After the consequences data were explained, panelists completed a questionnaire in which they were given the opportunity to recommend whether cutpoints should be changed to raise or lower the percentage of students performing at or above each level. A copy of the questionnaire appears in Appendix E.

In a separate room groups B and D reviewed and discussed the results and feedback from their Round 1 ratings and classifications. They did not, however, receive consequences data after Round 1.

Round 2

Because there were only two rounds of classifications, groups A and B received consequences data together at this time. Group D did not receive consequences data until after Round 3. The consequences data were updated to reflect the rating group's cutscores based on Round 2 ratings and classifications. They were given the chance to discuss this information within their group and then asked to complete a questionnaire in which they recommended new cutpoints that would raise or lower the percentages of students performing at or above each achievement level. A copy of the questionnaire appears in Appendix E.

Round 3

All panelists using the Reckase method received consequences data after Round 3. Group C received consequences data for the third time, and group D received consequences data with instructions for the first time.

Because groups A and B did not classify booklets a third time, there were no Round 3 cutpoints and consequences data for those groups.

DISCUSSION OF CONSEQUENCES DATA AFTER FINAL ROUNDS OF RATINGS AND CLASSIFICATIONS

After the consequences data from the final rounds were explained, panelists from groups A and B discussed this information and were encouraged to reach agreement on a group cutscore for each

achievement level. Panelists from groups C and D discussed group cutscores separately from groups A and B.

Panelists were asked to complete a consequences questionnaire. On this questionnaire, they had the opportunity to recommend new cutpoints that would raise or lower the percentages of students performing at or above each achievement level. All panelists were given this consequences questionnaire to record their reactions to and recommendations regarding the consequences associated with their final round cutpoints. A copy of the questionnaire has been included in Appendix E. New cutscores were computed, on the basis of these recommendations.

RECOMMENDED CUTSCORES

The new group cutscores were discussed separately by members of the different methods groups. Panelists were encouraged to reach common agreement on a final set of cutpoints that represent each group's recommendation to NAGB.

CONSEQUENCES DATA FEEDBACK AFTER DISCUSSION OF RECOMMENDED CUTSCORES

After the discussions, each panelist filled out a final consequences questionnaire regarding the consequences data and the final cutscores he/she would recommend to NAGB. Group C completed their fourth consequences questionnaire; group A completed their third; and groups B and D completed their second. Round 2 classifications were used to compute cutscores for panelists in groups A and B who recommended no changes in their cutscores. Round 3 cutscores were used to compute cutscores for panelists in groups C and D who recommended no changes in their cutscores. The individual cutscores marked on the final consequences data questionnaires were averaged to compute the final cutpoints for each of the four groups. These were the recommended cutpoints.

STEP 6: EVALUATIONS THROUGHOUT THE PROCESS

Panelists completed process evaluation questionnaires throughout the meeting. Groups A and B answered four evaluation questionnaires regarding their reactions to the Booklet Classification method, and groups C and D answered five questionnaires about the Reckase method. Responses to the questionnaires have been summarized and appear later in this report. When the last evaluation was finished, the panels were thanked for their work and the meeting was adjourned.

OUTCOMES OF WRITING FIELD TRIAL 2

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

Panelists generally were receptive to each method. They found the Reckase charts informative and interesting and seemed to have no problems with them. Having the booklets ordered seemed to simplify the classification task for panelists using the Booklet Classification method. Rather than first placing booklets in categories of achievement, as had been observed in NAEP ALS validation studies, panelists simply wrote their classifications on the booklets and on their "rating" form. They did not classify booklets strictly according to the ordering. That is, they classified some booklets from lower ranks at higher levels than other booklets that were ranked about the same. The classification task did not appear to be very challenging to panelists. ACT carefully estimated the pace that panelists would need to maintain and notified panelists of expected progress at regular intervals. These reminders seemed to help keep panelists focused on the holistic nature of the classification procedure and deterred their tendency to score the booklets.

EVALUATION OF CUTSCORES, STANDARD DEVIATIONS, AND RESULTING CONSEQUENCES DATA

The cutpoints, their standard deviations, and the percentages of students performing at or above each achievement level have been included in Table 8. Computational procedures for producing all outcome data for this field trial are described in Chen & Loomis (2000). The standard deviations are presented in line graphs in Figure 1. Those figures show standard deviations for each rating group across each round.

ROUND 1

Round 1 cutscores for the Basic level were quite low for all groups, relative to outcomes of other ALS studies. That is, the percentages of students scoring at or above the Basic cutscores set by the four rating groups were all quite high. The Round 1 cutscores for the Advanced level were quite high for all groups (few students had scores as high as the Advanced cutscore), and this has been an outcome of previous ALS studies. The Round 1 cutscores for the Basic and Proficient levels were *about* the same for groups A, B, and D. Group C set theirs considerably higher. In general, the Round 1 cutscores and the variability (SD) of the cutscores for the groups using the Reckase method were higher than those for the groups using the Booklet Classification method. The higher variability was especially evident for the standard deviations (SD) of the Round 1 cutpoints for the Basic and Proficient levels. Over all groups, the variability among panelists' cutscores (indicated by the SD) tended to be high at the Advanced level.

ROUND 2

Group C (Reckase method receiving consequences data) lowered all three cutpoints for Round 2. They lowered the Proficient and Advanced cutpoints considerably. The other Reckase group raised all three cutscores for round 2. Round 2 ratings by panelists in the Booklet Classification groups resulted in rather moderate changes in the cutpoints. Some were raised and some were lowered.

The standard deviations of several cutscores indicate even more variability among panelists' cutscores in Round 2 than in Round 1. The standard deviations increased for cutscores at all three achievement levels set by panelists who used the Reckase method and received consequences

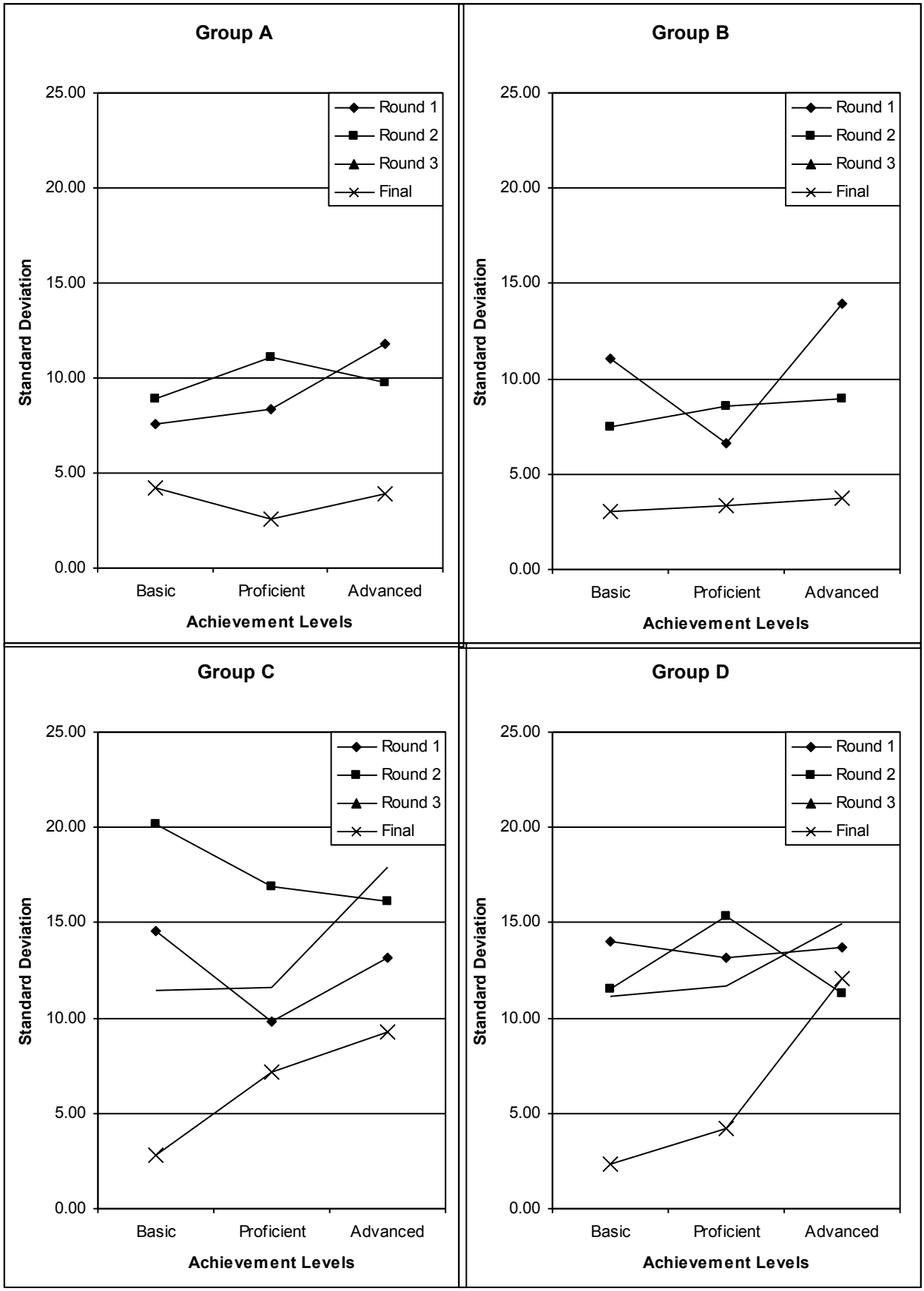
data after each round of ratings. The standard deviation of the Proficient level cutscore increased from Round 1 to Round 2 for all four groups. For each of the two method groups, the standard deviation of cutscores was higher for the group that received consequences data before Round 2 than for the other group.

Table 8
Comparison of the Outcomes from the Four Experimental Groups for Writing Field Trial 2

		Booklet Classification Method				Reckase Method			
		A (n=10) Early Consequences		B (n=10) Late Consequences		C (n=10) Early Consequences		D (n=10) Late Consequences	
Round	Level	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>	Cutpoint (SD)	%=>
1	Basic	128.1 (7.57)	96.9	129.7 (11.08)	***96.1***	138.5 (14.57)	97.8	123.1 (14.00)	***98.7***
	Proficient	157.1 (8.35)	44.8	160.6 (6.62)	***34.9***	179.4 (9.84)	3.6	164.6 (13.17)	***25.5***
	Advanced	194.5 (11.83)	0.2	187.1 (13.97)	***0.8***	217.5 (13.13)	0.0	211.3 (13.70)	***0.0***
2	Basic	131.6 (8.90)	94.9	131.5 (7.50)	***94.9***	134.1 (20.19)	92.9	124.8 (11.49)	***98.4***
	Proficient	156.8 (11.07)	44.8	154.3 (8.60)	***52.3***	167.5 (16.90)	19.0	166.3 (15.34)	***21.4***
	Advanced	196.5 (9.73)	0.1	190.0 (8.94)	***0.4***	202.9 (16.10)	0.0	213.5 (11.28)	***0.0***
3	Basic					136.2 (11.44)	90.7	122.6 (11.10)	98.8
	Proficient					171.5 (11.64)	11.9	165.7 (11.66)	22.3
	Advanced					208.6 (17.93)	0.0	215.7 (14.94)	0.0
Final	Basic	133.8 (4.19)	93.2	131.8 (3.06)	94.7	137.3 (2.78)	89.9	125.4 (2.36)	98.1
	Proficient	157.6 (2.59)	42.8	156.3 (3.34)	47.0	164.9 (7.16)	24.8	154.4 (4.22)	53.4
	Advanced	191.8 (3.92)	0.3	187.0 (3.77)	0.9	198.6 (9.25)	0.0	201.3 (12.05)	0.0

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

Figure 1: Standard Deviations of Cutpoints Set by Different Groups on Different Rounds



ROUND 3

Groups A and B did not classify booklets a third time. For Round 3, panelists using the Reckase method marked a cutscore on the Reckase charts. The scale scores for each panelist were averaged to produce the group cutscores. Panelists using the Reckase method and receiving consequences data after each round raised their cutscores for all three levels at Round 3. Panelists using the Reckase method but not receiving consequences data before Round 3 lowered the Basic cutscore slightly, kept the Proficient cutscore at about the same level, and raised the Advanced cutscore slightly. Standard deviations decreased for the Basic and Proficient cutscores for both Reckase method groups. These decreases were rather sharp for raters in group C.

FINAL

The general pattern of change in cutpoints for the final round was to raise the Basic cutpoint slightly and lower the Advanced cutpoint considerably. Both groups of panelists using the Booklet Classification method raised the Proficient cutpoint, and both groups of panelists using the Reckase method lowered the Proficient cutpoint. While all groups lowered the cutpoint for the final Advanced cutpoint, the Reckase group receiving consequences data for the first time after Round 3 lowered the Proficient cutpoint by more than 11 points on the ACT NAEP-like score scale and increased the percentage of students scoring at or above the cutpoint from 22% to 53%.

In general, panelists using the Booklet Classification method set cutpoints across the rounds that did not differ significantly by the timing of consequences feedback data. Panelists using the Booklet Classification method and receiving consequences data throughout the process set slightly higher cutscores than those receiving consequences data after the ratings/classifications were completed. On the other hand, panelists using the Reckase method and receiving consequences feedback data throughout the process generally set higher cutscores. At the Advanced level, cutscores set by panelists using the Reckase method were higher than those set by panelists using the Booklet Classification method, no matter when consequences data were introduced. Descriptive statistics of the results are reported in greater detail in Appendix N.

Of interest are the cutscores set for the Basic and Proficient levels, which seemed unusually low using both methods. The Advanced cutscore was quite high. The more extreme cutpoints could have been the result of using the 1992 Writing NAEP data with the 1998 ALDs. Perhaps the low cutpoints were due to the fact that the Basic level writing ALDs describe rather minimal performance. Panelists were asked to estimate borderline performance, and that would be lower still.

EVALUATION OF CLASSIFICATIONS RATING DATA

The percentages of items for which ratings and classifications were changed were not formally analyzed for field trial 2 due to the lack of time for extensive analyses. Rather, other indicators were used to gain an overall sense of the success of the two methods.

ACT staff inspected rater location charts to determine whether individual panelists shifted their cutscores in directions and magnitudes that appeared to be logical and reasonable. In general, the

changes in cutpoints made by most panelists were judged as indicating that they were responding to feedback in a logical manner. This was not true for all panelists, of course, but it was the prevailing pattern.

The rater location charts showed different patterns for the two methods. For Round 1, panelists using the Booklet Classification method tended to set cutscores that were spread across the ACT NAEP-like scale. For groups using the Reckase method, the Basic and Proficient cutscores followed a similar pattern. For the Advanced cutscores, however, the pattern was different. Most of the cutscores were clustered together at the high end of the scale, and there was little spread. (Please see Table 11.)

Table 11
Range of Individual Cutscores Displayed on Rater Location Charts for Writing Field Trial 2

	Range of Cutscores	
	Round 1	Round 2
Booklet Classification Method		
(With Consequences Data)		
Basic	118-143	120-143
Proficient	144-170	146-185
Advanced	169-208	180-209
Booklet Classification Method		
(Without Consequences Data)		
Basic	115-152	121-148
Proficient	147-167	134-167
Advanced	154-203	170-202
Reckase Method		
(With Consequences Data)		
Basic	114-163	119-162
Proficient	167-197	130-190
Advanced	205-220	170-219
Reckase Method		
(With Consequences Data)		
Basic	110-144	117-137
Proficient	140-180	140-197
Advanced	193-220	193-220

For Round 2, the panelists using the Booklet Classification method tended to set cutscores that were closer together within the intervals than their Round 1 cutscores. Interjudge consistency generally increased for panelists using the Booklet Classification method after discussing Round 1 classifications of booklets. Panelists using the Reckase method, however, continued to set cutscores that were spread across the intervals, including the Advanced cutscores. For Round 3, the Reckase panelists set cutscores that were closer together within the interval of each achievement level. The range of cutscores for Round 3 group B is large because of only one relatively low location for the Proficient and Advanced levels.

ACT staff prepared a chart to facilitate visual inspection of Booklet Classification data to determine whether panelists appeared to simply use the booklet ordering as the guide to

classifications. Those data are presented in Appendix L and clearly indicate that panelists were using something other than the performance ordering of booklets to make their classifications.

Finally, Reckase charts were visually examined from round to round to determine whether panelists seemed to be moving their ratings of student performance for each prompt to be closer to expected student performance at the cutscores based on their own ratings, and based on the ratings of the grade group.

All of these inspections suggested that panelists were generally able to use the methods and that they were making logical adjustments from round to round that were consistent with the feedback given to them.

EVALUATION OF PANELISTS' COMMENTS AND RESPONSES TO PROCESS EVALUATION QUESTIONNAIRES

To better understand panelists' perception of the rating process, they were asked to respond to process evaluation questionnaires, using Likert-type scale items. The questionnaires were administered throughout the meeting. The complete set of responses to the process evaluation questionnaires has been included in Appendix O.

THE METHODS AND PROCEDURES

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

In general, responses to questions evaluating the rating and classification processes, (e.g., understanding the rating task, instructions for classifying booklets) for all four groups were equally positive by Round 3. It is noteworthy that the Reckase group responded more positively when asked about the ease of applying the rating method than the Booklet Classification group when asked about the ease of classifying booklets. However, one of the Reckase groups responded less positively when asked whether the ALS process produced reasonable and defensible achievement levels. The same group responded less positively when asked whether the ALS process provided them an opportunity to use their best judgment in rating prompts. Panelists using the Booklet Classification method responded more positively than those using the Reckase method when asked if they would be willing to sign a statement recommending the use of achievement levels resulting from their ALS procedure. Two of 19 panelists from the Booklet Classification group responded negatively, whereas 6 of 18 panelists from the Reckase group responded negatively.

There was concern that the Reckase method held the potential for being too "data driven" and that the final cutpoints would be based on Reckase chart data rather than the standards represented by the ALDs. Panelists' reactions to the charts did not support this concern, however. When asked if

the Reckase charts influenced their judgments during the third rating session, 5 of the 20 panelists reported "greatly," 6 responded "to a modest degree" and 9 chose a level somewhere between the two. Please refer to Appendix O for a complete account of the panelists' responses and comments.

COGNITIVE COMPLEXITY OF PROCEDURES

The issue of cognitive complexity was included as an integral part of evaluating the ALS procedures. Responses to process evaluation questionnaires reflected the perceptions of panelists regarding the level of cognitive complexity required to perform the rating and classification tasks. By the final round, 37 of the 40 panelists indicated that they understood the tasks they were to accomplish during the session. Fifteen panelists from the Booklet Classification group and 11 from the Reckase group responded that they felt their level of understanding was "totally adequate." They did not appear to have difficulty performing any of the tasks involved in the process. Panelists reported their overall satisfaction with the rating and classification methods and their confidence in the results produced by the methods. Responses indicated that they were comfortable with the level of cognitive complexity required of them.

EVALUATION OF PANELISTS' RESPONSES TO CONSEQUENCES DATA QUESTIONNAIRES

Panelists had lengthy discussions about consequences data. Comments generally related to the fact that essentially no students performed at or above the Advanced level. None of the panelists seemed particularly concerned about the high percentage of students at or above the Basic level. To review panelists' comments about the impact of consequences data on their cutscores, please refer to Appendix O. Please see Table 12 for a summary of the consequences data.

Some panelists using the Booklet Classification method expressed confusion about the consequences data, relative to booklet classifications. Perhaps the confusion resulted from the fact that they were classifying booklets into categories. They felt that since they had classified more than 1% of their booklets at the Borderline Advanced and Advanced levels, more students should be at these levels. They reasoned that "a real student wrote each booklet," and they expected some students in the Advanced level. They were reminded several times that their 40 booklets did not reflect the national distribution of student performance, and their comments suggested that they tried to keep this in mind. Still, they had difficulty reconciling the consequences data with their classifications. When asked to recommend final cutpoints, the panelists tended to recommend percentages within levels rather than percentages at or above levels or actual cutpoints, as requested on the consequences questionnaires.

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

	Early Consequences Data		Late Consequences Data	
	Group A (BC) n=10	Group C(MR) n=10	Group B (BC) n=10	Group D (MR) n=10
Basic				
Round 1	96.9	97.8	96.1	98.7
Round 2	94.9	92.9	94.9	98.4
Round 3	NA	90.7	NA	98.8
Final	93.2	89.9	94.7	98.1
Proficient				
Round 1	44.8	3.6	34.9	25.5
Round 2	44.8	19.0	52.3	21.4
Round 3	NA	11.9	NA	22.3
Final	42.8	24.8	47.0	53.4
Advanced				
Round 1	0.2	0.0	0.8	0.0
Round 2	0.1	0.0	0.4	0.0
Round 3	NA	0.0	NA	0.0
Final	0.3	0.0	0.9	0.0

Changes Made to Cutscores in Response to Consequences Data

Table 13 displays the number of changes panelists made to their cutscores in response to consequences data. Panelists were still recommending changes to cutscores on the final consequences data questionnaire, particularly at the Advanced level. The majority of panelists in all but group C recommended that the Advanced cutpoint be lowered in the final consequences data questionnaire. Panelists in groups A and D tended to recommend a higher cutpoint for Basic, whereas panelists in groups B and C were satisfied with the Basic cutpoint as set. The majority of panelists in all but group A recommended that the Proficient cutpoint remain as set. Judges in group A recommended a higher Proficient cutpoint.

Table 13
Frequencies of Recommended Changes to Cutscores After Receiving Consequences Data by Group
Writing Field Trial 2

Round	Level	Group A (BC)				Group B (BC)				Group C (MR)				Group D (MR)			
		No Change	Lower	Higher	No Response	No Change	Lower	Higher	No Response	No Change	Lower	Higher	No Response	No Change	Lower	Higher	No Response
1	Basic	5	1	4	0	N/A				3	4	2	1	N/A			
	Proficient	5	5	0	0					0	9	0	1				
	Advanced	3	6	1	0					0	9	0	1				
2	Basic	6	0	4	0	6	3	1	0	3	0	7	0	N/A			
	Proficient	9	0	1	0	7	0	3	0	4	3	3	0				
	Advanced	3	7	0	0	5	4	1	0	4	5	1	0				
3	Basic	N/A				N/A				7	0	3	0	1	0	9	0
	Proficient									3	7	0	0	0	10	0	0
	Advanced									1	9	0	0	1	9	0	0
Final	Basic	1	1	8	0	10	0	0	0	10	0	0	0	6	0	4	0
	Proficient	1	2	7	0	7	3	0	0	9	1	0	0	7	3	0	0
	Advanced	1	9	0	0	4	6	0	0	6	4	0	0	0	10	0	0

SUMMARY OF FIELD TRIALS RESEARCH FINDINGS

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

WRITING FIELD TRIAL 1

The purpose of the first writing field trial was to compare the Item Score String Estimation (ISSE) rating method with the Mean Estimation (ME) method. Results of the first field trial indicated that panelists who used the ISSE method were more satisfied with the process than those who used the Mean Estimation method. The project staff observed differences between the groups, that were likely to have caused differences in the level of satisfaction between them. The ISSE method resulted in higher Advanced cutscores and lower Basic cutscores. Nearly 90% of students performed at or above the Basic level using the cutscores set by the ISSE method, in comparison with 83% for the Mean Estimation method. All panelists expressed concern about having no students scoring at the Advanced level. Follow-up analysis of the ISSE method found it to be biased in such a way that cutscores were higher for the Advanced level and lower for the Basic level when compared with the "true score" or "true" judgment of the panelist (Reckase & Bay, 1999). Because of this flaw, further research using the ISSE method was discontinued, and it was eliminated as a possibility for implementation in the writing ALS.

WRITING FIELD TRIAL 2

The fundamental purpose of the second field trial was to identify the method that would be used for the 1998 ALS process. ACT studied the new Reckase standard setting method, the Booklet Classification method, and the timing of consequences data on the outcomes of the process. Results of Field Trial 2 indicated that panelists had no difficulty using either the Booklet Classification method or the Reckase method.

It is notable that cutscores set for the Basic and Proficient levels seemed unusually low for both methods and those for the Advanced level were quite high. In general, panelists using the Booklet Classification method set cutpoints that did not differ significantly by the timing of consequences feedback data. On the other hand, panelists using the Reckase method and receiving consequences feedback data throughout the process generally set higher cutscores. At the Advanced level, cutscores set by Reckase panelists were higher than those set by Booklet Classification panelists, no matter when consequences data were introduced. This finding was rather surprising in that previous studies have all indicated that cutscores would be higher with the Booklet Classification method than with the Mean Estimation method, used since 1994 in the NAEP ALS process. This was the only time a Booklet Classification procedure was implemented with ordered booklets. Panelists knew that the last booklets had the highest scores. Perhaps this made it easier for panelists to judge the performance as Advanced. In earlier Booklet Classification studies, panelists frequently classified all booklets below the Advanced level. Those studies generally indicated that the cutscores would be higher at all levels with the Booklet Classification method (ACT, 1995; ACT, 1997b).

Panelists tended to focus on the low percentage of students at or above the Advanced level. Some panelists who used the Booklet Classification method expressed confusion between the booklets they classified and data reported as consequences feedback. Because of the concern regarding alternative computational procedures for the Booklet Classification method, the Reckase method was judged to be more promising for use in the writing ALS process.

FINDINGS REGARDING ISSUES IN NAEP STANDARD SETTING

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

THE ISSUE OF COGNITIVE COMPLEXITY

The charge has been made that item-by-item rating methods can not produce valid cutpoints because panelists are incapable of performing the cognitively complex task of estimating probabilities with reasonable accuracy (NAE, 1993; Shepard, 1995; Impara and Plake, 1998). ACT has collected considerable data during the writing field trials and previous ALS research where panelists have reported their capacity to perform the tasks associated with estimating student performance. Judges perceive that they are performing the estimation and judgmental tasks required by the item-by-item rating methods with relative ease. They report that they are

confident in their judgments and satisfied with the results. There is no evidence to indicate that panelists are unable to make these judgments.

THE ISSUE OF INTRAJUDGE CONSISTENCY ACROSS ROUNDS

One of the considerations for evaluating the various experimental methods explored by the field trials was based in part on indicators of "reasonable" intrajudge consistency across rounds. The ALS process designed by ACT provided panelists with extensive feedback and instructions for interpreting information when forming their judgments of student performance. Panelists were encouraged to reconsider their ratings and classifications and adjust them according to their interpretation of the many sources of information available to them. Judges were expected to adjust their ratings and classifications from round to round. It was reasoned that if panelists understood the method and the feedback produced by the method, they would adjust their ratings or classifications from round to round. If panelists did not adjust their ratings or classifications at all, this was an indicator that they probably did not understand the method or the feedback. On the other hand, if they changed all—or most—of their ratings or classifications after two rounds, this was also an indicator that they probably did not understand the method or the feedback. Statistical analyses of rating changes for field trial 1 and visual inspections of the results for field trial 2 led to the overall judgment that writing field trial panelists exhibited "reasonable" intrajudge consistency across rounds. Because the purposes of field trials were directed at issues of feasibility, rather than technical precision, this rather general judgment seemed adequate in the context.

THE ISSUE OF INTRAJUDGE CONSISTENCY WITHIN ROUNDS (RECKASE CHARTS)

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 relative to their overall ratings and cutscores. Although this information had been given to panelists in previous ALS meetings, there was no evidence to suggest that panelists either understood the information, or found it useful when forming their judgments about student performance. The Reckase method greatly advanced the effort to give panelists precise item-level information they could readily understand about ratings, relative to expected student performance at their individual cutscores and at the grade-level cutscores. Panelists who used the Reckase method in general adjusted their ratings or classifications for Round 2 to be more similar to the IRT-based performance estimates of students at the cutscores. This finding was consistent for all three achievement levels. It is important to note that for the third rating session panelists were instructed to select a single row or scale score to represent their ratings for each achievement level. There was concern that this procedure could overly influence judgments and that the final cutpoints would be based on Reckase charts data rather than the standards represented by the ALDs. Panelists' reactions to the charts did not support this concern, however. Panelists indicated that they were influenced by the achievement levels descriptions and other forms of feedback in addition to the charts when forming their final judgment of student performance.

THE ISSUE OF THE TIMING OF CONSEQUENCES DATA

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

THE ISSUE OF COMPUTING CUTSCORES FOR THE BOOKLET CLASSIFICATION METHOD

Of particular interest in the second writing field trial was the computational procedure used to calculate cutscores for the Booklet Classification method. Recall that each panelist classified a set of booklets into one of seven categories: Below Basic, Borderline Basic, Basic, Borderline Proficient, Proficient, Borderline Advanced, and Advanced. The cutscores were based on a weighted average that gave more weight to booklets classified at the borderline than to those classified within the interval of each achievement level.

At the request of TACSS, ACT computed two sets of cutpoints for each panelist. One set of cutpoints was computed using just the booklets classified at the lower borderline of each achievement level, and another set was computed using all booklets classified within the interval. In other words, the Borderline Basic booklets were combined with Basic booklets, Borderline Proficient with Proficient, and Borderline Advanced with Advanced.

The two sets of cutscores differed greatly, although theoretically the cutscores should have been essentially equivalent. Further, the cutscores computed for "borderline" versus "collapsed groups" resulted in very different percentages of students at or above each of the three achievement levels. As an alternative to this procedure, a cubic regression analysis was used to compute cutscores for the booklet classifications. Plake and Hambleton (1998) had worked with this method and found the results to be satisfactory.

TACSS concluded that given the considerable effort required to administer the Booklet Classification method, there should be strong support for the method to merit its implementation. Because TACSS members were uneasy with the alternative methods for computing the cutpoints for the Booklet Classification method, they recommend that it not be used for the next phase of planned research. Computing stable cutscores using the Booklet Classification method for NAEP remains an unresolved issue.

PLANNING FOR THE WRITING PILOT STUDY

The Reckase charts, introduced in the second field trial for writing, were judged to be a promising addition to the ALS process designed for NAEP. The charts appeared to have improved and strengthened the ALS process. Computing cutscores for the Booklet Classification method proved to be problematic, resulting in unstable cutpoints for the Booklet Classification method. After

reviewing the results of both the civics and writing field trials, it was agreed that research would continue on the Reckase charts. TACSS recommended using a modified version of the Reckase method for further research in the writing pilot study. The Reckase method, per se, was eliminated but the Reckase charts would be used in the pilot studies, with the expectation that they would also be used for the ALS. Panelists in future studies would omit selecting a single row on the chart to represent student performance at the cutscore for each achievement level. The Reckase charts would be presented to panelists as another form of feedback. This modified version of the Reckase method would be implemented in the ALS for writing, unless findings in the pilot study suggested otherwise. Finally, the modified Reckase method had been recommended for the civics ALS process, and this decision gave more weight to the choice of the same method for writing as well.

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

National Academy of Education (1993). *Setting performance standards for student achievement*, Robert Glaser, Robert Linn, and George Bohrnstedt, eds. Panel on the Evaluation of the NAEP Trial State Assessment. Stanford, CA: Author.

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

Plake, B.S. & Hambleton, R.K. (1998, April). *A standard setting method designed for complex performance assessments with multiple performance categories: Categorical assignments of student work*. Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

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

Reckase, M.D. (1998b). Converting boundaries between National Assessment Governing Board performance categories to points on the National Assessment of Educational Progress score scale: The 1996 science NAEP process. *Applied Measurement in Education* , *11,* (1): 9-21.

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.

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

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