# Developing Achievement Levels on the 1998 NAEP in Civics and Writing: Technical Report

Wen-Hung Chen
American Institutes for Research

Susan Cooper Loomis
and
Teri Fisher
ACT, Inc.

December 2000

# Developing Achievement Levels on the 1998 NAEP in Civics and Writing: Technical Report

Wen-Hung Chen
American Institutes for Research

Susan Cooper Loomis
and
Teri Fisher
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


Developing Achievement Levels on the 1998 NAEP in Civics and Writing: Technical Report ....1
Computational Procedures Used in Field Trials, Pilot Studies, and the Operational Achievement Levels-Setting Studies for the 1998 NAEP in Civics and Writing ....1
- Computational procedures Using Modified Angoff and Mean Estimation Methods....1
  - Computing Cutpoints for Multiple-Choice Items....1
  - Computing Cutpoints for Constructed Response Items....2
  - Information Weighting to Combine Multiple Choice and Constructed Response Item Ratings ....3
  - Transformation to the ACT NAEP-Like Score Scale ....4
- Computational procedure Using Item Score String Estimation (ISSE) Method....4
- Computational procedure Using an Item Mapping Method ....7
  - θ Location for the Multiple-Choice Items....7
  - θ Location for the Constructed Response Items....8
- Computational procedure for the Reckase Method ....9
- Computational procedure for Booklet Classification Method....11
  - Collapsed Categories Method ....11
  - Average Borderline Method....12
  - Weighted Collapsed and Borderline Method....12
  - Cubic Regression Method....12

Computational procedures for Feedback Used in Studies for the 1998 NAEP Achievement Levels-Setting Process in Civics and Writing....13
- Cutpoints and Standard Deviations ....13
- Rater Location Feedback....14
  - Computational procedure for Individual Cutpoint Using Modified Angoff and Mean Estimation Methods .....14
  - Computational procedure for Individual Cutpoints Using ISSE Methods ....15
  - Computational procedure for Individual Cutpoints Using the Item Mapping, Reckase, and Booklet Classification Methods....15
- Wholebooklet Feedback....15
- Consequence Feedback ....16
- Intrarater Consistency Feedback ....17
- Selection of Exemplar Items for the 1998 NAEP Civics and Writing Assessments....17
- Data Entry and Quality Control Procedures ....20
- The Main Computational Program....21
- Drawing Panels for the Pilot and ALS Studies....23
- Sampling of the School Districts....23
- Sampling of the Private Schools....25
- Sampling of the Panelist from the Pool of Nominees....25

References ....27
Key Recommendations by the Technical Advisory Committee on Standard Setting....39
Sampling Plan and Recruitment of Panelists ....40
Focus Groups for Review and Evaluation of Preliminary ALDs....40
Format for Recommendations to NAGB ....41
Scaling Procedures ETS Plans to Use for 1998 Writing NAEP ....41
Research Studies Included to Identify Methodology and Computational Procedures ....42
- Simulation Study....42
- Field Trial 1 for Civics and Writing....42
- Field Trial 1 for Writing....43
- Field Trial 2 for Civics and Writing....43
- Recommendations for Field Trial 2 for Civics ....44

Study Procedures ..... 44
Details for Implementing Reckase Charts ..... 46
Marking the Charts ..... 46
Facilitators' Instructions for the Reckase Method ..... 46
Field Trial 2 for Writing ..... 46
Study Procedures ..... 48
Booklet Classification Implementation Details ..... 49
ISSE Update ..... 50
Computing Cutscores with Booklet Classification Method ..... 50
Recommend ACT Add Field Trial 3 ..... 50
Recommended Procedures for Civics Pilot Study ..... 51
Recommended Procedures for Writing Pilot Study ..... 51
Recommendations for Civics ALS Process ..... 53
Recommendations for Writing ALS Process ..... 54
ACT Recommendation to NAGB that Consequences Data be Introduced Earlier ..... 55
General Discussion about Civics ALS ..... 55
TACSS Comments about Civics ALS Analyses ..... 56
TACSS Comments about Writing ALS Analyses ..... 56
TACSS Comments about Writing ALS Questionnaire Analyses ..... 57
Additional Analyses of ALS Data ..... 57
Follow-up Analyses for Civics ALS and Writing ALS ..... 57
TACSS Recommendations for Including or Rejecting Validations Studies ..... 58
Comparison of State Writing Standards to NAEP Study ..... 58
Congruence of Achievement Levels Descriptions for Civics ..... 59
Selecting Exemplar Items Study ..... 59
Effects of Motivation on Achievement Levels Study ..... 60
Standard Errors of Cutscores ..... 60
Performance Profiles ..... 61
Similarities Classification Study ..... 61
Comparison of 1992 and 1998 NAEP Writing Achievement Levels ..... 63
Plausible Values Study of Domain Coherence ..... 63
Person-Fit Statistics Study ..... 64
Civics Item Classification Study (CICS) ..... 64
Research Studies Related to the Redesign of NAEP and to International Benchmarking ..... 64
ALS Process Using TIMSS Items ..... 65
TIMSS-NAEP Linking Feasibility Study ..... 65
Rescale NAEP Using TIMSS Model ..... 66
International ALS for Writing ..... 66
Use of NAEP Short-Forms and Use of Domain Scores ..... 66
Using Domain Scores in Reporting Achievement Levels: Matt Schulz ..... 67
Examining Standard Setting ..... 67
The Evolution of the Achievement Levels-Setting Process ..... 68
Discussion of Response Probability Values for ALS Process and NAEP Reporting ..... 68
TACSS Research Proposals ..... 70
Standards-Based NAEP Score Reporting: Ron Hambleton ..... 70
Discussion of Achievement Levels Issues that Have Been Raised Recently ..... 71
Setting Additional Achievement Levels Below the Proficient Level ..... 71
Reporting Achievement Levels when Multi-Dimentional Scaling is Used ..... 71
Reporting Achievement Levels Data in Percent Correct Metric ..... 71
Changes in NAGB Policy Definitions for Achievement Levels ..... 72

# DEVELOPING ACHIEVEMENT LEVELS ON THE 1998 NAEP IN CIVICS AND WRITING: TECHNICAL REPORT

Wen-Hung Chen
American Institutes for Research
and
Susan Cooper Loomis
ACT, Inc.

## COMPUTATIONAL PROCEDURES USED IN FIELD TRIALS, PILOT STUDIES, AND THE OPERATIONAL ACHIEVEMENT LEVELS-SETTING STUDIES FOR THE 1998 NAEP IN CIVICS AND WRITING

### COMPUTATIONAL PROCEDURES USING MODIFIED ANGOFF AND MEAN ESTIMATION METHODS

#### COMPUTING CUTPOINTS FOR MULTIPLE-CHOICE ITEMS

ACT used the modified Angoff method for collecting panelists' judgments of student performance on multiple-choice items for 1998 National Assessment of Educational Progress (NAEP) in Civics during the second field trail, the pilot study, and the operational Achievement Levels-Setting (ALS) procedures. The panelists estimated the probability of a correct response by students performing at the borderline of each achievement level for each NAEP item in the assessment. Each panelist gave three ratings for each item, one for each of the three achievement levels Basic, Proficient, and Advanced.

Panelists were divided into two rating groups that were as equivalent as possible in terms of panelists' characteristics and demographics. Items were also divided into two rating pools that were as equivalent as possible in terms of item characteristics and item difficulty. (For a complete description of the process, see Loomis & Hanick, 2000e and 2000f.) Within each rating group, the average rating was calculated for each item. The sum of the average ratings was then used to project the overall cutpoint for the multiple-choice items onto the θ scale. To obtain the overall cutpoint for the multiple-choice items in group g on the θ scale, one must find the $\theta_g$ that satisfies the following equality:

$$\sum_{i=1}^{n_{gm}} \frac{\sum_{j=1}^{N_g} r_{ij}}{N_g} = \sum_{i=1}^{n_{gm}} p_i(\theta_{gm}) \qquad (1)$$

Where, $n_{gm}$ is the number of multiple-choice items to be rated by group $g$, $N_g$ is the number of panelists in group $g$, $r_{ij}$ is the rating of panelist $j$ for item $i$, $p_i$ is the IRT model predicted probability of a correct response for item $i$ given $\theta_{gm}$, and the IRT item parameter estimates. In NAEP, the 3PL model was chosen for the multiple-choice items. The 3PL model gives the probability of a correct response to item $i$ as,

$$p_i(\theta_{gm}) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta_{gm} - b_i)}} \qquad (2)$$

Where, $D$ equals 1.7, $a_i$ is the discrimination parameter, $b_i$ is the difficulty parameter, and $c_i$ is the guessing parameter.[1]

The right hand side of Equation 1 produces the true score for item $i$ at ability level $\theta_{gm}$. Essentially, the computational procedure is based on an assumption that the panelists' average ratings were the true score, and the task was to find the $\theta$ that yielded that true score. Because equation 1 cannot be directly solved, an interpolation procedure was applied to find the $\theta$. The interpolation procedure involved the arbitrary selection of a value of $\theta$ as the starting value, the use of IRT modeling to calculate the probability of that value, and a comparison of that value with the average of the ratings by each panelist. If the probability based on the IRT model was less than the average ratings for the panelist, then the theta value was increased. If the probability was greater than the average ratings for the panelist, then the theta value was decreased. This continued until the difference between the average ratings and the probability was less than the fixed criterion of .001.

Because the panelists gave three ratings for each item—one for each achievement level, Equation 1 was used to yield three overall cutpoints for the multiple-choice items for each group. These overall cutpoints were combined with the overall cutpoint at each achievement level for the constructed response items to form the overall cutpoints for each level within each grade.

### COMPUTING CUTPOINTS FOR CONSTRUCTED RESPONSE ITEMS

For constructed response items, the mean estimation method was used to collect judgments of panelists regarding the performance of students at the borderline of each achievement level. This method was used for the second field trial, the pilot study, and the 1998 Civics NAEP ALS process. It was also used during the first field trial, the pilot study, and the ALS process for the 1998 Writing NAEP. For each constructed response item in their rating pool, panelists estimated the mean score expected for students performing at the borderline of each achievement level. Each panelist gave three ratings for each item, one for each of the three achievement levels. Within each rating group, the average rating was calculated for each item. The sum of the average ratings was then used as the basis for projecting the overall cutpoint for the constructed response items onto the $\theta$ scale. The overall cutpoint for the constructed response items in group g on the $\theta$ scale is given by solving Equation 3 for $\theta_{gc}$.

$$\sum_{i=1}^{n_{gc}} \frac{\sum_{j=1}^{N_g} r_{ij}}{N_g} = \sum_{i=1}^{n_{gc}} \sum_{k=1}^{m_i} k p_{ik}(\theta_{gc}) \qquad (3)$$

Where, $n_{gc}$ is the number of constructed response items assigned to group $g$, $m_i$ is the number of response categories for item $i$, and $p_{ik}$ is the probability of obtaining response $m$ given $\theta_{gc.}$ In NAEP, the Generalized Partial Credit (GPC) model was chosen for the constructed response items. The GPC model is written as follows.

[1] The item parameters were provided by Educational Testing Service (ETS), operations contractors to the National Center for Educational Statistics.

$$p_{ik}(\theta_{gc}) = \frac{e^{\sum_{r=1}^{k} Da_i(\theta - b_i + d_{ir})}}{\sum_{u=1}^{m_i} e^{\sum_{r=1}^{u} Da_i(\theta - b_i + d_{ir})}} \quad (4)$$

Where, $a_i$ is the discrimination parameter, $b_i$ is the location parameter, and $d_{ir}$ is the threshold parameter for response category *r*. The threshold parameter is interpreted as the relative difficulty of getting response category *r* in comparison with other response categories. Again, Equation 3 cannot be directly solved, so an interpolation method was applied to determine the value of theta.

Equations 3 yielded three overall cutpoints for each group. These overall cutpoints for constructed response items were combined with the overall cutpoint for multiple-choice items to form the overall cutpoints for each achievement level within each grade.

### INFORMATION WEIGHTING TO COMBINE MULTIPLE CHOICE AND CONSTRUCTED RESPONSE ITEM RATINGS

Given the differences in characteristics of the multiple-choice and constructed response items, it was necessary to develop a means of combining ratings for the two types of item formats to form an overall composite cutscore for each level of achievement. In 1992, the recommendation of the Technical Advisory Committee on Standard Setting (TACSS) was to use information weighting. Using item information as the weights was thought to be reasonable because it took into account the measurement error at each θ level. The higher the level of item information is at a given θ level, the smaller the measurement error for the θ estimate. In effect, the group specific cutpoints of the combined multiple-choice and constructed response items were calculated as,

$$\theta_g = \frac{I_{gm}\theta_{gm} + I_{gc}\theta_{gc}}{I_{gm} + I_{gc}} \quad (5)$$

Where, $\theta_g$ is the combined cutpoint, and $I_{gm}$ and $I_{gc}$ are the information weights for multiple-choice and constructed response items, respectively. The information weight for the multiple-choice items is calculated as the sum of the information of individual items that were rated, given $\theta_{gm}$. The formula is,

$$I_{gm} = \sum_{i=1}^{n_{gm}} \frac{D^2 a_i^2 (1 - p_i)^2 (p_i^* q_i^*)^2}{p_i q_i} \quad (6)$$

Where, $p_i$ is the probability predicted by the 3PL model given $\theta_{gm}$, as in Equation 2, $p_i^*$ is the probability predicted by the 2PL model given by,

$$p_i^*(\theta_{gm}) = \frac{1}{1 + e^{-Da_i(\theta_{gm} - b_i)}}, \quad (7)$$

and $q_i$ and $q_i^*$ are $1 - p_i$ and $1 - p_i^*$, respectively.

For constructed response items, the information weights were calculated as,

$$I_{gc} = \sum_{i=1}^{n_{gc}} D^2 a_i^2 \left[ \sum_{r=1}^{m_i} r^2 p_{ir} - \left( \sum_{r=1}^{m_i} r p_{ir} \right)^2 \right] \quad (8)$$

Where, $p_{ir}$ is the probability predicted by GPC model for category $r$ of item $i$ given $\theta_{gc}$, as in Equation 4.

Equation 5 gave the three overall combined (multiple-choice and construct response items) cutpoints for a rating group. The overall cutpoints for the grade were computed as the simple average of the two rating groups of panelists. The three grade specific cutpoints were calculated as,

$$\theta_{overall} = \frac{\sum_{g=1}^{M} \theta_g}{M}, \quad (9)$$

where, $M$ is the number of groups of panelists. In the NAGB/ACT ALS process, two groups of panelists are used. In this situation, the standard deviation of the overall grade-level cutpoints were computed as,

$$STD_{\theta_{overall}} = \frac{|\theta_{g_1} - \theta_{g_2}|}{2} \quad (10)$$

### TRANSFORMATION TO THE ACT NAEP-LIKE SCORE SCALE

In the ACT/NAGB ALS process[2], all scale score data are reported on the ACT NAEP-like score scale. Since 1994, this scale has been set with a mean of 155 and standard deviation of 14.

The actual NAEP score scale is not used for several reasons. Using the ACT NAEP-like scores protects the security of the cutscores developed during the ALS process. The ALS process involves over 100 persons, and it seems highly probably that at least one of those persons might inadvertently "release" this information if data were reported on the NAEP score scale. A second advantage of using the ACT NAEP-like score scale is that it helps to maintain the focus on a criterion-referenced ALS process. With so many persons involved in the process of setting achievement levels for a given subject, it seems likely that someone would be familiar with achievement levels previously set in other subjects. If Civics ALS panelists knew about the cutscores set in US history or in geography, for example, they might judge the outcomes of the Civics ALS process relative to those other subjects rather than relative to the achievement levels descriptions for Civics.

The ACT NAEP-like scale is a simple linear transformation from the NAEP score scale, and these are both a linear function of θ. Please see Figure 1 for example cutpoints resulting from the ACT/NAGB method.

### COMPUTATIONAL PROCEDURE USING ITEM SCORE STRING ESTIMATION (ISSE) METHOD

During the first field trials for the 1998 NAEP ALS in Civics and Writing, an item score string estimation (ISSE) method was tested out. The method required the panelists to estimate the score for an examinee

[2] This name was recommended to distinguish the entire methodology of rating items and feedback used for setting NAEP achievement levels from the methodology more typically associated with a modified Angoff method (Reckase, 2000).

performing at the borderline of each achievement level. Panelists estimated whether the students performing at the borderline of an achievement level would obtain a score of 1 (correct) or 0 (incorrect) on multiple-choice items, and whether their score on constructed response items would be 1, 2, 3, 4, and so forth. This method was tried out because it was assumed that the task required for the panelists was easier than the modified Angoff and mean estimation method for item ratings. The panelists were to make discrete estimates rather than estimates on a continuous scale. The method yielded a panelist's ratings that resembled a student's response pattern to the items in NAEP. Panelists were essentially asked to estimate the response pattern that would match the performance of students at the borderline of an achievement level. For the ISSE method, the ratings cannot be assumed to represent the IRT true score, and the computational procedure was also different.

The use of the IRT-based techniques with the ISSE rating method has these positive attributes:
1. The discrete ratings appear to be easier than the probability ratings for the panelists to produce.
2. Item weights are determined by the item parameters implicitly.
3. It is robust with small sample sizes (Chen & Pommerich, 1998).

Each panelist gave an estimated response pattern (item score string), and the task was to use these ratings to estimate the overall cutpoint for the grade. The computational procedure estimated the mean of the population distribution from the panelists' item score string estimates. Notice that this population is the population distribution of cutpoints, $\tau$, not the population of $\theta$. The estimated mean of the population distribution is then the overall cutpoint, $\theta_{overall}$. The computational procedure for the mean of the $\tau$ distribution was described by Sanathanan and Blumenthal (1978) for the Rasch model, for the NAEP, we extended the procedure to the 3PL and GPC models.

First, assuming the overall cutpoint ($\theta_{overall}$) and its associated standard deviation ($STD_{\theta overall}$) are given by the population mean ($\mu$) and standard deviation ($\sigma$). The maximum likelihood estimates of $\mu$ and $\sigma$ are found by maximizing, with respect to $\mu$ and $\sigma$, the marginal likelihood based on the response patterns,

$$f(R;\mu,\sigma) = \int L(R \mid \tau)\, p(\tau;\mu,\sigma) d\tau\,, \tag{11}$$

Where $R$ is the data matrix of response patterns, $L(R|\tau)$ is the likelihood of observing $R$ given $\tau$, and $p(\tau;\mu,\sigma)$ is the density function of $\tau$ given $\mu$ and $\sigma$. When $\tau$ is assumed to be normally distributed, $\Sigma\tau_j$ and $\Sigma\tau_j^2$ are the sufficient statistics for $\mu$ and $\sigma$, where $\tau_j$ is the cutpoint set by judge $j$. Thus, in order to estimate $\mu$ and $\sigma$ we need to compute the expected values of $\tau_j$ and $\tau_j^2$. These are given by

$$E(\tau_j;\mu_{(k-1)},\sigma_{(k-1)}) = \int \tau g_j(\tau) d\tau\,, \tag{12}$$

and,

$$E(\tau_j^2;\mu_{(k-1)},\sigma_{(k-1)}) = \int \tau^2 g_j(\tau) d\tau\,, \tag{13}$$

where $\mu_{(k-1)}$ and $\sigma_{(k-1)}$ are the estimates of $\mu$ and $\sigma$ at iteration ($k$-1), and $g_j(\tau)$ is

$$g_j(\tau) = \frac{L(R_j \mid \tau)\, p(\tau;\mu,\sigma)}{\int L(R_j \mid \tau)\, p(\tau;\mu,\sigma) d\tau}. \tag{14}$$

Because one does not actually observe $\tau$, the EM algorithm is applied in an iterative process until the estimates of $\mu$ and $\sigma$ become stable. The iterations stop when the change in the estimates from the

previous to the current iteration is less than 0.001. One technical note is that since we assume $\tau$ is normally distributed, the density function of $\tau$, $p(\tau;\mu,\sigma)$ is simply the normal density function with parameters $\mu$ and $\sigma$. Another technical note: the likelihood function $L(R|\tau)$ is where the IRT model comes into play. Specifically, the likelihood function is,

$$L(R_j \mid \tau) = \prod_{i=1}^{n} \left( \prod_{r=1}^{m_i} p(x_{ij} = r \mid \tau)^{y_{ijr}} \right). \quad (15)$$

where $R_j$ is the response pattern by panelist $j$, $m_{i.}$ is the number of response categories for item $i$, and $y_{ijr}$ is a dummy variable whose value is 1 if $x_{ij}$ equals $r$ or 0, otherwise.

Once the expected values of $\tau_j$ and $\tau_j^2$ were obtained, the estimates of $\mu$ and $\sigma$, at iteration ($t$), were obtained via,

$$\mu_{(t)} = \frac{1}{N} \sum_{j=1}^{N} E(\tau_j; \mu_{(t-1)}, \sigma_{(t-1)}), \quad (16)$$

and,

$$\sigma_{(t)} = \sqrt{\frac{\sum_{j=1}^{N} E(\tau^2; \mu_{(t-1)}, \sigma_{(t-1)})}{N} - (\mu_{(t-1)})^2}. \quad (17)$$

After the estimation of $\mu$ and $\sigma$ converged, their estimated values were used for the overall cutpoints, $\theta_{overall}$, and their associated standard deviations, $STD_{\theta overall}$.

Although the panelist were divided into two groups and were assigned different item rating pools to ease the task of rating items, the simultaneous estimation of the population $\mu$ and $\sigma$ enabled us to treat the panelists as a single group. Thus, there was no need to compute the average cutpoints for the two rating groups. Finally, the $\theta_{overall}$ was transformed to the ACT NAEP-like scale with mean of 155 and standard deviation of 14.

The field test data suggested that the ISSE method resulted in more extreme cutpoints than the Mean Estimation method. This result is possibly due to the reduced numerical precision of the panelists' ratings—the reduction of information. The computation using the ISSE method is time-consuming and complex. It would be more difficult to explain to panelists.

TACSS reviewed the data. They expressed reservations and concerns regarding results produced by the method. Reckase and Bay (1999) conducted a theoretical experiment that indicated bias in the method. As a result, the method was not recommended for the 1998 NAEP ALS process. The findings raised a question of trade-off between complexity of rating method and loss of information in the ratings. Please see Figure 2 for example cutpoints resulting from the ISSE method.

So far, the three methods discussed all belong to the family of item-by-item rating methods. Another family of rating methods includes the holistic methods, some of which have gained popularity in recent years. ACT tried out some of the holistic rating methods in field trials for the 1998 NAEP ALS, and those are discussed next.

## COMPUTATIONAL PROCEDURE USING AN ITEM MAPPING METHOD

An Item Mapping method for setting cutpoints was tested during round 3 of the second field trial for the 1998 Civics NAEP. The panelists were given a list of items ordered on the measure of difficulty according to the relative location of the item difficulty on the θ scale. The panelists' task was to decide on a cutpoint for each achievement level, based on their judgements regarding the items and the feedback from the previous two rounds of item-by-item ratings. The items were presented both on a list with a short description of the item and on a large graph showing the score on the ACT NAEP-like scale associated with performance on each item. Panelists were to determine where each cutpoint should be and to mark the boundary between performance at adjacent achievement levels. With this method, each panelist gave only three cutpoint estimates, one for each achievement level. The overall group cutpoint was simply the numerical average of the individual cutpoints.

### θ LOCATION FOR THE MULTIPLE-CHOICE ITEMS

Items are located on the score scale according to a mapping criterion. A probability of correct response had to be decided upon to serve as the mapping criterion. Determining this was not trivial. The value selected largely determines the cutpoint. Once the criterion was decided and the locations of the items were set, the panelists made their judgments regarding the cutpoints. In the Bookmark Method used by CTB (Lewis, Mitzel, & Green, 1996) panelists put a mark between items to demarcate the boundary between two levels. Once that marker is placed, the numerical value of the cutscore is determined by the probability value used to locate items on the score scale.

The item mapping method is somewhat like the reverse of the NAGB/ACT method of item-by-item rating methods. In the NAGB/ACT method, panelists estimate the probability of correct response or mean score for students at the borderline of an achievement level. In the item mapping, that probability is decided for them. That is, once the panelist sets a cutpoint, it means that any examinee with ability at or above the cutpoint has at least the criterion-probability of correctly answering any items located below the cutpoint. In other words, the criterion-probability quantifies the minimal level of mastery needed for performance at each achievement level. For the field trial conducted for the 1998 Civics NAEP, a probability of correct response of 0.65 was selected for constructing the item maps. The 3PL model was used for the multiple-choice items, and it was deemed appropriate to correct for guessing because this correction is used in similar applications to NAEP data.[3] Following the NAEP convention for multiple-choice items having four options, a constant (0.25) was used as the correction for guessing. The correction was calculated as follows:

$$c(1-p)+p = 0.25(0.35)+0.65 \cong 0.74 . \qquad (18)$$

Having decided the criterion, the next step was to compute the θ value associated with the location of the item on the scale score map. The location of the item was determined by the value of θ that satisfied the following equality:

$$0.74 = p_i(\theta) = c_i + \frac{1-c_i}{1+e^{-Da_i(\theta - b_i)}} . \qquad (19)$$

[3] See, for example, U.S. Department of Education (1999).

The right hand side of Equation 19 was simply the 3PL model as shown in Equation 2. Because there is no exact solution for Equation 19, an interpolation method was used to identify the θ value.

### θ LOCATION FOR THE CONSTRUCTED RESPONSE ITEMS

The same criterion, 0.65, was applied to the constructed response items. Instead of being the probability of a correct response, however, the criterion was applied as the probability of obtaining a specific score and higher. In order to locate the constructed response items on the item map, it was necessary to "dichotomize" these items. For items having more than two score values, there is no unique scale score associated with a given probability of response that is neither incorrect nor completely correct. Thus, it was necessary to treat each "correct" response score as an item. The constructed response items with multiple score categories were recoded as several discrete items. The result was that each constructed response item with multiple score categories appeared multiple times on the item map, each representing a different score category.

To calculate the probability of obtaining a score of 2 or higher, one needed simply to accumulate the probabilities of obtaining each of the scores. That is, for obtaining scores of *k* and higher,

$$p_{i(>=k)}(\theta) = \sum_{t=k}^{m_i} p_{it}(\theta). \tag{20}$$

Specifically, since the GPC model was used for NAEP, the cumulated probability was calculated as

$$p_{i(>=k)}(\theta) = \sum_{t=k}^{m_i} \frac{e^{\sum_{r=1}^{t} Da_i(\theta - b_i + d_{ir})}}{\sum_{u=1}^{m_i} e^{\sum_{r=1}^{u} Da_i(\theta - b_i + d_{ir})}} \tag{21}$$

The right hand side of Equation 21 is the summation of Equation 4 from score *k* to $m_i$. The computation of the location of this *dichotomized* constructed response item was accomplished by setting Equation 21 equal to 0.65 and finding the θ that satisfied the equality. The interpolation method described previously was again applied to find the θ.

For example, a constructed response item with four score categories (scored 1 to 4) was *dichotomized* into three items: score of 2 and higher, score of 3 and higher, and score of 4. The item appeared three times on the item map to represent each of the score categories: ≥ 2, ≥ 3, and =4, respectively. To obtain the locations, we calculated the θs that satisfied $p_{(\geq 2)}(\theta) = 0.65$, $p_{(\geq 3)}(\theta) = 0.65$, and $p_{(=4)}(\theta) = 0.65$, respectively. Finally, the θ was transformed to the ACT NAEP-like scale with mean of 155 and standard deviation of 14. The items were then rank ordered by their values on the ACT NAEP-like scale and positioned on the item map accordingly.

Compared to the item-by-item rating methods, the item mapping methodology may yield a larger standard error for the cutpoints. This is a result of the fact that the item mapping method relies on the locations from only two items to determine the point estimate of each cutscore. Although the task of locating a cutscore on an item map appears easy for the panelists, it could be more cognitively challenging to determine one point than to estimate the probabilities for items. The item-by-item rating methods allow more flexibility. A few mis-judged item ratings do not jeopardize the validity of the final cutpoints.

A more serious criticism of the item mapping method is the necessity of having a fixed probability to use as the mapping criterion. In the NAEP field trial application, 0.65 was used because other areas in the NAEP program have used this value. That seemed sufficient for the field trials, but a policy decision by NAGB would have been needed to use this methodology in an operational ALS context. If panelists are allowed to make the choice, then the stringency of the standards and the operational definition of "mastery" vary. If NAGB decides or if the technical advisory committee decides, then an important aspect of standard setting is fixed and not subject to the judgment of the panelists.

A further complication of the item mapping methodology is the fact that the order of the items is not maintained with different probability values used as the criterion with the 3PL and GPC IRT models. Different criteria can produce different orderings of the items that change the locations of items on the map. If the cutpoint happens to be between two items that are reversed, the results cannot be reconciled. A sample item map and item mapping list are shown in Figures 3 and 4.

### COMPUTATIONAL PROCEDURE FOR THE RECKASE METHOD

The Reckase Method is another experimental method that was first tried out during the second field trail for both the 1998 NAEP in civics and in writing. The Reckase Method was introduced by ACT to provide panelists with useful and easy-to-understand information regarding the consistency in their own ratings in the IRT context. Ratings could be evaluated directly by the panelists to examine consistency for items of different types (multiple choice and constructed response), content (international relations versus foundations of American democracy), and difficulty levels. Panelists could also examine ratings for evidence of rater fatigue.

The centerpiece of the Reckase method is the Reckase Chart. The Reckase Chart is essentially a numerical representation of the item characteristic curves (ICC). The numerical entries for individual items were arranged in columns, with the item numbers identifying each column of data. Rows contained data for each scale score value. The data entries were expected performances on each item at each score point. The discrete ACT NAEP-like scale scores identified the rows, and they were arranged in descending order. For multiple-choice items, each entry was the probability of a correct response at a specific scale score: The expected proportion of correct responses for students performing at a given scale score. For constructed response items, each entry was the expected score: the average score expected for students performing at a given scale score. In each case, the "expectation" is derived from the IRT model.

Each panelist was provided a Reckase Chart that included each item in his or her item rating pool. For the field trials and civics pilot study, panelists marked their Reckase Chart by circling the entries corresponding to their item-by-item ratings in the previous round. The ratings were marked electronically for panelists in subsequent implementations. The panelists also marked their individual cutpoints and the grade level cutpoints corresponding to the NAEP-like scale score on the chart. The pattern of their ratings across items thus became visually clear. The panelists were able to inspect their ratings for each item with respect to their own cutpoints and the grade level cutpoint. The Reckase Charts were so arranged that only one block of items was displayed on each page.

The Reckase method included rounds of item-by-item ratings. In the field trials, panelists completed two rounds of ratings using the modified Angoff and mean estimation methods. For the third round, they were asked to select a single row on the Reckase Chart to represent their individual cutpoint for each of three

achievement levels. The average of the individual cutpoints was computed to serve as the overall grade level cutpoint.

To prepare the Reckase Charts, each ACT NAEP-like scale score *S* was transformed to the theta scale.

$$\theta = \frac{(S - 155)}{14}. \tag{22}$$

For example, an ACT NAEP-like score of 160 was converted to 0.357 on the θ scale. This θ was then used to obtain the entries on the Reckase Charts corresponding to the ACT NAEP-like scale score. For the multiple-choice items, the probability of a correct response was calculated as

$$p_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} \tag{23}$$

which simply applied the 3PL model as in Equation 2. The expected score for the constructed response items was computed as

$$T_i(\theta) = \sum_{k=1}^{m_i} k p_{ik}(\theta) = \sum_{k=1}^{m_i} k \left( \frac{e^{\sum_{r=1}^{k} Da_i(\theta - b_i + d_{ir})}}{\sum_{u=1}^{m_i} e^{\sum_{r=1}^{u} Da_i(\theta - b_i + d_{ir})}} \right) \tag{24}$$

That is, the expected score $T_i$ given θ was the accumulation of the probabilities (associated with each score category) times the corresponding score categories, given θ. To construct the Reckase Charts, the obtained $p_i(\theta)$ or $T_i(\theta)$ were entered in the chart for each item at the corresponding ACT NAEP-like scale score.

The Reckase Chart was designed to provide information regarding rating consistency across items within rounds for the panelists, and this information could be used to bring ratings more in line with the item difficulty and other item characteristics in the region of the borderline for each achievement level. There was concern that the panelists' rating might be too influenced by the Reckase Charts—that panelists would concentrate on being consistent and ignore the correspondence between the ratings and the achievement level descriptions. Thus, the actual score scale was not printed on the charts given to panelists following the first round of ratings. Instead, letters of the alphabet were used to identify each row on the chart for panelists to evaluate after the first round of ratings. Following the second round of ratings, panelists were given charts marked with the ACT NAEP-like cutscore. At that time, panelists were asked to select three rows to represent their cutscores for each of the three NAEP achievement levels.

ACT decided *not* to implement the Reckase Method, per se. Instead, panelists continued with item-by-item ratings throughout three rounds of item ratings. Reckase Charts were provided to each panelist following the first two rounds of ratings. The charts were marked with panelists' ratings for each item, and the charts served as feedback to panelists on intrajudge consistency. Reckase Charts are now a part of the NAEP ALS process and they are an integral part of the ACT/NAGB method. See Figure 5 for a sample Reckase Chart.

## COMPUTATIONAL PROCEDURE FOR BOOKLET CLASSIFICATION METHOD

The Booklet Classification method was implemented as another rating method tested during the second field trail for the 1998 Writing NAEP. Unlike the previous rating methods, the booklet classification method was a holistic and examinee-centered rating method. ACT had experimented previously with booklet classification procedures in validation research studies conducted for NAGB, but the method had never been implemented as an achievement levels-setting procedure. Because the Writing NAEP contains only two constructed response items for each student, the booklet classification procedure seemed appropriate. Findings from previous studies suggested, however, that the cutscores would be set higher with this method than with the item-by-item procedure typically used by ACT in the NAEP ALS process (ACT, 1995; ACT, 1997b; Hanson, Bay, & Loomis, 1998; Loomis, 2000).

A photocopy of a student's responses to the exercise questions in a NAEP test form was made for each booklet used in the study. Student responses to background questions were omitted from these booklets. The items had already been scored, and the ACT NAEP-like scale scores had been computed for each booklet. Booklets were selected to represent performance rather evenly along the score scale. Each panelist classified each of the booklets in his or her set into seven categories: below Basic, borderline Basic, Basic, borderline Proficient, Proficient, borderline Advanced, and Advanced.[4] The booklets were ordered in terms of performance levels from lowest to highest, and panelists were notified of this fact. ACT had not ordered booklets in previous studies, and the ordering seemed to alter the procedure rather sharply.

Based on the panelists' classifications, the cutpoints were computed. Four methods of computing the cutpoints were considered and applied for evaluation purposes: collapsed categories, average borderline, weighted collapsed and borderline, and cubic regression methods. The weighted collapsed and borderline method was used to compute cutpoints during the field trial for purposes of providing feedback to panelists. Please refer to Hanson & Bay (1999) for more details.

## COLLAPSED CATEGORIES METHOD

With this method, the borderline categories were collapsed with the corresponding achievement level categories so that the seven classification categories were reduced to four: below Basic, Basic, Proficient, and Advanced. Each cutpoint was computed separately using a two-category procedure. That is, each cutpoint was computed based on the booklets that were classified in the same category by the decision rule and the panelists, and booklets that were classified in the next lower category by the decision rule and the panelists. The cutpoint for each achievement level, $\theta_l$, is defined as the value that minimizes the Bayes risk, that is,

$$r(\pi, \theta_l) = \pi_{l-1} p(\eta \geq \theta_l \mid L = l-1) + \pi_l p(\eta < \theta_l \mid L = l) \qquad (25)$$

Where, $r$ is the Bayes risk, $\pi$ is the prior probability of a booklet being classified into one of the four achievement level categories (l = 0, 1, 2, 3 for below Basic, Basic, Proficient, and Advanced), and $p(\eta \geq \theta_l \mid L=l\text{-}1)$ is the probability of a scaled score of $\eta$ for a booklet being greater or equal to the cutpoint for achievement level $l$ given that the booklet is classified at level $l$-1. The prior probability $\pi_{l-1}$ is calculated as $n_{l-1}/(n_{l-1}+n_l)$, the number of booklets being classified as at level $l$-1, $n_{l-1}$, divided by the number of booklets being classified as at level $l$-1 or $l$, $n_{l-1}+n_l$. Likewise, $\pi_l$ is calculated as $n_l/(n_{l-1}+n_l)$. Equation 25

[4] Please see Loomis, Hanick, Bay and Crouse (2000b) for a complete description of this procedure.

produces cutpoints for each panelist based on his or her classifications. The overall cutpoint for the group is then the average of the individual cutpoints.

### AVERAGE BORDERLINE METHOD

This method of setting cutpoints uses only booklets that are classified in a borderline category. The cutpoint for each level is the mean $\eta$ of the borderline booklets for that level. Again, the overall cutpoint is then the average of the individual cutpoints across panelists. In other words, the overall cutpoints computed with this method are the grand averages of the scores of the booklets that are classified in a borderline category, based on the classifications of all panelists.

### WEIGHTED COLLAPSED AND BORDERLINE METHOD

With this method, the cutpoint is calculated as the weighted average of the cutpoints obtained with the collapsed category and the average borderline methods. The weight given to the collapsed category cutpoint is

$$\frac{n_{l-1} + n_l}{n_{l-1} + n_l + n_b}, \tag{26}$$

where $n_b$ is the number of booklets classified at the borderline of the level, which is also included in $n_l$. The weight given to the average borderline cutpoint is

$$\frac{n_b}{n_{l-1} + n_l + n_b}. \tag{27}$$

The value of the weighted collapsed and borderline cutpoint for a particular achievement level is the collapsed cutpoint multiplied by the weight in Equation 26 plus the average borderline cutpoint multiplied by the weight in Equation 27.

The process of combining the two sets of cutpoints results in more weight being given to booklets classified as borderline than those classified as within levels. The borderline booklets are used twice in the computation of the cutpoints: once in the collapsed category method and once in the average borderline method. Booklets that are classified as within-levels are used only once in the collapsed category method.

### CUBIC REGRESSION METHOD

Cutpoints using the cubic regression method are obtained by fitting a cubic regression model to the booklet classification data (Hambleton & Plake, 1995). The numerical values of the classifications, i.e., 1, 2, 3, …, 7, serve as the independent variables in the regression equation, and the dependent variable is the scale score, $\eta$, associated with each booklet. The regression curve represents the conditional mean $\eta$ as a function of the achievement levels. Cutpoints are given by the values of the regression curve corresponding to the borderline categories. The averages of the individual cutpoints are then used as the overall cutpoints for the grade.

Of all the methods that have been tested, the booklet classification method was probably the most time consuming and mentally demanding one. In order to produce reliable classifications, panelists must read multiple booklets that are representatives of the entire range of proficiency. For example, for the field trial

for the NAEP Writing assessment, ACT proportionally selected booklets based on the distribution of the raw scores. ACT selected six booklets each for the score range of 1-3 and 10-12, and 12 booklets each for the score range of 4-6 and 7-9. This was a total of 36 intact booklets to be classified into four major performance levels: Advanced, Proficient, Basic, and below Basic. Booklets were ordered from low to high performance for panelists to use in the classification.

The method was not used in further studies. Although the field trial cutpoints resulting from this procedure were not higher than those from the Reckase method, the consistent evidence from previous studies suggested that these results might not be repeated. Further, there was no consensus among the members of TACSS on the choice of computational procedure to use for the booklet classification method. Different computational procedures produced quite different cutpoints. (See Table 1.) Further, the outcomes of the method are subject to the impact of several factors related to the study design such as the number of booklets to be used, the distribution of the booklets across the score scale, and the number of categories to be classified. Those reasons, coupled with considerations of the heavy logistic burden for implementation represented by this procedure and the heavy cognitive burden for panelists led to the decision to abandon further studies with this methodology.

## COMPUTATIONAL PROCEDURES FOR FEEDBACK USED IN STUDIES FOR THE 1998 NAEP ACHIEVEMENT LEVELS-SETTING PROCESS IN CIVICS AND WRITING

The NAEP ALS process includes an extensive set of feedback that is produced for panelists following each round of ratings. The feedback data are generally produced for each grade level, but some are produced for each panelist. The wholebooklet data are produced for panelists in each rating group. In field trials, when several rating methods and experimental procedures were implemented, feedback data were produced for each group: a total of four sets of feedback data. Please see reports of the studies conducted for the 1998 NAEP ALS process reported by Loomis, et al. (2000a – 2000f).

### CUTPOINTS AND STANDARD DEVIATIONS

The overall grade level cutpoints were provided for the panelists as feedback after each round of ratings. In the field trials, for which multiple methods were used, the cutpoints were computed for each group using a different rating method and having a different type of consequences feedback. For the second field trials in each subject, four different sets of feedback were computed.

The computational procedures of the grade level cutpoints were described in the previous section for each of the different rating methods. Along with the overall grade level cutpoints, the feedback includes the standard deviations of the cutpoints. As noted previously, the standard deviations of the overall grade-level cutpoints are the standard deviations based on the individual cutpoints of each panelist with one exception, the mean estimation method. For the mean estimation method, the standard errors of overall cutpoints are half of the absolute difference between the grade group cutpoints (see Equation 10). However, in the grade level cutpoint feedback, the standard deviation of the individual cutpoints was used. The purpose of showing the standard deviation was to provide information to panelists regarding the spread of the individual cutpoints, and the standard deviation of the individual cutpoints was appropriate for this purpose.

The creation of feedback had two steps. First, the main computational program computed the cutscore and standard deviation using the procedure appropriate to the rating method used to collect panelists'

judgments. The main program, written in Fortran, output many statistics including the overall cutpoints, the individual cutpoints, and the standard deviations associated with each, for example. The second step was to input the overall cutpoints and their standard deviations into an Excel spreadsheet. An Excel chart was automatically created. In this chart, the grade level cutpoints were shown, along with the actual value of each cutpoint. The variability in the cutscores was shown as vertical lines one standard deviation above and below the points representing the cutscores. Figures 1 and 2 are examples of this feedback.

### RATER LOCATION FEEDBACK

Rater location feedback is a graphical display of individual cutpoints of each panelist. The feedback provided each panelist with information regarding his or her own cutpoints relative to the grade level cutpoints and relative to other panelists in his or her own group or grade. In other words, rater location feedback was used to show interrater consistency. Additionally, it also provided the relative locations of the three cutpoints that each panelist set. Panelists could evaluate whether the differences between cutpoints set for each achievement level seemed appropriate, given their understanding of the achievement levels descriptions.

The construction of the rater location feedback was very easy to accomplish. Rater location charts are histograms with the ACT NAEP-like scale scores as the horizontal axis, and the frequency count for cutpoints set by panelists at each score point as the vertical axis. Each panelist's identification was used as the symbol of the data value so panelists could identify the location of their cutpoint for each, specific achievement level. The data necessary for constructing the rater location feedback were the individual cutpoints.

### COMPUTATIONAL PROCEDURE FOR INDIVIDUAL CUTPOINT USING MODIFIED ANGOFF AND MEAN ESTIMATION METHODS

The computation of individual cutpoints when using the mean estimation method was similar to the computation of the group level cutpoints, except the individual panelist's item-by-item ratings were used rather than the average ratings of the group. For the multiple choice items the task was to find the $\theta_{jm}$ that satisfied

$$\sum_{i=1}^{n_{gm}} r_{ij} = \sum_{i=1}^{n_{gm}} p_i(\theta_{jm}), \tag{28}$$

where $\theta_{jm}$ is the individual cutpoint for multiple-choice items of panelist *j*. For constructed response items, it was to find the $\theta_{jc}$ that satisfied the following:

$$\sum_{i=1}^{n_{gc}} r_{ij} = \sum_{i=1}^{n_{gc}} \sum_{k=1}^{m_i} k p_{ik}(\theta_{jc}) \tag{29}$$

where $\theta_{jc}$ is the individual cutpoint for constructed response items of panelist *j*. Once the cutpoints for the multiple choice items $\theta_{jm}$ and constructed response items $\theta_{jc}$ were computed, the individual cutpoints for panelist *j* were calculated as the weighted averages of $\theta_{jm}$ and $\theta_{jc}$ using item information as the weights, just as for the group level cutpoints. The information weights were calculated using Equations 6 and 8.

### COMPUTATIONAL PROCEDURE FOR INDIVIDUAL CUTPOINTS USING ISSE METHODS

The computation of individual cutpoints when the ISSE method was used required no computation other than the computation of the grade level cutpoint. The grade level cutpoint *was* the average of the individual cutpoints. More accurately speaking, in computing the grade level cutpoint, the individual cutpoints were obtained as the byproducts. For details, please see the description of the computational procedure using the ISSE method above.

### COMPUTATIONAL PROCEDURE FOR INDIVIDUAL CUTPOINTS USING THE ITEM MAPPING, RECKASE, AND BOOKLET CLASSIFICATION METHODS

No computations were needed for the individual cutpoints when either the item mapping procedure or the Reckase method was used. Panelists were required to set their own cutpoint directly on either the item map or on the Reckase Chart. The three locations or score points that each panelist selected for each of the achievement levels were the corresponding cutpoints. These values were used directly to produce the rater location feedback.

The actual creation of the rater location feedback was accomplished through a subroutine embedded in the program that performed all the computations. This program was written in Fortran. The individual cutpoints were inputs to the rater location feedback subroutine of the main program used to calculate the frequency distribution. This frequency table was then converted to a histogram and output as a text file. The feedback was the printed hardcopy of this text file.

The actual rater location feedback consisted of three histograms, one for each of the three achievement levels. The three individual cutpoints set by each of the panelists were shown in all three of the histograms. The difference between each of the histograms was that only the individual cutpoints of one level were labeled with the personal codes assigned to each of the panelists, while the individual cutpoints for other two levels were shown with shading symbols and no personal codes. For example, the histogram for the Basic level showed the individual raters' cutpoints for the Basic level identified by their personal codes. Their individual cutpoints for Proficient and Advanced levels were just shown with generic symbols. This design of rater location feedback enabled the panelists to concentrate on one level at a time and to notice the positions of their own cutpoints relative to those of other panelists. The use of personal codes kept the individual's cutpoint confidential and known only to the panelist. Because the individual cutpoints for all levels were represented on the same histogram, the panelists were able to compare the relative positions of the cutpoints across all three levels. An example of Rater Location Charts is given in Figure 6.

## WHOLEBOOKLET FEEDBACK

Unlike the rater location feedback, the computational procedure for the wholebooklet feedback was the same regardless of the rating method used. The wholebooklet feedback utilized only the grade level cutpoint. The wholebooklet feedback gave the panelist a sense of the typical performance associated with just reaching each of the three achievement levels. Wholebooklet data were reported in terms of the percentage of score points for a particular booklet form needed to qualify at the cutpoint of each achievement level. During the operational ALS process, the actual grade level cutpoints were used to produce the wholebooklet feedback because only one rating method was used. During the field trail,

different rating methods were used, and the wholebooklet feedback was based on the group cutpoints for each rating method used.

As part of the training process, panelists take the NAEP in the subject using actual test booklets. Two different booklet forms were used for each grade level so that the raters in each rating group were assessed over items not included in their item rating pool. The NAEP booklet forms used in this initial training exercise were used to provide the wholebooklet feedback. The wholebooklet feedback was the percentage of total points possible in the particular test booklet that were required for performance at the cutscore at each achievement level. The percentage of score points that students would need to obtain was simply the ratio between the IRT expected score and the total number of score points for items included in that booklet. The IRT expected score for the booklet, $T_{booklet}$, given the overall cutpoint, $\theta_{overall}$, was calculated as,

$$T_{booklet}(\theta_{overall}) = \sum_{i=1}^{n_m} p_i(\theta_{overall}) + \sum_{i=1}^{n_c}\sum_{k=1}^{m_i} kp_{ik}(\theta_{overall}) \qquad (30)$$

where $n_m$ was the number of multiple choice items in the booklet, $n_c$ was the number of constructed response items, $p_i$ was the probability of a correct response to multiple choice item i based on the 3PL model, and $p_{ik}$ was the probability of a response in category k of constructed response item i based on the GPC model.

The actual creation of the wholebooklet feedback required two steps, just as for the production of the overall cutpoint feedback. First, the percentages of score points were computed by the main Fortran program. The main Fortran program output the percentages along with many other statistics such as the overall cutpoint, the standard deviations of the cutpoints, and the rater location data. The second step was to input these percentages into an Excel spreadsheet. Three pie charts were generated from the input values of the percentages. The three pie charts represented the performances (percentage of total score points) necessary to reach the achievement level cutpoints computed from the item ratings. An example of Wholebooklet feedback is given in Figure 7.

### CONSEQUENCE FEEDBACK

As with the wholebooklet feedback, the computational procedure for the consequence feedback was the same regardless of which rating method was used. It involved only the grade level cutpoints. The consequence feedback was simply the percentage of students expected to perform at or above each grade level cutpoint. When empirical data were available, the percentages of students who obtained scores at or above the achievement level cutpoints (the percentage of plausible values at or above the cutpoints) were reported. When empirical data were not available, a normal population distribution was assumed, and numerical integration was applied to obtain the percentages. The percent at or above the cutpoint is,

$$P_{consequence} = \sum_{i:X_i \geq \theta_{overall}}^{N_q} W(X_i), \qquad (31)$$

where $N_q$ is the number of quadrature points for the numerical integration, and $W(X_i)$ is the weights are quadrature point $X_i$.

The consequences feedback provided the panelists with the percentage of students expected to reach each of the three achievement levels. This helped panelists to judge whether the cutpoints they set to represent

the standards were higher or lower than their expectations of performances. It did not, however, help panelists to discern whether the standard was too high or student performance was too low.

As was the case for feedback previously described, the actual creation of the consequence feedback was a two-step process. First, the percentages of students at or above each achievement level were computed in the main Fortran program. The program was designed to output these percentages along with other statistics required for other feedback. The second step was to use these percentages as input to an Excel spreadsheet. Three bar graphs were created from the percentages. The first bar graph showed the percentage of students at or above the Basic level with no information to delineate percentages at or above the Proficient and Advanced levels. The second bar graph showed the percentage of student at or above the Basic level overlaid by the percentage of students at or above the Proficient level. The third bar graph showed the percentage of students at or above the Basic level, overlaid by the percentage of students at or above the Proficient level, overlaid by the percentage of students at or above the Advanced level.

The consequence feedback was shown to the panelist in a group session. The bar graphs were shown one at a time so that the panelists could see the change of percentages from one level to the next. In addition, on the top right corner of the bar graph a pie chart was also shown. The pie chart, however, showed only the percentages within each achievement levels. The pie chart reported the percentages of student *within* each achievement levels, whereas the bar graph reported the percentage *at or above* each achievement levels. An example of consequences feedback in included in Figure 8.

### INTRARATER CONSISTENCY FEEDBACK

The Reckase Charts were used as feedback to allow panelists to examine the internal consistency of their ratings. The item-by-item ratings of each panelist can be marked on the Reckase Chart to reveal the variability, or lack thereof, of judgements for each item for each of the three achievement levels. If a panelist was consistent with his or her judgement and perceived the item difficulty well, the pattern of the ratings would form a relatively flat line across the Reckase Chart. Panelists could judge the consistency of their ratings by examining the charts and the extent to which the pattern was relatively flat or characterized by "peaks and valleys." Panelists could compare ratings for multiple choice items with those for constructed response items to determine the consistency of ratings by item format. They could also compare ratings for items within different content categories. An example of a Reckase Chart is included in Figure 5.

### SELECTION OF EXEMPLAR ITEMS FOR THE 1998 NAEP CIVICS AND WRITING ASSESSMENTS

Exemplar items are used to facilitate the interpretation of performance relative to the achievement levels. After panelists reach agreement on the cutpoints to be recommended to NAGB, they engage in the process of selecting exemplar items. An item meets one criterion for consideration as an exemplar item at a given level if the weighted average conditional probability of a correct response for that item exceeds 0.5 across the range of the achievement level. If a multiple choice item *i*, meets the following criterion at achievement level *l*, it will be listed for review by panelists as an exemplar item.

$$\frac{\sum_q p_i(X_q)W(X_q)}{\sum_q W(X_q)} \geq 0.5, \text{ where the sum is over } q \text{ such that } \theta_l \leq X_q < \theta_{l+1} \qquad (32)$$

In equation 32, $\theta_l$ is the cutpoint for level *l*, and $\theta_{l+1}$ is the cutpoint one level above level *l*. $W(X_q)$ is the weight at quadrature point $X_q$, and $p_i(X_q)$ is the probability of a correct response to item *i*, given ability $X_q$, that is calculated using the 3PL model of Equation 2. $X_q$ is the discrete quadrature point chosen to represent the value on the θ scale. Currently, 600 quadrature points between –6 and 6 on the θ scale are used for these computations.

In addition to the 0.50 performance/difficulty criterion, a discrimination criterion is also applied to items that are to be considered in the exemplar item selection process. One must first calculate the difference between the weighted average conditional probabilities of adjacent achievement levels. After the weighted average conditional probabilities are computed for the three achievement levels, the differences between the probabilities at the below Basic and Basic, Basic and Proficient, and Proficient and Advanced levels are computed. For all the items that meet the weighted average conditional probability criterion for a given achievement level, the differences are rank-ordered. The value associated with sixty percent of the items in the grade pool having the highest differences in probabilities is used to list released items for primary consideration as exemplar items, and the remaining items are listed as the secondary candidates. A simple example with three items illustrates this step. If there are only three items A, B, and C in an assessment and all have probabilities that exceed 0.5 within the Basic range, they all meet the "difficulty" criterion of being a Basic level exemplar items. In addition, if item A has the highest difference in the average conditional probability between the Basic and below Basic levels, followed by item B and then by item C, then items A and B (60% of the items) would be on the primary list and item C (the remaining 40%) on the secondary list. The difference of the weighted average conditional probabilities between the adjacent achievement levels are computed as,

$$\left(\frac{\sum_q p_i(X_q)W(X_q)}{\sum_q W(X_q)}\right) - \left(\frac{\sum_v p_i(X_v)W(X_v)}{\sum_v W(X_v)}\right), \qquad (33)$$

where the sum is over *v* such that $\theta_l \le X_v \le \theta_{l+1}$, and the sum is over *q* such that $\theta_{l+1} \le X_q \le \theta_{l+2.}$

Computation of the weighted average conditional probability for the constructed response items is a rather complex proposition. The problem is that it is rare to find a θ value where the probability of obtaining the given response exceeds 0.5; hence the average across the range would never exceed 0.5. Further, such θ values can only be found for the lowest and highest response categories—none in between. Therefore, the constructed responses are recoded into dichotomized items, one for each credited response, as described previously. The dichotomized items coded *correct* for scores of 2, 3, or 4 will cover a wider range of ability and represent a lower level on the achievement/performance score scale. The dichotomized items coded *correct* for scores of 3 and 4 cover a shorter range of ability and represent a higher level on the achievement/performance score scale. The dichotomized items coded *correct* for a score of 4 cover the shortest range of ability and represent the highest level on the achievement/performance score scale. The recoding creates three pseudodichotomized items to represent the 4-point polytomous item: 1 vs. 2, 3, or 4; 1 or 2 vs. 3 or 4; and 1, 2, or 3, vs. 4. The computation requires a summation over the conditional probabilities of a given response categories. For the pseudodichotomized item 1 vs. 2, 3, or 4, one simply sums over the conditional probabilities of obtaining a score of 2, 3, and 4. The formula is,

$$\frac{\sum_{q}\sum_{k=1}^{m_i} p_{ik}(X_q)W(X_q)}{\sum_{q}W(X_q)} \geq 0.5, \text{ where the sum is over } q \text{ such that } \theta_l \leq X_q < \theta_{l+1}. \quad (34)$$

Where, $\theta_l$ is the cutpoint for level $l$ and $\theta_{l+1}$ is the cutpoint one level higher, $\phi(X_q)$ is the normal density function at $X_q$ that is used as the weights, and $p_{ik}(X_q)$ is the probability of obtaining a score of $k$ for item $i$, given ability $X_q$, that is calculated as the GPC model. $X_q$ is the discrete quadrature point chosen to represent the value on the θ scale. Currently, ACT uses 600 quadrature points between the –6 and 6 range of the θ scale.

Again, these pseudodichotomized constructed response items are subject to the discrimination criterion for further screening in the exemplar item selection process. The differences in the average conditional probabilities are calculated as,

$$\left(\frac{\sum_{q}\sum_{k=1}^{m_i} p_{ik}(X_q)W(X_q)}{\sum_{q}W(X_q)}\right) - \left(\frac{\sum_{v}\sum_{k=1}^{m_i} p_{ik}(X_v)W(X_v)}{\sum_{v}W(X_v)}\right), \quad (35)$$

where the sum is over $v$ such that $\theta_l \leq X_v \leq \theta_{l+1}$, and the sum is over $q$ such that $\theta_{l+1} \leq X_q \leq \theta_{l+2}$.
One point worthy of notice is that once an item meets the 0.5 criterion for an achievement level it would certainly meet the criterion for a higher level, since the conditional probability is even higher at higher θ values. Given that, an item is only listed one time, and that is at the lowest level at which it has qualified. After the candidate items list is determined, based on the screening criteria, the achievement level-setting panelists are given the responsibility of selecting from the list those items that best represent the achievement levels descriptions. Three lists of items are presented to the panelists in each grade group, one for each achievement level. Items that reach the 0.5 criterion within the below Basic range are not to be considered in the selection of exemplar items. The panelists may recommend items on the secondary list, but they are urged to do so only if they cannot find (enough) suitable ones from the primary list.

For the 1998 Writing NAEP achievement levels-setting process, there was no distinction made between items with respect to the primary and secondary lists. Only three item prompts were released for each grade so there were very few "items" to consider in the exemplar selection process. Items presented to panelists were ordered on a single list according to their discrimination, from highest to lowest.

Occasionally, an item might not meet the 0.5 criterion within any of the specified performance ranges. These are the very difficult items, and they tend to be the highest score category for constructed response items. In such cases, the weighted average conditional probability of obtaining the highest score does not reach 0.5 within the Advanced range (i.e., from Advance cutpoint to θ=6.0). Even at this very high level of performance, these items do not meet the criterion and would not appear on the exemplar item list. They are far too difficult to qualify as exemplar items—even for Advanced level performance.

Another technical issue for further discussion is the decision to use the normal density function as the population distribution for weights as opposed to the empirical distribution. Previously, the NAEP ALS process was conducted before the empirical distribution data were available. Given this circumstance, the

normal distribution was the best and *only* choice because no other source of information was available. This turned out to be the best approximation to the empirical distribution for the NAEP Writing and NAEP Civics administered in 1998. For both the 1998 NAEP in writing and in civics, unidimensional scales were used. Additionally, the unidimensional scale was set to have a mean of zero and standard deviation of one. The empirical distributions of the two administrations were very similar to a normal distribution. Given that the normal distribution was a good approximation, the decision was made to continue using the normal distribution as the population distribution for purposes of assigning weights in the computation of conditional probabilities.

The creation of the printouts of the exemplar items was performed by a subroutine embedded in the main Fortran program. The lists of the exemplar items were output as text files. The printed hard copies were then distributed to the panelists. An example of an Exemplar Item list is included in Figure 9.

## DATA ENTRY AND QUALITY CONTROL PROCEDURES

The NAEP ALS Process designed by ACT has always consisted of several rounds of ratings with feedback provided between rounds. Because feedback is based on the previous round of ratings, it is necessary to process and analyze the ratings on site and produce feedback *very* rapidly and efficiently. To assure accurate and efficient data entry of the ratings for computation of the cutpoints and production of the feedback, both the data entry program and the computational program were written for the NAEP standard setting project.

One data entry program was written for each method of item ratings implemented. The fundamental features of the data entry program have been modified over the years to increase the ease, speed and accuracy of data entry. The data entry program was only slightly different for each method in terms of the constraints on the values that were considered "legitimate" values for ratings. The data entry program was written in Basic programming language. When executed, the computer monitor displayed a screen exactly like the rating form used by the panelists in each rating group. The data entry personnel could accurately and efficiently enter the ratings into the database.

To further ensure the accuracy of the data, constraints were built into the program so that logically incorrect ratings could not be entered. First, the ratings had to be in ascending order for Basic, Proficient, and Advance levels. A rating for the lower level of achievement could not be greater than that for a higher level. If an incorrect value were entered, the program would issue a warning beep and the cursor would be locked on the incorrect value. The data entry personnel would then check whether there was a data entry error or an error by the panelist in marking the rating. In the latter case, the data entry personnel could by-pass the field and the program would automatically insert a flag in the field. Other project staff would talk to the panelist regarding the rating in question. Once the problem was resolved, the data entry personnel could go back to the database and enter the correct value(s).

The second constraint was that the rating could not below the minimum or above the maximum score for the items. Again, if out-of-range values were encountered, the program would issue a warning and lock on the field until further instruction from the data entry personnel. The procedure to correct the problem was exactly the same as previously described. For a multiple-choice item, the ratings should be between 0% and 100%. For short constructed-response items, the ratings should be any value (with one decimal place) between 1 (the minimum score) and 3 (the maximum score. For the ISSE method, the ratings for multiple

choice items could only be 0 (incorrect response) or 1 (correct response). For a short constructed response item, for example, ratings could only be whole integers 1, 2, or 3.

For an even higher level of assurance of quality control, a hard copy of ratings was printed for each panelist after his or her ratings had been entered into the database. The staff, in pairs, would proofread the data by comparing the original rating form and the printout. Any inconsistency between the printed copy and the original rating form was corrected to represent the value on the original rating form. The correction was performed directly in the database. The database was, in fact, a text file output directly by the data entry program. One text file was saved for each of the panelists. These text files were the input files for the main computation program to produce the cutpoints and the feedback.

As mentioned earlier, this data entry program was used only for the item-by-item rating methods. The large number of items made the quality control checks in the data entry program a necessity. For the other rating methods—Item Maps, the Reckase method, and the Booklet Classification method, no special data entry program was needed because the number of "ratings" was small.

There were only three cutpoints to be recorded for panelists who used either the Item Maps or the Reckase method. Data entry was easily accomplished by using an Excel spreadsheet to enter ratings into a database and output the data as a text file. Two of the staff checked the database and the original rating forms (Item Maps or Reckase Charts) to ensure the ratings were entered correctly. The text files were used as the input to the main computation program.

The "ratings" for the Booklet Classification method were of a totally different nature than those of any of the methods previously described. Rather than probability or score estimates, the "ratings" were the achievement level classifications for each of the booklets. ACT used 7 different classification levels: the three achievement levels, borderlines of the three achievement levels, and below Basic. There was one and only one classification associated with each booklet for a given panelist. Thus, data entry was somewhat easier. These classifications were entered into an Excel spread sheet, one for each panelist. A text file was output for each panelist as his or her datafile. The datafiles were used as input for the main computation program to produce the cutpoints and feedback. Again, before the computation, the datafiles were proofread for accuracy against the original classifications.

### THE MAIN COMPUTATIONAL PROGRAM

In order to produce data on-site throughout the process, it was necessary to have a program that could compute the cutpoints and produced feedback accurately and efficiently. Because of the complexity of the different rating methods and the specifics of the NAEP, this program had to be custom-made and virtually guaranteed to work. The main computational program was written in Fortran, and the Lahey Fortran 77 was used as the compiler for this purpose.

The main program used the text data files (see the data entry program section) as input for computing the cutpoints. It then output all the statistics that were required to produce the feedback. The main program directly produced some feedback, such as the Reckase Charts, rater location charts, and the exemplar item lists. Some feedback were produced using data from Excel spreadsheets as input to compute the statistics that the main program computed such as the overall cutpoints, the wholebooklet feedback, and the consequence feedback. Excel was selected so that ACT could provide the panelists with graphical feedback that was easy to edit and easy to understand.

When the main program was executed, it first asked for the basic information such as the rating methods, the NAEP subject, the number of panelists, and the name of the control file. This information enabled the program to allocate the memory space and to activate the correct computational procedures. The main program generated several output files. First, the main result file that contained the individual cutpoints, the group level cutpoints, the grade level cutpoints, and the standard deviations associated with each. It also contained the percentages for the wholebooklet feedback and the percentages for the consequence feedback. Another output file contained the rater location feedback. The last output file contained the exemplar item lists.

To ensure that the analyses were correct, the main program also output a log file that contained many intermediate steps as output from the computational procedures. For example, the log file contained the number of panelists, the number of items, and the number of groups that the program had read as the input. It also contained the item parameters based on the IRT model. The user could examine the log file to determine if there were any inconsistencies between the intended input and the actual input. This was part of the quality control built into the computational process.

Rigorous tests of the main program were performed before it was actually applied in the NAEP standard setting process. To ensure the accuracy of the main program, NAEP data from previous ALS procedures were used to test whether the program could duplicate the results. In instances for which there was a new rating method but no old data, simulated data were generated. These data were also analyzed with statistical software such as SAS to ensure that the results were duplicated. The feedback was reviewed by the project staff to ensure the accuracy and completeness of feedback. In summary, the main computational program was tested to be accurate and efficient before it was used for the NAEP standard setting. All computations were found to be accurate.

A feature of the computational procedure should be mentioned here because it was embedded in the main program and did not fit in other sections. Under two scenarios the program will automatically put constraints on the individual cutpoints for the panelists.

The first scenario can occur only in the modified Angoff procedure with multiple-choice items. Recall that there is a probability of guessing associated with the multiple-choice items (see Equation 2, the 3PL model). This implies a minimum probability of answering the question correctly. It happens that if the average of the panelist's ratings (for multiple-choice items) is less than the average of the guessing parameters, then no cutpoint can be computed. In that case, the projection from the sum of ratings does not intercept the TCC, hence there is no corresponding $\theta$. For the modified Angoff procedure, the main program always tests whether any average for multiple-choice item ratings is less than the average of the guessing parameters. When the program finds one, it sets the panelist's average rating very, very slightly higher than the average of the guessing parameters and computes the cutpoint accordingly. This is a rare, but possible, scenario; and it only happens when computing an individual panelist's cutpoint based solely on multiple-choice items. These individual cutpoints are computed for research purposes and not for feedback in the process. Additionally, if this situation should arise, it will almost always be for the Basic level cutpoint. An even more remote possibility is that this would happen at the grade group level cutpoint. In this case, the average rating of all panelists in the rating group would be lower than the average of the guessing parameters. This would indicate an *extremely* low cutscore for the entire group of panelists. This scenario has never occurred in a NAEP standard setting process, however.

In any event, the "fix" for individual raters' cutpoints described above is only a technical fix so that a cutpoint can be computed and feedback can be produced. The real solution comes from discussing the extremely low item ratings with the panelist to determine whether that is, indeed, his or her judgment of student performance or whether the low ratings were a result of a misunderstanding about the rating methodology.

The other problem scenario could only occur with the ISSE method, which was dropped from consideration due to inherent bias. A problem can occur if a panelist gives all correct or all incorrect ratings to the multiple-choice items, and all lowest score or all highest score estimates to constructed response items. In this case, the estimated individual cutpoint is either in the range of negative infinity or positive infinity. If this should occur, the program gives an arbitrary minimum or maximum cutpoint to the panelist in question. The boundaries are set at –4.0 and +4.0 on the θ scale. Again, this is a temporary fix. A discussion with the panelist in question is necessary to resolve the problem. If the panelist indeed believes in his or her judgement of student performance on the items, given the achievement level descriptions, then the extreme cutpoints would be allowed to stand unchanged.

### DRAWING PANELS FOR THE PILOT AND ALS STUDIES

The National Assessment Governing Board requires that the NAEP achievement levels-setting panelists be broadly representative. NAGB requires that 70% of the panelists be educators and 30% non-educators. Additionally, 55% of the educators should be classroom teachers in the K-12 level. For the pilot study, there were 20 panelists per grade level; and for the operational ALS procedure, there were 30 panelists per grade.

The lists of candidates for the panels were derived through a nomination process. The process of selecting panelists has four stages. In the first stage, samples of public school districts and private schools were drawn to serve as the basis for identifying the nominators. The school districts and private schools were drawn on the basis of the geographical region and demographic characteristics[5]. In the second stage, the potential nominators were identified. The potential nominators were persons who hold certain positions in the school districts and private schools that were sampled—mayors, superintendents, curriculum directors, and so forth. The potential nominators were sent packets of materials including nomination guidelines and forms, and they were invited to nominate up to four candidates per grade level. When the nomination process was completed, at the third stage of the panelist selection process, the nominees were screened to form the pool of candidates. Several criteria were specified in the guidelines for each type of panelists, and nominees had to meet those criteria in order to enter the pool for selection. At the last stage, the panelists were sampled, based on the targeted composition of the panel. The targeted composition of the panel included criteria such the proportion of panelists of each type (teacher, other educators, and general public), percentages of panelists in various racial/ethnic groups, percentages of geographical regions, and percentages of males and females.

### SAMPLING OF THE SCHOOL DISTRICTS

The 1997 Market Data Retrieval (MDR) dataset was used to draw the samples of school districts and private schools. The sampling had three criteria. First the number of districts sampled should be proportional to the distribution of the population across the four NAEP regions. Second, 15% of the

[5] Please see the *Design Document* (ACT, 1997a) for a more complete description of the process.

districts should have student enrollment of 25,000 or larger. Third, 15% of the districts should have at least 25% of their population living below the poverty level. The percentages for the large school districts and the low income school districts were applied to the marginals rather than within each region. That is, both 15% requirements (enrollment and low income) applied to the total sample of school districts drawn. The enrollment and income requirements were not applied to districts within each region.

The 1995 estimates of the census data showed the population distribution by region as 22%, 24%, 24% and 30% for Northeast, Southeast, Central, and West, respectively. However, instead of drawing school districts according to the proportion based on the population distribution, ACT decided to draw the sample based on the proportions of school districts of all four regions. This decision was made for two important reasons. First, the size of school districts varies sharply across the regions. Typically, the school districts in the Southeast region were larger, compared to the school districts in the Northeast. If 24% of the school districts were drawn from the Southeast, a much larger proportion of the population of the Southeast would be represented relative to that of other regions. It would also over-sample the school districts with enrollment of 25,000 or more from the Southeast. The second reason was there were more school districts with 25% or more of their population below the poverty level in the Southeast region. If 24% of the school districts were drawn from the Southeast region, the low income school districts were actually over-sampled. Indeed, almost all of the school districts in the low income category. Given these reasons, ACT revised the percentages of school districts to be drawn to be based on the proportion of the school districts represented in each region. These were, 22%, 12%, 36%, and 30% for Northeast, Southeast, Central, and West regions, respectively.

To draw the school districts, every school district was first assigned an identification code number using numbers randomly generated by an SAS random number function. When the random number matched the identification code, the corresponding school district was checked. The attributes of this school district were examined to determine whether they met the distribution requirements of the sample. For example, if the school district were from the West region, then the tally of the West region increased by one. The same was applied to the tallies of the large (enrollment of 25,000 or more) and the low income school districts (25% below poverty level). However, if any of the three marginal percentages had met the criteria, this school district would not be included in the sample and was eliminated from the pool. This was sampling without replacement. For example, if it were a school district from the West, and the percentage of West school districts already in the sample had reached 30%, this school district was thrown out. For another example, if the school district drawn were a larger school district from the West, this district would be thrown out if the percentage of the large school districts in the sample had reached 15% already. In other words, if one of the three marginal percentages had been met, then a school district with the attribute would not be included in the sample. Given this sampling procedure, the selection process became slower as a match was harder to find and the candidate pool became smaller. When the required number of school districts was met, the process stopped.

When a school district was selected it was randomly assigned to one of the three panelist types for identifying nominators: the teacher, the non-teacher educator, and the general public. If the required number of school districts for one of the panelist types had been met, the district was randomly assigned to one of the two remaining types. Finally, there would be only one type that had not been filled, and the remaining selections would all be assigned to this type.

The sampling of school districts for the 1998 NAEP Civics and Writing was conducted independently. There were only a few school districts that appeared in both samples and those were replaced. In all, there

were 130 school districts selected for Writing to nominate teachers, 15 to nominate nonteacher educators, and 100 to nominate the general public panelists. For Civics, 130 school districts were selected to nominate teachers, 20 to nominate nonteacher educators, and 114 to nominate the general public panelists. In addition to the districts sampled, nominators were identified in the nonteacher educator category from post-secondary institutions. Chief state school officers, state curriculum directors, and state assessment directors were identified as nominators of panelists from any district within the state for states having at least one district drawn in the sample.

Territories are not included in the MDR database. In order to include them in the sampling procedure, each territory was randomly assigned to a state. Each of the four regions was represented in the assignments, however. If Guam were assigned to Wisconsin and if a district in Wisconsin were selected for example, then Guam was included as a nominator location.

### SAMPLING OF THE PRIVATE SCHOOLS

Selection of the private schools followed the same procedure as selection of the public school districts, except on a smaller scale. In addition, there was no nomination for the general public type from the private school samples. The same MDR file was used since it included information for both the school districts and private schools. In all, there were 33 private school selected for Writing to nominate teachers, and 5 to nominate non-teacher educators. For Civics, there were 52 private school selected to nominate teachers, and 5 to nominate non-teacher educators.

### SAMPLING OF THE PANELIST FROM THE POOL OF NOMINEES

At the end of the nomination process, when all the information for the nominees had been collected, the sampling of the panelists started. The first task was to screen the qualifications of the nominees. Points were assigned to each of the nominees according to the screening criterion. Nominees with the highest ratings were given highest priority for selection.

There were six candidate pools each for the three grade levels for the two assessments. The sampling was conducted independently for each candidate pool. The sampling of the panelists was more constrained than the sampling of the school districts. This resulted from the skewed distribution of the characteristics of the nominees. For example, there were far fewer male teachers in grade four than there were female teachers. This made the balance between the genders difficult for selecting grade four panel members. Similarly, there were far fewer minority teachers and non-teacher educators nominated in the Central region than in the Southeast, and this made the balance among the regions and across racial/ethnic groups difficult.

Each nominee was assigned an identification code number. A random number was generated by the SAS random number function. When the two numbers matched, the nominee was checked as a potential panelist. The constraints on the panel composition included these.

1) equal proportion from all four regions;
2) equal proportion of males and females;
3) at least 20% minority panelists with different racial/ethnic identities;
4) 55% teacher, 15% non-teacher educator, and 30% general public.

All four constraints were tallied given the characteristics of the nominee selected. When one constraint had been met, any nominee with the same characteristic was excluded. The process continued until the required numbers were met.

Because participation in the NAEP achievement levels-setting process was voluntary, the selected panelists had to first agree to participate. Since some panelists would not be able or willing to participate, the potential panelists were over-sampled. In all, 50 panelists were sampled for each grade level of each content area for the operational ALS meeting. Additionally, after the selection of the ALS panel, a 30 member panel was selected for the pilot study for each grade level of each content area.

The sampling process did not stop when the potential panelists had been selected. First, the potential panelists were contacted regarding their willingness to participate. Despite the over-sampling, the composition of the panels did not always meet the criteria. The particular characteristics needed on the panel were evaluated, and a list of the nominees with the specific characteristics was printed out. The staff then contacted a nominee from this list and recruited a "replacement" panelist. The final stage of the panelist selections was not a random process because it was necessary to meet the requirements of both well-qualified panelists and panelists who were broadly representative. The panels were selected to maximized the qualifications of the panelists **and** the representativeness of the panels with respect to race/ethnicity, sex, and region while also meeting the NAGB requirements for panelists of different types (teachers, other educators, noneducators), and the ACT goal of selecting only one panelists from the same district.

## REFERENCES


ACT (1995). *Research studies on the achievement levels set for the 1994 NAEP in geography and U.S. history.* Iowa City, IA: Author.

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

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

Chen, W.H. & Pommerich, M. (1998).**?

U.S. Department of Education. Office of Educational Research and Improvement. National Center for Education Statistics. *The NAEP 1998 Reading Report Card for the Nation and the States,* NCES 1999-500, by P.L. Donahue, K.E. Voelkl, J.R. Campbell, & J. Mazzeo. Washington, DC: 1999.

Hanson, B.A. & Bay, L. (1999). *Classifying student performance as a method for setting achievement levels for the Writing NAEP*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, Montreal.

Hanson, B.A., Bay, L., & Loomis, S.C. (1998, April). *Booklet classification study.* Paper presented at the meeting of the National Council on Measurement in Education, San Diego, CA.

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.

Lewis, D.M., Mitzel, H.C., & Green, D.R. (1996, June). Standard setting: A bookmark approach. Paper presented at the CCSSO Large Scale Assessment Conference, Phoenix.

Loomis, S.C. (2000, June). *Research on the use of holistic methods for the NAEP achievement levels-setting process.* Paper presented at the Large-Scale Assessment Conference, Snowbird Resort, Utah.

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

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

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

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

Loomis, S.C. & Hanick, P.L. (2000e). *Developing achievement levels for the 1998 NAEP in civics: Final report*. Iowa City, IA: ACT.

Loomis, S.C. & Hanick, P.L. (2000f). *Developing achievement levels for the 1998 NAEP in writing: Final report*. Iowa City, IA: ACT.

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

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

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

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

**Table 1**
**Cutpoints Computed for Booklet Classifications:**
**Four Computational Procedures Compared**

| Round | Group | Level | Cutpoints | | | |
|---|---|---|---|---|---|---|
| | | | Collapsed Categories | Average Borderline | Weighted Collapsed & Borderline | Cubic Regression |
| 1 | A | Basic | 123.70 | 138.46 | 128.11 | 136.75 |
| | | Proficient | 154.26 | 167.26 | 157.10 | 165.96 |
| | | Advanced | 195.36 | 190.31 | 194.47 | 190.33 |
| | B | Basic | 124.71 | **137.33** | 129.67 | 137.15 |
| | | Proficient | 156.26 | 168.40 | 160.62 | 166.43 |
| | | Advanced | 185.31 | 190.65 | 187.14 | 190.27 |
| 2 | A* | Basic | 129.03 | 135.87 | 131.56 | 137.55 |
| | | Proficient | 151.88 | 169.46 | 156.84 | 167.40 |
| | | Advanced | 197.32 | 191.62 | 196.51 | 192.94 |
| | B | Basic | 127.57 | **138.93** | 131.48 | 135.73 |
| | | Proficient | 151.42 | 162.13 | 154.29 | 163.86 |
| | | Advanced | 189.28 | 191.64 | 189.96 | 189.69 |

Note: **Bold**-faced numbers were averages from fewer than ten judges.

* Percentages of students performing at or above the cutscores set in each round were reported to panelists in group A. Panelists in group B received this feedback *after* round 2 ratings were collected.

**Figure 1**
**Sample Cutpoints Using Angoff Method**

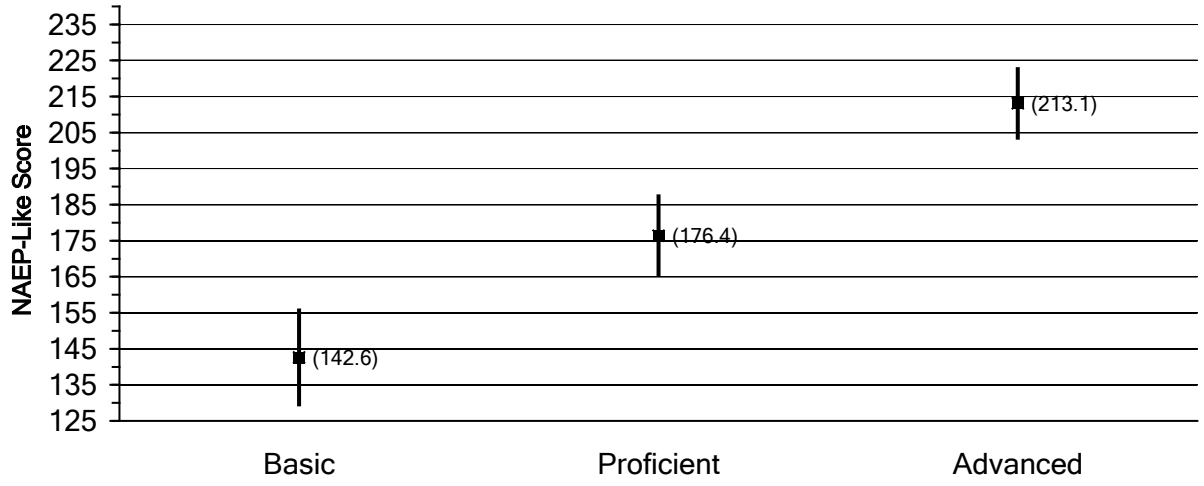



**Figure 2**
**Sample Cutpoints Using ISSE Method**

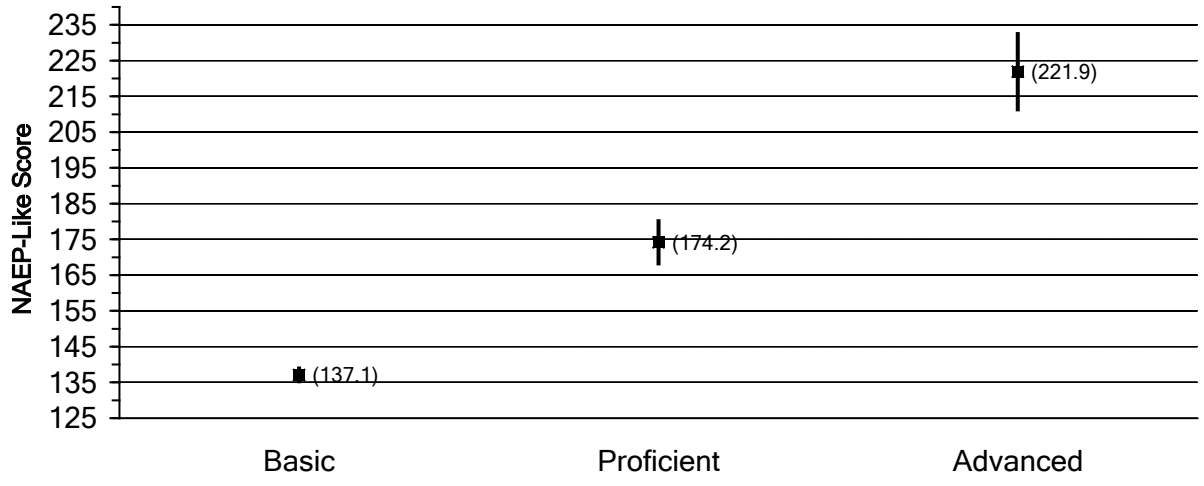


**Figure 3**
**Sample Item Map**

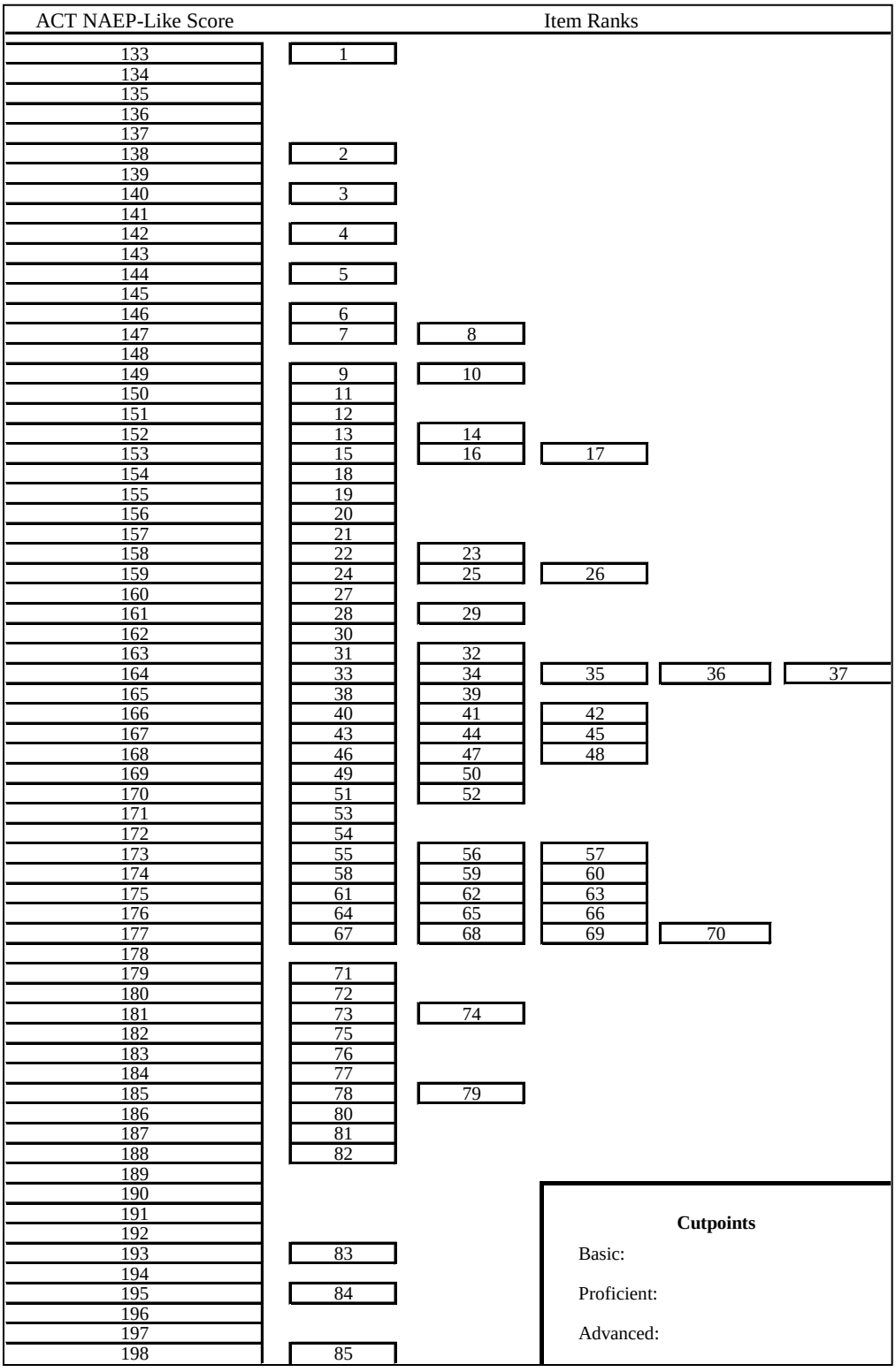


**Figure 4**
**Sample Item Mapping List**

| Rank | Block | Item Number | Score | ACT NAEP-Like Score | Item Description |
|---|---|---|---|---|---|
| 1 | Q12G8 | 1 | >=2 | 133 | Map: draw X where live OE |
| 3 | Q12G8 | 12 | | 140 | What is needed to make land arable MC |
| 5 | Q12G8 | 4 | | 144 | % S. America with 240 days for growth MC |
| 7 | Q12G8 | 3 | | 147 | Map: identify Mississippi River |
| 9 | Q2G5 | 4 | | 149 | How is earthquake intensity measured MC |
| 11 | Q2G5 | 3 | | 150 | Earthquake map: city most damaged MC |
| 13 | Q12G8 | 15 | >=2 | 152 | Draw map given certain features |
| 15 | Q12G8 | 2 | | 153 | Map: identify Lake Superior MC |
| 17 | Q2G4 | 14 | >=2 | 153 | One product: why does U.S. import OE |
| 19 | Q2G4 | 2 | | 155 | Why were river valleys settled |
| 21 | Q2G3 | 16 | >=2 | 157 | Diagram: identify landforms |
| 23 | Q2G3 | 8 | | 158 | Map: U.S. trade balance in 1990 MC |
| 25 | Q2G5 | 11 | >=2 | 159 | Why tropical deforestation |
| 27 | Q12G8 | 6 | | 160 | What is world's largest ocean |
| 29 | Q2G3 | 9 | | 161 | Maps: which shows most area MC |
| 31 | Q2G3 | 14 | >=2 | 163 | Compare country life expectancies OE |
| 33 | Q2G4 | 13 | | 164 | Factor in greenhouse effect |
| 35 | Q2G4 | 6 | | 164 | Why ancient towns built on hills MC |
| 37 | Q2G4 | 10 | | 164 | Map: locate city with most people MC |
| 39 | Q12G8 | 15 | >=4 | 165 | Draw map given certain features |
| 41 | Q12G8 | 11 | | 166 | Why do countries join the UN MC |
| 43 | Q2G5 | 15 | | 167 | Effect of El Nino on Peru's economy MC |
| 45 | Q2G3 | 5 | | 167 | Where Islam originated |
| 47 | Q2G5 | 14 | | 168 | How does El Nino affect Peru MC |
| 49 | Q2G5 | 18 | | 169 | What defines the Corn Belt MC |
| 51 | Q2G3 | 6 | | 170 | Influence of Islam |
| 53 | Q2G4 | 8 | >=2 | 171 | Pros and cons of nuclear energy OE |
| 55 | Q12G8 | 7 | | 173 | Map: where would large city develop MC |
| 57 | Q2G3 | 4 | >=2 | 173 | Graph: why urban populations changed OE |
| 59 | Q2G3 | 12 | >=2 | 174 | Explain pop. distribution in Egypt OE |
| 61 | Q2G3 | 10 | | 175 | Consumer demand conflicts with |
| 63 | Q2G4 | 12 | | 175 | Why is Quebec cultural region MC |
| 65 | Q2G5 | 11 | >=3 | 176 | Why tropical deforestation |
| 67 | Q2G4 | 3 | | 177 | How to predict where acid rain falls MC |
| 69 | Q2G5 | 9 | >=2 | 177 | Diagram: compare population patterns OE |
| 71 | Q2G4 | 15 | >=3 | 179 | Map: explain language patterns OE |
| 73 | Q2G4 | 4 | >=3 | 181 | Why is tundra hard to settle OE |
| 75 | Q2G5 | 16 | >=4 | 182 | Map: Why changes in U.S. pop. center OE |
| 77 | Q2G5 | 6 | >=3 | 184 | Which intl. product shown on map OE |
| 79 | Q2G3 | 14 | >=4 | 185 | Compare country life expectancies OE |
| 81 | Q2G5 | 12 | | 187 | What is the major religion in India MC |
| 83 | Q2G5 | 9 | >=3 | 193 | Diagram: compare population patterns OE |
| 85 | Q2G3 | 16 | >=3 | 198 | Diagram: identify landforms |

**Figure 5**
**Sample Reckase Chart**

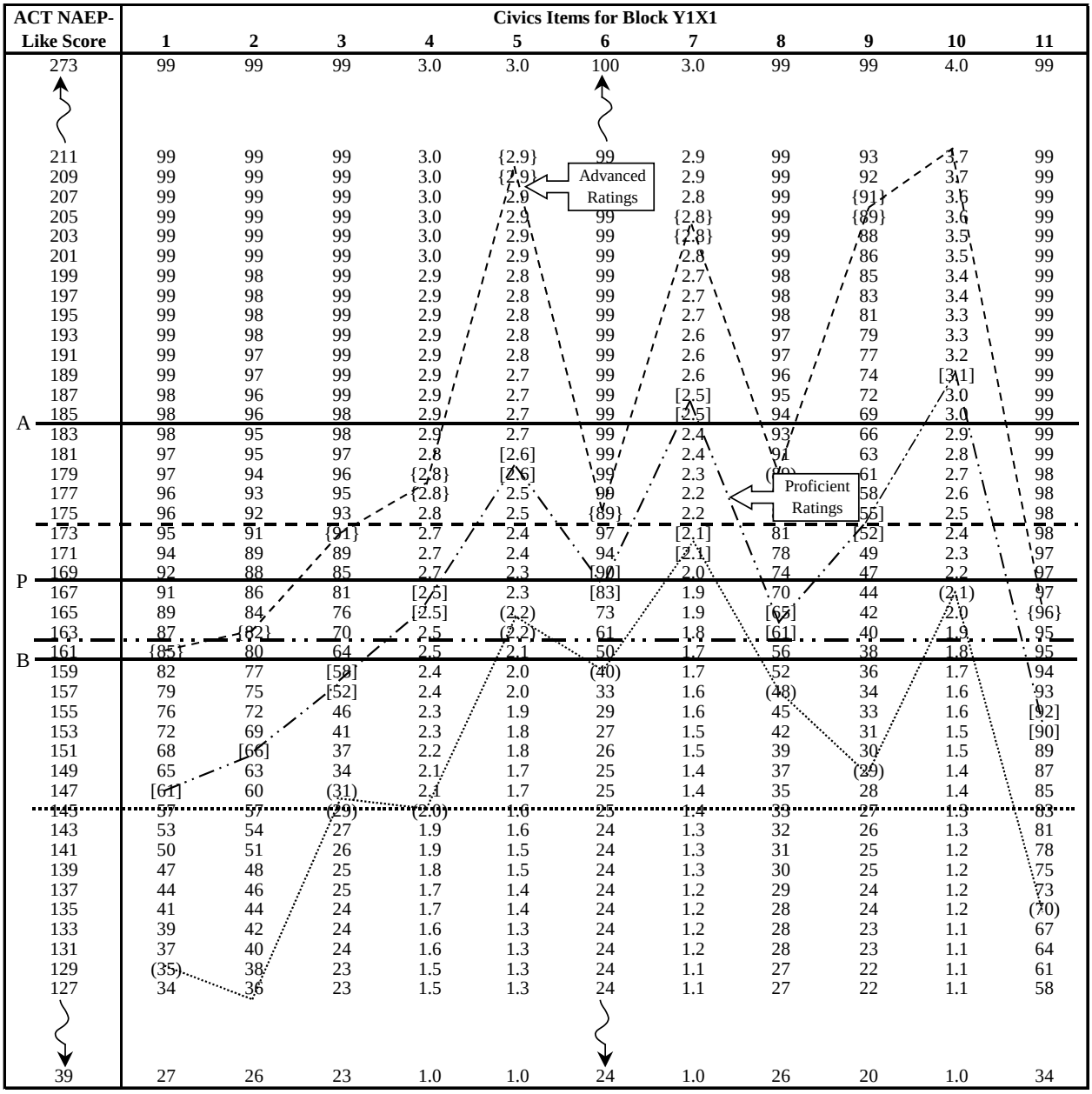
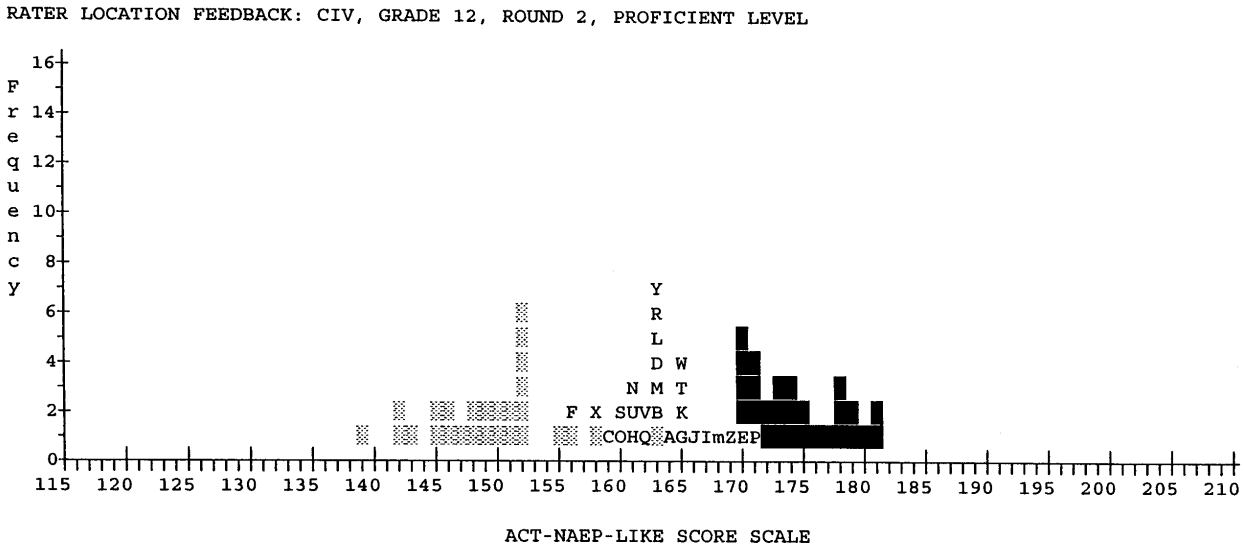


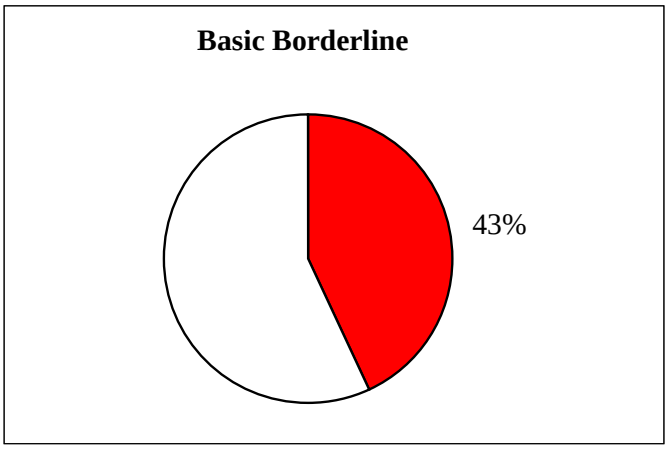
**Figure 6**
**Sample Rater Location Feedback**

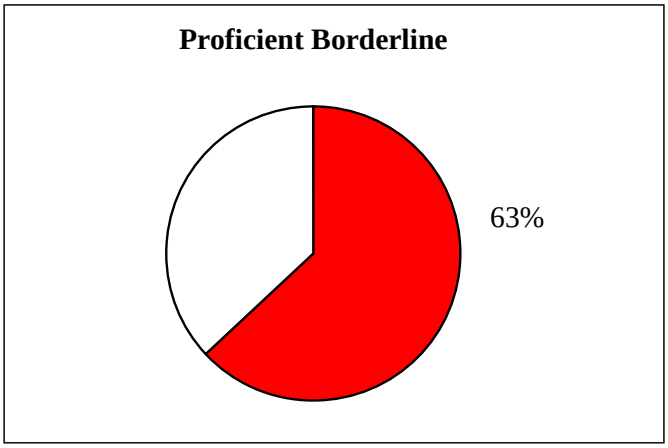


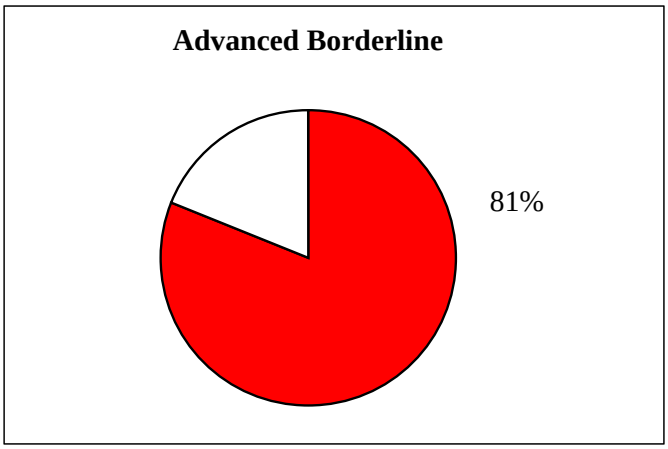

Legend: A-Z are secret IDs of panelists.
▒ represents rater location for Basic Achievement Level.
▮ represents rater location for Advanced Achievement Level.

**Figure 7**
**Sample Wholebooklet Feedback**

Grade X Group X Round X




These pie charts report information for the NAEP test booklet that you completed on Day One. The shaded portions of the charts show the percent of total possible points that students performing at the borderline of each achievement level would need to earn. The percents were estimated using the achievement levels cutpoints that your group set in this round.

**Figure 8**
**Sample Consequences Feedback**

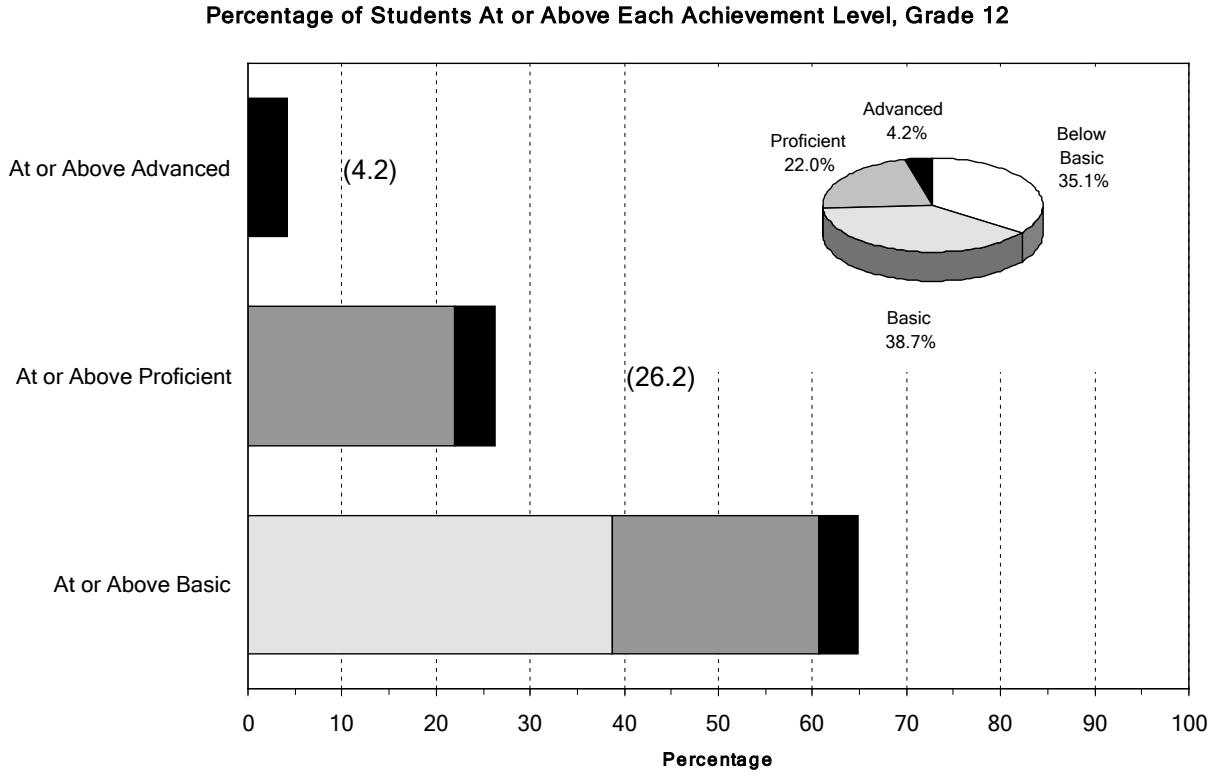


**Figure 9**
**Sample Exemplar Item List**

EXEMPLAR ITEM LIST FOR GRADE X
BASIC LEVEL EXEMPLAR ITEMS: PRIMARY LIST

| BLOCK | ITEM | | AVE PROB | ITEM DESCRIPTOR |
|---|---|---|---|---|
| U1C6 | 14 | | 84 | KNOWS WHO HAS RIGHT TO VOTE IN U.S. |
| U1C7 | 14 | | 65 | KNOWS WHO MAKES LAWS |
| U1C7 | 12 | >=3 | 70 | CAN IDENTIFY GOOD REASONS FOR BEING COP |
| U1C6 | 15 | | 72 | KNOWS PURPOSE OF UNITED NATIONS |
| U1C7 | 8 | | 71 | KNOWS REQUIREMENTS FOR CITIZENSHIP |
| U1C6 | 8 | | 68 | KNOWS MEANING OF PLEDGE |

BASIC LEVEL EXEMPLAR ITEMS: SECONDARY LIST

| | | | | |
|---|---|---|---|---|
| U1C6 | 1 | | 71 | KNOWS DEMOCRATIC METHOD OF CHOICE |
| U1C6 | 9 | | 74 | UNDERSTANDS CIVIC RESPONSIBILITY |
| U1C7 | 1 | | 64 | KNOWS PURPOSE OF LAW IS TO PROTECT PUBLIC |
| U1C6 | 5 | | 55 | KNOWS BENEFIT OF HIGHER TAXES |
| U1C7 | 3 | | 59 | KNOWS PURPOSE OF A RULE |

# KEY RECOMMENDATIONS BY THE TECHNICAL ADVISORY COMMITTEE ON STANDARD SETTING

Teri Fisher, Editor
ACT, Inc.

The following is a compilation of key recommendations by the Technical Advisory Committee on Standard Setting (TACSS) to the ACT Achievement Levels-Setting (ALS) project. The recommendations included here are taken from notes on TACSS meetings during ACT's contract with the National Assessment Governing Board (NAGB) beginning July 1997 and closing December 2000.

This is only one of many ways in which this information could be organized. The organization selected here is by topic, and the topics are arranged in the order in which issues arose in designing and implementing the ALS process. More complete notes are available in TACSS briefing materials, available from ACT.

TACSS members, along with their tenure on TACSS, are listed below. Please note that Professors Robert Brennan and Mark Reckase were both members of ACT's Technical Advisory Team (TAT) at the start of this contract. Dr. Brennan remained a member of TAT, and he served as one of two TAT representatives to TACSS. When Dr. Reckase left the ACT staff to join the faculty of Michigan State University, he was transferred from membership on the TAT to membership on the TACSS. Dr. Nancy Petersen became the second representative of TAT to TACSS. All members of the Technical Advisory Team are listed below.

ACT Project Staff are also listed below. Other regular observers and participants at TACSS meetings were the Contract Officer Sharif Shakrani and the Contract Officer's Technical Representative Mary Lyn Bourque, both of the NAGB. Dr. Andrew Kolstad represented the National Center for Education Statistics at the TACSS meetings.

**TECHNICAL ADVISORY COMMITTEE ON STANDARD SETTING (TACSS)**
William Brown, Educational Services, North Carolina
Barbara Dodd, University of Texas-Austin
Robert Forsyth, University of Iowa
Ronald Hambleton, University of Massachusetts-Amherst
John Mazzeo, Educational Testing Service
William Mehrens, Michigan State University
Jeffrey Nellhaus, Massachusetts Department of Education
Mark Reckase, Michigan State University
Douglas Rindone, Connecticut Department of Education
Wim van der Linden, University of Twente
Rebecca Zwick, University of California-Santa Barbara

**INTERNAL TECHNICAL ADVISORY TEAM (TAT)**
Robert Brennan, University of Iowa
Bradley Hanson, ACT
Nancy Petersen, ACT
Richard Sawyer, ACT
Catherine Welch, ACT

## SAMPLING PLAN AND RECRUITMENT OF PANELISTS

(12/97)

- To improve current sampling plan, ACT recommended using proportion of school districts rather than proportion of population; modification would yield a proportion of population at least as representative as that yielded by sampling districts proportional to population.
- Current method of sampling complies with NAGB guidelines and the general desire to have panels be broadly representative; but method does not guarantee that selected nominees will be as diverse as nominators—a concern of TACSS.
- TACSS concerned that all minorities might not be represented adequately using the current process for drawing panelists from nominee pool.
- Suggestion to improve scientific rigor of sampling process when selecting nominees in the future; sampling process involves three steps:
  - Step 1 - identifying nominators (this part of process is fine)
  - Step 2 - selecting from pool of nominees (this part of process needs improvement)
  - Step 3 - inviting nominees
- Recommendation to define parameters of interest in sampling population not only for nominators, but also panelists selected.
  - Specify what is a "good mix."
  - Elaborate plan to assure broad representation.
  - Target 20% minority representation among various racial/ethnic groups.
  - Perhaps use linear programming technique for sampling.
  - Precisely define "expert" quality of "outstanding" panelists.
- ACT must accurately report the sampling process that is currently used for drawing district samples to identify nominators without implying that the same scientific rigor is being used to select panelists.
  - **Word of caution**: Improved process must be realistic, given consistently low proportion of nominees recommended from certain regions and ethnic groups.
  - Suggest conduct post hoc examination of every step of sampling process to see why low response rate, including analysis of non-response.
  - **Decision:** NAGB should consider complete set of sampling issues and establish a policy on sampling requirements.

(6/98)

- TACSS requested more information about the point system ACT used for quantifying nominees' qualifications when selecting panelists.
- It was generally agreed that when reporting the proportions of school districts sampled by region, the approximate number of children who would be represented by these proportions should be included.
- It was suggested that ACT consider increasing the Hispanic/Latino representation to 10%.

## FOCUS GROUPS FOR REVIEW AND EVALUATION OF PRELIMINARY ALDS

(12/97)

- Proceeding as planned; generating recommendations for improving usefulness and reasonableness of preliminary ALDs for consideration by expert review panels.
- TACSS' concern pertains to possibility that because different types of writing are being assessed, there could be serious scaling problems with no clear alternatives in place if there are scaling difficulties.

- ❖ Also concern about possibility that judges could have strong opinions that do not fit with strictly conjunctive or compensatory model; no clear alternatives if judges' opinions show variations not easily modeled.
- ❖ Suggest ask focus groups to provide input on compensatory/noncompensatory issue.

(2/98)
- ❖ Send copy of recommended ALDs to framework planning committee members to review.
- ❖ Review item pool as it relates to recommended ALDs.
- ❖ Prepare complete, detailed account of the finalizing process to be available to larger community.

(3/98)
- ❖ Considerable discussion took place about potential problems in setting achievement levels for the Writing Assessment.
  - The scoring rubrics focus on measurable writing qualities and the ALDs focus mainly on more qualitative features. Therefore, there could be a disjuncture between the ALDs and the application of the scoring rubrics.
  - ACT has received comments from writing reviewers that indicate there might be a gap in the description of writing skill progression between grade 8 Basic and Proficient.
  - A final issue was discussed concerning how well the ALDs depicted first draft writing, as compared to finished writing.
    - It was generally agreed that if the writing content experts were aware of the fact that papers would be draft quality and wrote the ALDs with that in mind, then there is no problem with explicitly stating such.

## FORMAT FOR RECOMMENDATIONS TO NAGB

(12/97)
- ❖ discussion topics (no recommendations at this time)
- ❖ present a range of cutscores (e.g., Basic = 220-225)
- ❖ provide NAGB with point estimates (mean, median)
- ❖ provide distribution of judges' ratings (rater location data)
- ❖ include summary statistics
- ❖ "confidence level" is inappropriate and should not be included in reports to NAGB

## SCALING PROCEDURES ETS PLANS TO USE FOR 1998 WRITING NAEP

(2/98)
- ❖ Past results indicated very little difference in performance across types of writing.
  - In 1992 Writing NAEP, 4th grade narrative and persuasive were somewhat separate from other grades and types of writing.
- ❖ ETS still plans to scale unidimensionally.
- ETS wants scale based on largest number of prompts.
  - It is difficult to define scale with small sample of prompts.
- There are too few prompts in any one writing type to scale as a subscale.
- ❖ If results appear unusual for certain classes of prompts, then ETS will study irregularities.
- ❖ ETS encourages discussion and debate about scaling concerns <u>before</u> administration.

## RESEARCH STUDIES INCLUDED TO IDENTIFY METHODOLOGY AND COMPUTATIONAL PROCEDURES

### SIMULATION STUDY

(12/97)

- **Decision:** ACT will proceed with ISSE method for civics, but more simulation research is required to understand how ISSE rating method would work for writing; concern based on increased error due to limited number of writing prompts.

### FIELD TRIAL 1 FOR CIVICS AND WRITING

(12/97)

- Purpose of Field Trial 1 is simply to test out rating methods under consideration in order to determine whether panelists **can** do them, what problems are associated with each, and the relative ease/confidence in using each method.
- Because item mapping and paper selection are methods being used more frequently for standard setting, NAEP should consider them.
- **Decisions:**
  - FT1 for **civics** will compare **two** rating methods—ISSE and Mean Estimation.
  - FT1 for **writing** will compare **four** rating methods—ISSE, Mean Estimation, Grid, and Booklet Classification
    - ISSE method will have judges estimate score for each prompt.
    - Mean Estimation method will have judges estimate mean score for each prompt.
    - Grid method (6 x 6 matrix) will have judges indicate the score combinations on the matrix that would represent each level of performance for the two prompts, taken together.
    - Booklet Classification will classify student responses for both prompts contained in one form, with and without scores marked on booklet (one study group each).
      - Panelists sort writing into four piles (Below Basic, Basic, Proficient, and Advanced) using achievement level descriptions.
      - Panelists sort booklets within each achievement level into three levels (high, middle, low).
      - Booklets must be well-chosen: double-scored (*do double-scored booklets exist?*) or somehow represent high confidence in consistent scoring.
      - Panelists need general sense of scoring guides to understand papers, but do not need to be trained as scorers themselves.
      - **Word of caution**: Indications are that Booklet Classification method will result in higher cutscores.
- ACT will report to TACSS any difficulties with carrying out procedures of four methods.
- If results are inconclusive for writing, Paper Selection method will be used as fall back procedure.
- TACSS discussed numerous issues relative to the implementation of the various rating methods under consideration:
  - Computational procedures: Use same procedure to aggregate over panelists for Grid method and Booklet Classification method.
  - Standardize terminology in Design Document (e.g., define rounds more clearly as far as what happens, type of feedback given, and so forth).
  - Operationalize definitions of intrajudge and interjudge consistency in Design Document.

- Analyze similarity of cutpoints for multiple choice items vs. constructed response items in comparing rating methods.
- Concern raised by one member regarding need to determine if judges are rating well enough to be considered a "good judge."
- Search for ways to reduce time required by ALS process.

### FIELD TRIAL 1 FOR WRITING

(2/98)

❖ TACSS recommended to proceed with Writing Field Trial 1 even if fewer participants than anticipated.
- ACT needs a minimum 12 panelists to split into 2 groups (minimum group size about 6).
- If there are 12 panelists or more, compare ISSE and ME rating methods.
- If there are less than 12 panelists, have only one group that uses ISSE method.

(3/98)

The purpose of the field trial was to compare two rating methods, the mean estimation method (ME) and the item score string estimation method (ISSE). It was generally agreed that the results of the field trial for writing were not definitive in determining which of the two methods to drop from further consideration and which to carry forward for more research. There was some evidence to suggest that the raters found it easier to use ISSE than ME. It was suggested that ACT conduct further analyses of the results to examine combinations of factors that would relate the rating method to other variables of interest, such as rater consistency, panelist type, level of satisfaction with the process, comfort with the cutscores, and so forth.

### FIELD TRIAL 2 FOR CIVICS AND WRITING

(12/97)

❖ Purpose of Field Trial 2 is to examine use of item map with best rating method from Field Trial 1.

❖ Research on feedback conducted to inform NAGB in policy decisions with regard to the effect of consequence data on setting cutscores; impact of consequence data is more than a statistical question.

❖ Concern about establishing a consistent and acceptable response probability value that can be agreed upon by various NAEP groups for item mapping.
- Until agreement among all groups, ACT will use 74% for multiple choice items (using a correction for guessing) and 65% for extended response items.
- Eight members voted for RP=.74 (MC) and .65 (CR) and two members voted for RP=.80 for all items.

❖ For consequences feedback, panelists will use item maps to adjust cutscores. They will not be asked for recommendation about the distribution, *per se*, across achievement levels.

(3/98)

The upcoming field trials were planned to examine the interface between an item-by-item rating method when combined with an item mapping procedure. TACSS members generally agreed that more data were needed to inform them about the rating methods before they could make any recommendation regarding a method to study in conjunction with an item mapping procedure. There was a discussion related to how to map polytomous items to the score scale, and it was generally agreed that item maps should correspond in meaning to the form and format in which ratings are collected.

## RECOMMENDATIONS FOR FIELD TRIAL 2 FOR CIVICS

(3/98)

- After a lengthy and detailed discussion, TACSS recommended that ACT develop several different experimental designs for them to consider. ACT agreed to send the various designs to TACSS with an explanation of how the designs differ from each other and what direct comparisons can be made.
- It was generally agreed that there would be two experimental groups: one will use the ME rating method and the other will use the ISSE method. In addition, consequences data, distribution data, and item maps will be part of the experimental design.
- Of interest is when and how often the consequences information will be introduced in the achievement levels-setting process. One design should introduce consequences feedback after round two ratings, but not before round one.
- There should be 40 or more participants with a minimum of 10 per experimental condition.
- The booklet classification rating method will not be included in the field trial for civics. The protocol document will contain the precise instructions that will be given to panelists in addition to prepared answers to anticipated questions from the panelists.

(6/98)

There are three primary goals of the study:

- To compare the Mean Estimation (ME) and the Mark Reckase (MR) rating methods;
- To investigate the impact on ratings of consequences data provided at different times during the achievement levels-setting process; and
- To examine the interface between item maps and the Mean Estimation method.
- Strategies were discussed to increase the number of panelists who would be willing to participate in the field trials.
  - Sharif Shakrani, the Contract Officer, approved offering a $300 honorarium to each field trial panelist.
- It was generally agreed that, of the proposed Field Trial 2 designs, design #3 be implemented.
- Forty panelists will be required, ten for each of four experimental groups.
- Geography data from the 1994 NAEP will be used for the study.

### STUDY PROCEDURES

Step 1**:** Training the panelists, including taking the NAEP exam.

Step 2**:** Round One Ratings
All groups will rate each item using the ME rating method.

Step 3**:** Feedback after Round One Ratings
All groups will receive the standard forms of feedback:

1. P-value feedback
2. Rater location charts
3. Whole booklet exercise and feedback
4. Cutpoints and their standard deviations based on Round One ratings computed for each separate group (4 cutscores)

Groups A and C will receive consequences data after Round One:
The percentage of students who score at or above the cutscore computed from Round One ratings

Groups B and D will not receive consequences data after Round One.

Step 4**:** Round Two Ratings
Groups A and B will continue using the ME rating method, while groups C and D will switch to the Reckase rating method.

1. Groups A and B will adjust their ratings on the original ME rating sheets.
2. Groups C and D will circle their first round ratings on the Reckase charts, evaluate them in light of feedback data, and adjust their ratings on the original ME rating sheets.

- Panelists will receive one Reckase chart for each block of items at each achievement level.
- Charts will be color-coded for each achievement level.
- Panelists will receive a total of 12 Reckase charts (three achievement levels times four blocks).

Step 5**:** Feedback after Round Two Ratings
All groups will receive the standard forms of feedback:

1. Rater location charts
2. Whole booklet feedback
3. Cutpoints and their standard deviations based on Round Two ratings for each separate group (4 cutscores)

Groups A and C will receive consequences data for the second time:
The percentage of students who score at or above the cutscore computed from Round Two ratings.

Groups B and D will not receive consequences data after Round Two.

Step 6: Round Three Ratings
Groups A and B will be given Item Maps to assist them in adjusting their ratings.
They will mark their adjusted cutscores on the spaces provided on the IM forms.

Groups C and D will continue using the Reckase charts.

- Panelists will select a single row for each block that represents their ratings for each achievement level and mark it on the Reckase charts.
- The row will indicate their cutscore for each achievement level for each block.

Step 7: Feedback after Round Three Ratings
All groups will receive the standard forms of feedback:

1. Rater location charts
2. Whole booklet feedback
3. Cutpoints and their standard deviations based on Round Three ratings
   - Cutscores from the selected row for the Reckase method will be averaged for each rater across blocks and for all raters within each group (C and D).
   - Again, there will be four separate sets of cutscores.

Groups A and C will receive consequences data for the third time.

Groups B and D will receive consequences data for the first time.

Step 8: Group Agreement on the Final Cutpoints and Compute Averages

### DETAILS FOR IMPLEMENTING RECKASE CHARTS

- ❖ Provide a straight edge for panelists when referencing rows.
- ❖ Widely space columns to simplify reading the charts, particularly for civics.
- ❖ Italicize the midpoint in a string of five or more identical p-values to make it easier for panelists to locate the value that needs to be circled.
- ❖ Include a name or short content description for each item at the top of every column on the chart.
- ❖ Reckase charts for round two ratings will use a non-numerical row identifier, rather than the ACT-NAEP like scale score. Including the scale score could possibly influence the judges' ratings.
- ❖ Reckase charts for round three ratings will use the ACT-NAEP like scale score as the row identifiers.

### MARKING THE CHARTS

- ❖ Panelists will transfer their Round 1 and Round 2 ratings to the Reckase charts by circling the p-values that correspond to their ratings.
- ❖ Panelists should circle the lowest value if their rating is lower than the lowest p-value included on the chart.
- ❖ Panelists should circle the highest value if their rating is higher than the highest p-value included on the chart.
- ❖ Panelists should mark between p-values if their exact rating does not appear on the chart.
- ❖ Panelists should mark the midpoint (to be italicized) if their rating is within a string of identical p-values that appears an odd number of times.
- ❖ Panelists should mark between the two middle values if their rating is within a string of identical p-values that appears an even number of times.

### FACILITATORS' INSTRUCTIONS FOR THE RECKASE METHOD

- ❖ Facilitators will explain item discrimination as related to a string of identical p-values.
- ❖ Facilitators will encourage panelists to focus on the items: their content, distribution, relationship to other items, and overall characteristics.
- ❖ Facilitators will explain the relationship between rater consistency and a single row on the Reckase charts.
- ❖ Facilitators will explain the need for panelists to scrutinize "outliers."
- ❖ Facilitators will direct panelists to check for variation in their ratings for dichotomous and polytomous items.
- ❖ Facilitators will direct panelists to consider those ratings for which they have the most confidence when drawing their cutscore line.

## FIELD TRIAL 2 FOR WRITING

(2/98)

Key points in discussion about which methods should be investigated:

- ❖ It is still unknown how to compute cutscores for Grid and Booklet Classification methods.
- ❖ Some members in favor of studying unknown methods since already familiar with ME and ISSE.
- ❖ Grid method introduces the potential for non-compensatory strategies that would be contrary to the scaling model.
- ❖ ACT questioned whether TACSS only wants to consider compensatory methods since those are consistent with scaling process.
- ❖ A suggestion was made to examine Grid method using simulation data to map cutscores. If able to compute cutscores, then continue to examine Grid method with real panelists.
- ❖ ACT wants to decide rating method before Pilot Study.

- Several members expressed support to study Booklet Classification method.
  - Support is based on fact that method is truly different from ME and ISSE.
- Finding appropriate student booklets remains problematic for field trials.
  - Flawed prompts from 1992 Writing NAEP distract and confuse panelists.
  - "Virtual booklet" idea rejected as possible source of student booklets.

❖ It was agreed to schedule Writing Field Trial 2 for summer to study Booklet Classification method.
- Scored booklets from 1998 Writing NAEP will be available by June.

❖ TACSS debated the issue of using test booklets with scores vs. without scores. The issue was not resolved.
- Points in favor of using booklets with scores:
  - Scores help overcome tendency to set higher cutscores.
  - Scores help to put booklets into "correct level" and avoid the fatal flaw of using non-compensatory model.
- Points in favor of using booklets without scores:
  - No scores help to keep panelists focused on task and curb debate on scoring.
  - No scores increase demand by panelists to score items.

❖ One member suggested different procedures to use in Field Trial 2 study for Booklet Classification method.
- Rank order booklets, based on scores.
- Panelists match ordered booklets with performance described in ALDs.
- Procedure is similar to Bookmark method, but at booklet level.
- Procedure would bring efficiency to process without having scores marked on booklets.
- Although there are many details to work out, group seemed to like general idea.

❖ Suggestions for how to address proposed research on "order effect" in rating levels:
- If use "Bookmark method" described above, order effect is no longer an issue.
- Use factorial design to determine what order is the "right" order.
- One member recalled that research has already been conducted on this topic by the National Academy of Education. ACT will check research available.

❖ ACT should determine what states have been using, with respect to order by type of writing task.
- There was agreement that examining order effect was not worth conducting a separate study. Instead, ACT should build order condition into study already planned.

(3/98)

❖ TACSS recommended that ACT conduct another field trial for writing using the booklet classification rating method. TACSS was encouraged to think about recommendations for the experimental design for the writing field trial using the Booklet Classification rating method. Design issues include ranking the booklets, using booklets with scores and/or without scores, the appropriate mix of booklets, the number of booklets panelists should read, how many times panelists classify a given booklet, using a sorting or a rating task, and so forth.

❖ There was general agreement that the 1992 writing data should be used for the writing field trial, rather than the 1998 data. Because of indecisive results from Field Trial 1 and because of questions about the calibration of Basic and Proficient ALDs for grade 8, TACSS recommended that an additional grade level be included in Field Trial 2 for writing. Panelists for the study will be recruited for grades 8 and 12.

(6/98)

- ❖ Given the added number of panelists that would be necessary to recruit for Field Trial 2 for Writing, and the general uncertainty associated with the Grid rating method, it was agreed that the Grid method be dropped from further investigation by ACT at this time. Efforts will focus on researching the Booklet Classification method and the Mark Reckase method in Field Trial 2 for Writing.

- ❖ The discussion centered on making recommendations for the implementation of the Booklet Classification method. It was agreed that design #2 will be implemented for Field Trial 2 for Writing. The 1992 eighth grade writing NAEP data will be used for the study. No data from 1998 NAEP available for feedback. The achievement level descriptions will be those developed for the 1998 Writing NAEP. It was recommended that computational methods be explored further for calculating a cutscore using the Booklet Classification (BC) method.

There are two primary purposes of the study:

- ❖ To compare the Booklet Classification (BC) and the Mark Reckase (MR) methods.
- ❖ To investigate the impact on ratings (MR method) and classifications (BC method) of consequences data provided at different times during the achievement levels-setting process.

**STUDY PROCEDURES**

Step 1: Training the panelists, including taking the NAEP exam

Step 2: Round One
Groups A and B will use the BC method (classifying booklets into seven categories).
Groups C and D will use the MR method (rating performance on individual prompts).

Step 3: Feedback after Round One
All groups will receive the standard forms of feedback:

1. P-value feedback at the booklet level (percent of total possible points), as well as at the prompt level
2. Rater location charts
3. Whole booklet exercise and feedback
4. Cutpoints and their standard deviations based on Round One, computed for each separate group (4 cutscores)

Groups A and C will receive consequences data after Round One:
The percentage of students who score at or above the cutscore computed from Round One

Groups B and D will not receive consequences data after Round One.

Step 4: Round Two
Groups A and B will continue using the Booklet Classificaiton method.

1. Groups will discuss their booklet classifications.
2. Groups will evaluate their classifications in light of feedback data.
3. Groups will adjust their classifications by marking the original classification sheets.

Groups C and D will continue using the Reckase method.

1. Groups will circle their first round ratings on the Reckase charts, evaluate them in light of feedback data, and adjust their ratings on the original rating sheets.

2. All writing prompts will appear on one Reckase Chart for a given level.
3. Charts will be color coded for each achievement level.

Step 5: Feedback after Round Two
Groups will receive the standard forms of feedback:

1. Rater location charts
2. Whole booklet feedback
3. Cutpoints and their standard deviations based on Round Two for each separate group (4 cutscores)

Consequences data will be given to both Booklet Classification groups after Round 2:

1. Group A will receive consequences data for the second time.
2. Group B will receive consequences data for the first time.

Consequences data will be provided to the same Reckase group:

1. Group C will receive consequences data for the second time.
2. Group D will not receive consequences data after Round Two.

Step 6: Round Three

Groups A and B have finished classifying booklets and will reach agreement on Final Cutpoints.

Groups C and D will continue using the Reckase method:

- Panelists will select a single row that represents their ratings for each achievement level and mark it on the Reckase charts.
- The row will indicate their cutscores for each achievement level.

Step 7: Feedback after Round Three

Groups C and D will receive the standard forms of feedback:

1. Rater location charts
2. Whole booklet feedback
3. Cutpoints and their standard deviations based on Round Three "ratings" for which panelists select a cutscore for the grade level for their rating group.

Both Reckase groups will receive consequences data after Round Three:

1. Group C for the third time.
2. Group D for the first time.

Step 8**:** Groups C and D using the Reckase method will reach agreement on Final Cutpoints.

### BOOKLET CLASSIFICATION IMPLEMENTATION DETAILS

*Booklets:*

- ❖ Booklets will be rank ordered within each form, but not scored.
- ❖ Booklets will represent a uniform distribution that will cover the full range of achievement for each form.
- ❖ All prompts will be included in the sample of booklets, but not every combination of prompts will be included.

- ❖ Each panelist will rate 40 booklets in all; 20 for each of 2 forms.
- ❖ There are 36 different score combinations for which theta values will be estimated.
- ❖ Estimated theta values will be used as the booklet scores when creating the distribution of booklets.
- ❖ The distribution of student performance represented by consequences data will be based on plausible values.

*Classifications:*

- ❖ Booklets will be ordered from lowest to highest performance.
- ❖ Panelists will classify booklets directly into seven categories on first reading, with allowance for individual differences in the exact classification method.
- ❖ Panelists will be instructed to keep notes on each booklet for discussion purposes.
- ❖ Panelists must decide a category for every booklet: there will be no "undecided" booklets.

*Computations of Cutscores:*

- ❖ The algorithm used for computing cutscores will be based on all of the booklets, rather than the data from only the borderline booklets.
- ❖ Several alternative methods for computing the cutscores for the Booklet Classification method should be compared for "robustness" of model fit.
- ❖ Cutscores should be computed for each panelist individually, creating a distribution of cutscores which will be averaged.

### ISSE UPDATE

(4/98)

- ❖ General consensus of TACSS that would be very easy to get cutscores that would be lower for Basic and higher for Advanced. Item maps, booklet classification, and mean estimation believed to be more promising. Continued concern with RP issue in creating item maps. Mean estimation experience shows that it would probably work for civics; possibly use booklet classification for writing. ACT agreed to provide TACSS with a description of the possibilities for review the following week.

(6/98)

- ❖ The bias inherent in the ISSE rating method was reviewed and discussed. It was determined that the ISSE method produces cutscores that will be lower for Basic and higher for Advanced relative to cutscores produced by the modified-Angoff method and relative to the "true" cutscores for raters. Further examination of the relationship between panelists' ISSE ratings and their responses to evaluation questions indicated no striking patterns. The decision was retained by TACSS to drop plans for further research using the ISSE method.

### COMPUTING CUTSCORES WITH BOOKLET CLASSIFICATION METHOD

(9/98)

- ❖ Brad Hanson's paper was discussed. The paper describes the application of the Hambleton, *et al* method for computing NAEP achievement level cutpoints using the booklet classification data collected in the Writing Field Trial 2. This computational method was not endorsed by TACSS. Using the collapsed category and a cubic regression were thought to be unsound practices.

### RECOMMEND ACT ADD FIELD TRIAL 3

(12/97)

- ❖ Purpose of Field Trial 3 is to examine the impact on the cutscores set when panelists do not start ratings at the Basic level.

- **Decision:** use whichever rating method emerges from Field Trials with two rounds of ratings under four manipulations for order effect: 1) BPA, 2) PBA, 3) PAB, 4) APB.
- **Suggestions**: check with states to determine the order they use; discuss with panelists to determine sources of differences (if any).
- Extensive discussion regarding instructions for modifications in ratings following round one.

## RECOMMENDED PROCEDURES FOR CIVICS PILOT STUDY

(7/98)

The decision was made to use Reckase Charts, but not the Reckase Method. All aspects of the ME method with the Reckase Charts will be implemented in the Civics Pilot Study the same way that they were implemented in Field Trial 2, *except* the following:

- Panelists will write borderline achievement levels descriptions.
- Panelists will participate in a paper selection exercise using 3 student papers at each score point for each constructed response item in their item rating pool.
- In addition to the "standard" forms of feedback, panelists will be given the Reckase Charts to mark their ratings from rounds one and two. The charts will display the ACT NAEP-like scale for all rounds, rather than an alphabet scale. Panelists will not be instructed to select a single row on the chart that represents their ratings for round three, as they did for the field trial.
- After the third round of ratings, panelists will be presented with information about the consequences of their own cutscores for each level based on their round three ratings.
- The cutscores based on individual consequences data will be the ones recommended to NAGB, and will be used for identifying items in the pool of exemplars.
- In addition to the standard forms of feedback, each grade group will receive individual consequences data for members of their group, using the same format as that used for rater location charts.
    - Note: Loomis later reported that only individual consequences data were given to panelists during the civics pilot study. The group consequences data were inadvertently omitted in the round 3 feedback session.

## RECOMMENDED PROCEDURES FOR WRITING PILOT STUDY

(7/98)

The decision was made to use Reckase Charts, but not the Reckase Method. All aspects of the ME method with Reckase Charts will be implemented in the Writing Pilot Study the same way as will be implemented in the Civics Pilot Study, *except* the following:

- Training for panelists will include a Paper Selection Exercise (PSE), and the student papers used during the exercise will be available to panelists for reference during the rating periods.
- For the PSE, there will be three papers at each score point for a total of 216 student papers.
- The PSE will require panelists to classify one essay at the borderline from the set of student papers.
- Training in the ALDs for writing will be modified. Panelists will be asked to decide whether a student response best fits the achievement description of writing at the Basic, Proficient, or Advanced level. The papers used for this part of the exercise will be one of two prompts included in the booklet form used to take the NAEP exam on Day One and later for the Whole Booklet Exercise and Feedback. Panelists will then be asked to classify the booklet, as a whole. Staff will work out details.
- The Reckase Charts for writing will include the prompt type for each item.
- ACT will consider the option of panelists marking their ratings for all three achievement levels on one Reckase Chart by using colored markers for the different levels.

- ❖ Instructions to panelists will be refined further, especially the explanation of the Reckase Charts. Particular attention will be paid to the instructions that include the concepts of "consistency," "weighting," and "implicit."

(9/98)

- ❖ The instructions for the Reckase Charts were finalized, as follows: "A student varies in performance on prompts. To be considered Basic/Proficient/Advanced, his or her average performance would have to be above *Y*. This means that some performance can be below *Y*, but other performance would have to be above *Y* so that the average would be above *Y*. If you believe the average level of performance is too low, you need to raise your ratings. If you believe the average performance is too high, you should lower your ratings."
- ❖ TACSS reviewed the division of prompts for the Writing Pilot Study along with the overall training exercises for panelists. A modification was suggested to the ALD Exercise when panelists classify the first prompt for 10 student papers followed by classifying the whole booklet of the same 10 students. Rather than use a booklet that was scored 3 for one prompt and "off-task" for the other, it was suggested that a booklet scored 3/3 should be substituted for training Group B, 4$^{th}$ Grade.
- ❖ It was generally agreed that the Paper Classification Exercise should be conducted in two parts. The first part will require panelists to classify papers into four categories (Below Basic, Basic, Proficient, and Advanced). The second part will require panelists to select the papers from each category to represent performance at the borderline of each of the three achievement levels. If no paper represents the borderline, none will be classified there. In addition to the usual purposes of the Paper Classification Exercise, the activity also will provide input for selecting exemplar performances. ACT will record the classification of papers that were written in response to common blocks, which will be used in the exemplar performance selection process.
- ❖ The term "P-value data" will be changed to "student performance data."
- ❖ Because the score distribution is "lumpy," it will be difficult to find papers to fit the criterion for selecting booklets at or near the borderline to show for the Whole Booklet Exercise. It was recommended that theta estimates be used to identify booklets and to report the percentage of total points associated with the cutscores for Whole Booklet Feedback. It was also recommended that theta estimates would be used to identify booklets and to report the percentage of total points associated with the cutscores for Whole Booklet Feedback.
- ❖ It was decided not to provide booklet level performance data to panelists as an added source of information.
- ❖ It was recommended that instructions for the Reckase Charts should include an explanation of *average* performance as a means of addressing the issue of compensatory scoring. The exact instructions for this issue are included on page 3 in the *Discussion Summary* ("Plan for the Writing Pilot Study").
- ❖ The same statistical criterion used to select exemplar items since 1994 will be used to select exemplar performance for the Writing Pilot Study. Papers from the Paper Classification Exercise with scores meeting the statistical criterion for consideration will be available for panelists to review and accept or reject.
- ❖ The booklets and score combinations to be used for panelists' training exercises were presented to TACSS for evaluation and were approved. These exercises are being modified for writing so that panelists review papers and then booklets written by the same students. The ALD training will consist of presenting panelists with the general definitions of achievement levels, scoring rubrics, student papers, and student booklets. Panelists will classify the first prompt of ten student booklets. Next they will classify the performance of the student as a whole. Panelists will not see examples of all of the possible score combinations for student booklets for this exercise because they will not be available. After classifying student papers, panelists will have the opportunity to discuss the reasons for their

classifications. Panelists will mark their classifications on a special form as part of the exercise. They will be given classifications for the single prompt to compare to their classifications of the student's booklet. Discussion will focus on comparing the two classifications.

- ❖ The distribution of scores on the Writing NAEP is unusual in that scores cluster in clumps rather than spread out across a normal distribution curve. This will cause some variation in the usual "within two percentage points" criterion followed to select booklets at or near the borderline to show in the Whole Booklet Exercise. Panelists may have difficulty in differentiating between levels if the criterion is increased to four percentage points. To further complicate the matter, there are limited booklet score combinations. TACSS recommended that ACT pick the booklets based on theta estimates that would best represent the cutscores. The pie charts used for Whole Booklet Feedback, however, would not change. It was agreed to try out this procedure in the pilot study and see how it works.
- ❖ In selecting exemplar performances for writing, ACT will follow the same statistical criterion used since 1994. That is, the 50% average RP across the range will be applied to writing tasks for selection of exemplar performances. Panelists will be presented a list of papers having scores that meet the statistical criterion. Three common block prompts, one for each type of writing, will be identified for use in selecting exemplar performances. Panelists will be given feedback information in the form of the frequency with which specific papers having the qualifying scores were classified at different achievement levels or borderlines. The "veto" method will allow panelists to exclude papers deemed to be poor or inappropriate illustrations of student performance for each level.

## RECOMMENDATIONS FOR CIVICS ALS PROCESS

(9/98)

- ❖ Panelists will write borderline achievement levels descriptions.
- ❖ To reduce the number of papers reviewed by panelists and to provide more time for group discussion of the papers, only the common blocks of items will be used for the Paper Selection Exercise. Scored papers from other constructed response items in the rating pool will be provided for review prior to Round One ratings.
- ❖ After taking the NAEP exam, panelists will review the scoring rubrics to check their work (as usual) without receiving additional feedback.
- ❖ The order of rating item blocks will not be reversed during the rating process (to control for a possible fatigue effect in panelists).
- ❖ ACT should test the following order of presenting feedback before implementing the ALS:
    a. Cutpoints
    b. Reckase Charts and P-values presented together
    c. Rater Location Data
    d. Whole Booklet Feedback and Exercise
- ❖ Individual Reckase Charts will be generated electronically and marked for panelists. Panelists will connect their ratings by hand to help identify rating patterns on the Charts.
- ❖ Instructions for the Reckase Charts will remain the same for the ALS as those used in the PS. Instructions focus on panelists' confirming that their ratings accurately reflect their judgment by considering the item content, item format, group ratings, and individual ratings.
- ❖ The Preview Rating Session was a difficult exercise for most of the panelists to understand. It was thought that the difficulty stemmed from panelists working with items and ALDs, and then switching to working with student performance. It was suggested that the Preview Rating Session be replaced by a group presentation early in the ALS meeting giving examples that illustrate the rating process.
- ❖ It was questioned if the Reckase Charts were introduced in the proper sequence. There remains an ongoing concern about panelists receiving too much feedback information at one time. TACSS

recommended that the Reckase Charts be used as a form of feedback for the ALS meeting, rather than as a step in the rating process. TACSS instructed ACT to test sequences of feedback to determine the optimal order.

- ❖ TACSS approved ACT's plan to provide feedback from the Paper Selection Exercise. It was suggested that a tally be included for the Paper Selection Exercise. Not only would panelists discover which papers had been selected to represent performance at the borderline of Basic, Proficient, or Advanced by other panelists, but they would also have a context for more meaningful discussion about the selections.

## RECOMMENDATIONS FOR WRITING ALS PROCESS

(10/98)

- ❖ Use the term "types" of writing rather than "genre."
- ❖ Do not use booklets that contain a prompt that is scored zero for "off-task" or "incorrect."
- ❖ The paper classification exercise will be conducted the same way for the ALS as was done for the PS. Only common blocks will be used. Panelists will look at all of the constructed response items from the two common blocks and classify papers from the two (out of five) blocks. Scored papers from the other three blocks will be available to panelists for reference.
- ❖ ACT will continue to keep tallies of panelists' classifications to use as part of the exemplar selection process.
- ❖ ACT will label and explain cutscores and standard deviations feedback information.
- ❖ A legend will be added that states that the box represents the cutpoint (the average for the grade group) and the line represents the variation around the mean.
- ❖ The term *Standard Deviation* will be removed from the title of the feedback sheet.
- ❖ It will not be necessary to revise the labels to include "Borderline" for each level, or replace "Cutpoint" with "Achievement Level."
- ❖ ACT will add a legend to the rater location charts to define the symbols.
- ❖ The axes labels (i.e. Frequency and ACT-NAEP Like Scale) will be in bold.
- ❖ If possible, all 3 charts will appear on one page.
- ❖ ACT will change the whole booklet feedback from a 3-D pie chart to a 2 dimensional circle. Even though many other suggestions were discussed, it was agreed that no further changes need to be made to the whole booklet information format.
- ❖ Instructions to panelists will be revised to improve understanding of "average scores" earlier in the process. ACT will add a pencil and paper task for how to compute an average.
- ❖ ACT will continue to conduct the three ALD exercises that involve types of writing, student papers, and student booklets.
- ❖ When training panelists in ALD exercises, ACT will include statements to panelists about what is NOT intended, as well as the purpose of the exercises (i.e. "It is not sensible to associate a score of 3 only with Basic achievement.").
- ❖ Panelists need to understand the concept that a student paper scored 3, for example, could be 2.8 if the prompt is easy or 3.4 if the prompt is hard. Facilitators must emerse panelists in this logic until they understand the concept fully. To clarify this concept, ACT should consider instructions using teachers' already-formed understanding of weighting particular essays or test items more heavily than others. Perhaps using the Olympic diving example included in one panelist's comments or an example from the figure skating scoring method would be helpful.
- ❖ Content facilitators must be sure that when panelists write borderline descriptors, the descriptors calibrate accurately with the ALDs.

- ❖ The Reckase Charts will be considered part of feedback information, as in the PS, rather than a separate step in the ALS process.
- ❖ Showing the Charts to panelists earlier in the process could help panelists understand the concept that student scores are variable at the same level of performance.
- ❖ ACT will revise the individual cutpoints and consequence data sheet. The percentages at or above Basic, Proficient, and Advanced should be less than 6 decimal places, which is how it is displayed now. The percents will appear in their own column next to the cutpoints and both will be labeled. The rows should be double space so that there is a blank row between raters' data. The columns should be labeled with the headers Basic, Proficient, and Advanced.
- ❖ The panelists will not receive added sources of information from related assessments, like TIMSS data, AP data, State NAEP data, and so forth.
- ❖ It is not of great importance for panelists to discuss their individual consequences data. However, there should be enough time for panelists to discuss the group consequences data.
- ❖ ACT should ask the compensatory/conjunctive questions at all levels, not just Advanced.

## ACT Recommendation to NAGB that Consequences Data be Introduced Earlier

- ❖ ACT will recommend to NAGB that grade level consequences data should be presented to panelists after Round 2 along with the other forms of standard feedback. Panelists will not recommend cutscores at that time, but continue rating for Round Three.
- ❖ Panelists will be able to use the Reckase Charts if they wish to see the ratings that are associated with higher or lower cutscores when producing Round 3 ratings.
- ❖ Panelists will receive consequences data again after Round 3, when they will be asked for cutscores.
- ❖ To improve the timing sequence, panelists will complete the cutscore recommendations section of the consequences questionnaires, but will write comments to the open-ended questions later, at the conclusion of the process.
- ❖ Consequences data will be presented in the format used for rater location charts with the addition of the percentages at or above each achievement level.
- ❖ Panelists should see consequences data for all grade groups.
- ❖ A conference call will be arranged with the ALC from NAGB to discuss their response to this recommendation before the ALS.

## General Discussion about Civics ALS

(1/99)

- ❖ One TACSS member questioned what was done differently in the 1998 ALS process that could explain the more "realistic" results. He observed that standards usually start out very high, but this was not the case for the 1998 ALS results. In response to the question, it was noted that panelists were given about the same information they have always been given. They were informed of student motivational issues, as usual. Bourque did discuss "reasonableness" during her Day 1 presentation. The term "world-class standards" was not mentioned when instructing panelists, but that is not a change in procedure. TACSS judged it to be unlikely that these factors would have had a great impact on the overall ALS results. Loomis noted that many ALS panelists have been involved in setting state standards. The 1998 panelists arrived at the ALS meeting with a much different mind-set than panelists from earlier meetings. This could effect "realistic" cutscores from the start. As part of the advance materials, panelists were sent the new *NAEP Guide,* which is a very informative booklet that

answers many questions about NAEP. Perhaps panelists are better informed now and have more "realistic" expectations.

## TACSS Comments about Civics ALS Analyses

(1/99)

- TACSS recommended that when analyzing data related to the discrepancies between round 3 actual ratings and expected ratings (Reckase Charts), the differences should be expressed in terms of the ACT NAEP-like scale. ACT should consider comparing the "peaks and valleys" of the Reckase Charts for both writing and civics ratings. Bourque suggested that ACT look at the empirical "difficulty" level of the items and compare that to how panelists rated the items. The comparison will determine if panelists "appreciated" the level of difficulty of the items. It was suggested that the National Academy of Education NAEP Evaluation Report be used for reference when planning this comparative analysis.

- Bourque also inquired about analysis of standard error (SE), which is a study that is currently underway by ACT staff. The SE study will not be completed before NAGB makes a decision about achievement levels. Because there are many dependencies in the data, it would be difficult to do extensive analyses with computations of SE. The Group A versus Group B study seems to be the best approach. It was suggested that rather than standard error analysis, it should be called "sensitivity analysis" of error.

## TACSS Comments about Writing ALS Analyses

(1/99)

- Conducting tests using inferential statistics was not advised. Loomis asked TACSS for advice in determining which analyses are appropriate, given the volume of data generated during the ALS meeting. Because panelists discuss their ratings after round 1, the data from subsequent rounds become statistically dependent. It was generally agreed that for the purpose of informing TACSS, the analyses are very useful, even if the data do not represent statistically independent events.

- TACSS suggested a visual examination of the means to determine whether there were any unusual results or likely patterns of results. If, for example, means appeared to differ by as much as .5 SD, then that might signal a need for further analyses.

- TACSS noted that the Group A versus Group B comparisons provide compelling data regarding the replicability of the results.

- TACSS recommended that the data be converted to the ACT NAEP-Like score scale for computing changes in ratings between rounds and for computing discrepancies between "expected" and "actual" ratings. The "expected" values are based on Reckase Chart data, and the "actual" values are the subsequent rating data. Using the ACT NAEP-Like scale values would provide a common unit of analysis across all subjects and it would eliminate the need to transform constructed response data to compare to multiple choice data. Further, TACSS recommended that the mean squared deviation be used in the analysis of discrepancies between round 3 actual ratings and expected ratings.

- This discrepancy has been referred to as an indicator of the impact of the Reckase Charts on panelists' ratings. Because many factors could contribute to the discrepancy, TACSS recommended that ACT

avoid this terminology. The word "absolute" should be added to the title for reporting these discrepancies since the absolute differences are used in the analyses.

- One TACSS member suggested the use of person-fit analyses to discover which panelists judged item difficulty most and least accurately. The suggestion was also made that the analyses must impose some limits by using controls for floor/ceiling effects. Finally, TACSS suggested that if the focus were only on two rounds of ratings, it would be reasonable to plot the data for "expected" and "actual" ratings in scatter plots.

- TACSS noted that by Round 3, the rater location feedback resembled distinct clusters of cutscores with no overlap between levels. The data were remarkably tight, showing very little spread. These data are quite different from those for the 1992 Writing ALS which might be, in part, the result of a higher quality assessment.

## TACSS Comments about Writing ALS Questionnaire Analyses

(1/99)

- It was suggested that the questionnaire data and the rater consistency data be linked. Analysis should be done of the individual raters whose ratings most nearly formed a row on the Reckase Charts and of raters whose ratings were most variable relative to a row on the Reckase Charts. The individual responses by such panelists to the evaluation questionnaires should be studied for patterns.

## Additional Analyses of ALS Data

(4/99)

- A small proportion of civics ALS panelists was classified as "extreme raters." Most of the extreme ratings occurred at the Basic level for multiple choice items. No extreme ratings were found for writing panelists. There were no noticeable differences in cutpoints after round 3 when comparing Group A with Group B.

## Follow-up Analyses for Civics ALS and Writing ALS

(2/99)

- TACSS generally agreed that the results of the follow-up analyses conducted for the Civics and Writing ALS data did not indicate any unusual patterns that would cause concern about either the ALS process or the results of the process. Overall the follow-up results indicated that the ALS procedure was implemented exceptionally well and that the data were remarkable consistent. TACSS recommended to NAGB that they adopt the achievement levels as set for both writing and civics.

(2/99)

- **Methodology used in exemplar item selection.** There was a technical discussion about how the theta metric for NAEP data varies, depending upon the existence of subscales for the separate subjects. Since the item parameters relate to the subscales, calculations become very complex for many of the analyses that must be performed using the NAEP data. It was noted that the specific calculations must be reevaluated for each new data set, given the variations in the Assessments from year to year. TACSS recommended that ACT report the details of each computational method used during data analysis for all procedures that are involved in the achievement levels setting process.

(4/99)

- Chen summarized the steps ACT followed to compute the average probability of correct response for the purposes of identifying items in the Exemplar Item Selection set. Zwick and Mazzeo explained their concerns about the metric used to compute the average p values. There was a technical discussion about the relationship between the metric of the item parameters and the metric of the reporting scale, which varies across assessment years and subjects. The current computational procedures are particularly problematic when a composite scale is involved. When a composite scale is involved, the current procedures will be very inaccurate.

## TACSS Recommendations for Including or Rejecting Validations Studies

(12/97)

### Comparison of State Writing Standards to NAEP Study

Purpose: To examine the degree of similarity between state writing standards and NAEP writing standards and to indicate the level of difficulty of NAEP ALDs relative to state ALDs.

Rationale: The study will generate valuable information that will determine the extent to which NAEP standards can be considered reasonable when compared to state standards. If the expected skills and performances for NAEP are similar to those for the state documents reviewed, then the NAEP standards will be regarded as reasonable. They will also be useful because they allow the potential for comparisons of student performance on state assessments and the writing NAEP.

(12/97)

- Will be conducted by Cathy Welch.
- NAGB currently involved in studying this issue.
- Alignment of North Carolina and Maryland state standards with NAEP standards.
- Also looking at how to develop model that can be used for repeating this kind of study for other states.
- Study not only should include intended use for standards, but also how standards ultimately were used; for instance, high stakes.

(2/98)

- Authors are seeking guidance from TACSS for further development of study.
- It was noted that some states that are included in Table 1 are missing from Table 2.
- It was unclear how states were selected for sample.
- There was some confusion regarding the terms “quantitative” vs. “qualitative” language when describing writing performance.
- It was suggested that the authors consider the following research plan:
  - First examine results of state assessment.
  - Determine percentage of students performing at B, P, A on state assessment.
  - Then examine results of State NAEP.
  - Determine percentage of students performing at B, P, A on State NAEP.
  - If state curriculum standards are the same as NAEP achievement levels descriptions, performance on state assessment should be about the same as performance on State NAEP.
  - If state curriculum standards are lower than NAEP achievement levels descriptions, performance on state assessment should be better than performance on State NAEP.

(9/98)

- ❖ Study nearly complete.

### CONGRUENCE OF ACHIEVEMENT LEVELS DESCRIPTIONS FOR CIVICS

(formerly Behavioral Anchoring)

Purpose: To examine the degree of similarity between what students *should know and be able to do* and what students *actually do.*

Rationale: The study will provide information that will determine the extent to which NAEP standards can be considered reasonable when compared to what students "can do." If the descriptions of student performance written for items that map within achievement levels cutscores are similar to ALDs, then the NAEP standards will be regarded as reasonable. This will be evidence in support of the validity of the achievement levels.

(12/97)

- ❖ Will be conducted for civics, but not writing.
- ❖ Some discussion about difficulty in applying mapping procedures to writing.
- ❖ Use another name to distinguish this study from ETS uses of term.

(9/98)

- ❖ There was general agreement to delay conducting this study.

(1/99)

- ❖ The Congruence Study (previously referred to as Behavioral Anchoring) will be rescheduled for March, when the SCS was originally scheduled. The Congruence Study involves panelists writing descriptions of student performance for items that map within achievement level cutscores. These descriptions will then be evaluated by content experts for their degree of similarity with the ALDs. ACT will provide a description of the study for TACSS review and recommendations during the February meeting.

(2/99)

- ❖ TACSS agreed that the CAS should be postponed and reconsidered. Although several concerns were expressed about the study, it would seem that the primary objection was the lack of criteria for Part II in determining the similarities and differences between the two sets of achievement level descriptions. TACSS recommended that ACT redesign the study for further consideration.

(4/99)

- ❖ Loomis reported that the Item Classification Study has been cancelled because TACSS's concerns regarding the study could not be overcome. At the March board meeting the ALC asked Loomis if some form of validation evidence would be available for their deliberations that would match civics items and achievement level descriptions. Loomis reminded them of the ETS review that matched the revised ALDs with the civics item pool. TAT, however, recommended that if the ETS review lacked reliability, then ACT should conduct a small classification study in response to the ALC request.

### SELECTING EXEMPLAR ITEMS STUDY

Purpose: To study how well different sets of exemplar items communicate the achievement levels to educators and the public.

Rationale: The study will provide information that will determine the extent to which judgmental classifications of NAEP items reflect different statistical criteria used in selecting the items. The response criterion best matching the item classifications most frequently selected will be considered the "best" criterion for selecting exemplar items. This study will help identify the criterion that will result in the selection of exemplar items that seems most reasonable.

(12/97)

- Considerable discussion of criteria used for selecting exemplar items based on discrimination and difficulty of items.

(9/98)

- Rebecca Zwick will conduct the RP criteria component of the study, and ACT will conduct the survey research component.

### EFFECTS OF MOTIVATION ON ACHIEVEMENT LEVELS STUDY

(12/97)

- Will be omitted.
- Purpose of study was unclear

### STANDARD ERRORS OF CUTSCORES

Purpose: To examine how to expand present statistical procedures used for estimating standard errors of the cutscores to include differences in panels, occasions, and instructions.

Rationale: The study will determine how the results of the ALS process could be reported with the standard error estimates incorporated into the reports. This information would be useful when reporting and interpreting the results of the ALS process to stakeholders and the general public.

(9/98)

- Is being conducted by Brad Hanson and Dave Woodruff.
- This study is underway.

(12/99)

- The purpose of this research is to examine procedures to estimate and report the variability of cutscores across a set of replicate NAEP ALS studies in order to allow such variability to be taken into account in the use and interpretation of the NAEP achievement levels. This research idea is the outgrowth of concern expressed several years ago that the Brennan method of estimating the standard error (SE) of NAEP cutscores actually underestimates SE. This study is an attempt to identify and parcel out other sources of error.

- TACSS had many questions about the details of the study. No new data will be collected for the study. One suggestion was to use the "boot strapping" technique, when data from one rater are omitted and the remaining data are reanalyzed. After considerable discussion directed at clarifying the purpose of the study, it was decided that Hanson and Woodruff should outline the positive and negative features of new and existing estimators of SE that would work with standard setting procedures. Kolstad added that rating data should be published so that it would be available for secondary analysis.

## PERFORMANCE PROFILES

(12/97)

- ❖ Will be conducted for writing only.
- ❖ Not clear whether distinction in mathematical models is real between conjunctive/compensatory judgments.
- ❖ Considerable discussion about conjunctive/compensatory judgments and varying levels of difficulty in three types of writing.
- ❖ Suggest include 20 duplicate profiles as reliability check in performance profiles study.
- ❖ Be sure to get information from panelists about decision process.

(2/98)

- ❖ Redo analysis using a more descriptive approach because it will produce more useful information.
  - Examine key combinations of uneven student performance on different prompts.
    - Key combinations represent crux of research issue.
  - Determine % of students rated Basic, Proficient, Advanced.
  - Produce distribution of ratings.
  - Look to see if judges used different strategies at different achievement levels.
- ❖ Study judges individually to gain insight into judgment process of each participant.
  - Connect ratings to strategy items from questionnaire (items 11-22).
  - Determine what model type each judge used.
- ❖ Omit terminology questions (items 27-30) because difficult to understand.
- ❖ Report residuals as well as standard error and the number of cases included in the analysis.

## SIMILARITIES CLASSIFICATION STUDY

Purpose: To investigate the validity of interpretations of the NAEP achievement levels with respect to student performance, and to explore some of the issues related to using a shortened form of NAEP.

Rationale: The study will provide information that will determine the extent to which NAEP standards can be considered reasonable when compared to how teachers classify their own students in the subject areas assessed by NAEP. If the correspondence between teachers' judgments and empirical classifications is similar, then the NAEP standards will be regarded as reasonable. Also, the ALDs will be judged as useful statements against which to judge student performance.

(12/97)

- ❖ Will be conducted at NAGB's request.
- ❖ Study includes valuable research for redesign of NAEP (market basket approach).
- ❖ Discussed use of schools/classrooms designated for advanced students in validation process for Advanced level (AP classes, etc. samples as added study for SCS).
- ❖ ACT needs to develop study further before TACSS can make recommendations

(9/98)

- ❖ This is a high priority study that NAGB requested ACT to conduct.

(1/99)

- The original design of the SCS Validation Study will be changed to incorporate a Booklet Classification component. Although TACSS voiced concern about the problems associated with teachers and students discussing the assessment before the teachers are convened, there is no logistical way to include the Booklet Classification component without this trade-off. It has yet to be determined if writing will be included in the SCS Study. TACSS agreed that it seemed unlikely that a writing form could be developed to meet the criteria required by the study.

(1/99)

- TACSS discussed adding a booklet classification component to the SCS study to examine earlier findings that booklets are classified one level lower than empirical classifications. By adding booklet classification, however, students would have to be assessed before the teachers are convened.

- Westat is heavily booked in March and will administer the SCS version of NAEP at the beginning of May. This would cause the classification study with teachers to be conducted in July (after school terms end), which is later than originally planned.

- Although the SCS easily applies to the civics assessment, it is doubtful that it will be useful for the writing assessment. Four prompts most likely will not be enough to provide statistical reliability, but too much for a student to produce in one sitting. Logistically it would be difficult to administer the assessment over two days. Further, that would not reflect the NAEP administration conditions.

- Concern was expressed about the reliability of teachers' classifications. ACT was encouraged to think of ways to account for and increase the reliability of teacher's classifications. A suggestion was also offered to have teachers classify students using more than four categories. This was done in 1995, and it was planned for the 1998 study. The method of scoring "omits and not reached" could misrepresent SCS students' scores, and ACT was urged to give careful consideration to that.

- It was generally agreed that the study should add Booklet Classification. There is concern about students and teachers discussing the assessment before teachers are convened, but this is unavoidable if the Booklet Classification task is included. Because NAGB will reach a decision on the achievement levels at their May meeting, the study will not be completed in time to inform the board before they set the levels.

(2/99)

- Selection of Item Blocks for SCS**.** TACSS generally agreed with the methodology that ACT used to select item blocks and estimate reliability for the SCS study. TACSS recommended that the test and item characteristic curves for the selected block also be examined.

(6/99)

- In general, the SCS/BCS research design will follow that used in the 1995 validation study for geography and U.S. history. The study is planned for 8$^{th}$ grade civics only. All grade 8 teacher panelists who participated in the civics pilot study and ALS meeting have been invited to participate in the SCS/BCS research. Panelists will be asked to estimate the achievement level category for each of their students in general, and on the Civics NAEP in particular. Teachers also will classify the level of achievement represented in student booklets. TACSS made many recommendations for the design of the study. See the following "Discussion Summaries" for details.

(9/99)

- ❖ General findings from the teachers' classifications of their students indicate that teachers are likely to classify their students' level of overall civics achievement and expected performance on the special NAEP at a higher level than the student's empirical performance. Teachers tend to classify student booklets at the same level, or one level lower than the empirical level.

- ❖ These findings replicate results from similar, although not identical, studies. In 1995, the findings for the SCS and BCS in geography and U.S. history were similar to the current findings. In addition, the findings for the 1998 BCS in science were similar to those for the current BCS.

### COMPARISON OF 1992 AND 1998 NAEP WRITING ACHIEVEMENT LEVELS

Purpose: To evaluate the results of the two ALS processes, the magnitude of differences in the results, and the factors that account for those differences.

Rationale: Every reasonable aspect of the ALS processes will be examined as a potential source of differences in the results. The study will produce information that will determine the reasonableness, usefulness, and validity of the achievement levels set for the 1998 ALS as compared to the results of the 1992 Writing ALS. (Note: The results of the 1992 ALS process were judged as neither reasonable nor reliable.)

(12/97)

- ❖ Necessary as part of report on 1998 ALS process.

(9/98)

- ❖ This is a high priority study that ACT must conduct.

(9/99)

- ❖ Comparing the generalizability of the 1992 Writing NAEP with the 1998 Writing NAEP reveals improved measurement precision and increased universe-score correlation among the three types of writing assessed for the 1998 Writing NAEP. TACSS agreed that the reliability estimates were adequate to produce stable cutscores to use for setting the achievement levels for the 1998 Writing NAEP. Further, the findings support the decision to use unidimensional scaling.

### PLAUSIBLE VALUES STUDY OF DOMAIN COHERENCE

(Formerly Psychometric Domain Coherence)

Purpose: To study the degree of similarity between classifications of student performances on different subscales and relative to overall cutpoints.

Rationale: This study will examine construct validity by using plausible value scores of students who took the 1994 NAEP in U.S. History and Geography, and the 1996 NAEP in Science. This issue has lost most of its urgency since the operations contractor decided that scaling for the 1998 NAEP will be performed on one scale. The information produced by the study will be useful to inform the overall ALS process regarding panelist perception of domain coherence.

(9/98)

- ❖ There was general agreement that this is a low priority study since the issue has lost most of its urgency.

### PERSON-FIT STATISTICS STUDY

Purpose: To examine the assumptions that panelists' ratings fit the IRT model of student performance on NAEP.

Rationale: The study will provide information that will determine the extent to which the assumption can be considered correct that the same model governs panelists' ratings that governs student responses. If the two response patterns are similar, then the NAEP standards will be regarded as useful and reasonable because judges' estimates of student performance accurately reflect actual student performance.

(12/97)
Will be conducted for AERA presentation.

- ❖ General issue: Is it appropriate to view judges ratings as fitting model used for examinee data?

(3/98)
The primary purpose of the study was to determine if panelists' ratings fit the IRT model of student performance on NAEP. Results suggested that panelists did not fit the model well. It was generally agreed, however, that the data needed to be reanalyzed using three separate ANOVA's, one for each RP criterion. In this way the main effects will not be confounded with the different RP values.

### CIVICS ITEM CLASSIFICATION STUDY (CICS)

(6/99)

- ❖ Results of the study should be reported cautiously, as only 3 teachers participated in each grade group. In general, teachers' classifications of the civics items were similar in content areas and cognitive abilities to those specified in the Civics Framework. Teachers generally agreed on the classification of items by achievement levels. Using .65 as the response probability value produced classification outcomes that were similar to those produced using .50 as the RP value. There was low agreement between the outcomes of classifications made separately by ACT and ETS. See the following "Updates" for details.

## RESEARCH STUDIES RELATED TO THE REDESIGN OF NAEP AND TO INTERNATIONAL BENCHMARKING

(2/98)
General Discussion

- ❖ There is a need to formulate process of awarding funds.
  - TAT has recommended funding 1 or 2 studies now, with remaining funds available later.
- ❖ ACT needs TACSS involvement in conducting some research studies that have been proposed to NAGB. Reports on research studies due July 2000. The following are topics suggested for TACSS to consider:
  - Overall summary of what has been learned about standard setting for NAEP
  - Computation of standard error in ALS process
    - It was suggested to coordinate research on standard error with recommendation from ACES to research sources of error.
- ❖ There is renewed research interest in TIMSS data. Some topics of interest include:
  - Compare science NAEP results with science TIMSS results

- Recompute science NAEP data using TIMSS procedure
- Particular interest in performance of grade 8 on TIMSS when it will be administered in 1999 because grade 8 took TIMSS as 4th graders in 1995.

### ALS PROCESS USING TIMSS ITEMS

Purpose: To set NAEP achievement levels on TIMSS to relate the results of the 1996 NAEP Science achievement levels and TIMSS.

Rationale: The study will generate valuable information that will determine the extent to which NAEP standards can be considered reasonable with respect to performance of students in international assessments. The set of cutscores resulting from the ALS process rating TIMSS items will provide the percentage of students in the U.S. and in other countries performing at or above each achievement level. The NAEP standards will be regarded as reasonable if performance on TIMSS results in approximately the same proportion of U.S. students performing at or above each achievement level, as was the case for NAEP. We expect the cutscores set on TIMSS will be relatively higher than those set on NAEP.

Now that TIMSS R99 is scheduled for administration, this study seems increasingly relevant and important. We recommend that this study be conducted for grade 8, at a minimum.

(9/98)

❖ This is a study that some recommended as a low priority. NAGB staff recommended considering conducting the study for math, but not science. Several problems were anticipated in conducting the study described. TACSS will reconsider the study with a different subject (math).

### TIMSS-NAEP LINKING FEASIBILITY STUDY

(12/99)

❖ Nichols explained the rationale of the study, "Evaluating the Use of TIMSS as an International Benchmark for NAEP." The study relates the NAEP achievement levels to TIMSS using the mathematics NAEP achievement levels and the TIMSS mathematics framework and items. In this study, content experts will compare the two assessments at different levels of specificity. The purpose is:

- To compare the NAEP framework and the TIMSS framework for math;
- To compare the NAEP achievement levels descriptions for math with the TIMSS math framework;
- To compare the NAEP achievement levels descriptions for math with the TIMSS math item pool.

❖ Two or three panels comprised of 6-12 content experts in mathematics would be recruited from different regions of the country. The 1999 TIMSS items would be used for the comparisons. It was suggested that panels consist of experts who not only would be familiar with students, but assessment frameworks as well. Experienced item-writers would also be desirable panel members.

❖ Shakrani explained the AIR study that attempted to link the science and math NAEP with TIMSS for grades 4 and 8. The general conclusions from the study were that TIMSS math and NAEP math measured similar skills at grade 8. The math assessments seemed to be written at about the same level of difficulty for both grades. However, the TIMSS science and NAEP science measured different

skills. The science NAEP seemed to be more difficult than the TIMSS. Linking the two assessments seemed possible for grade 8, but not grade 4.

- ❖ Loomis remarked that she would like to read the AIR study before making a decision about the future of this research. Reckase questioned how to set the criteria that would determine the degree to which the two assessments are similar. If the comparison of the assessments were conducted in detail, differences inevitably would emerge. Perhaps NAGB should review the comparisons and decide if the differences are small enough to allow linkage, or large enough to prohibit linkage. Bourque added the general "rule of thumb," that there must be 85% commonality in content and level of difficulty for linkage to be appropriate for the two assessments.

### RESCALE NAEP USING TIMSS MODEL

Purpose: To examine the relationship between the results of the 1996 NAEP Science achievement levels and TIMSS.

Rationale: The study will generate valuable information that will determine the extent to which NAEP standards can be considered reasonable by using the IRT model that was utilized to scale items from TIMSS. (We have learned of plans for rescaling TIMSS with the NAEP IRT model.) The set of cutscores resulting from the ALS process using the new set of parameters ratings will provide the percentage of students in the U.S. and in other countries performing at or above each achievement level. This study allows comparisons of U.S. student performance on TIMSS and NAEP, and comparisons of performance of students from the U.S. with students from other countries. As such, it provides a useful benchmark.

(9/98)

- ❖ This is a low priority study and perhaps should be omitted.

### INTERNATIONAL ALS FOR WRITING

Purpose: To replicate the ALS process in writing in an English-speaking or largely English-speaking country, and to determine if NAEP standards reflect "world class standards."

Rationale: The study will provide information that will determine the extent to which NAEP standards can be considered reasonable when compared to expectations of student performance in other countries. If cutscores set by international panelists are similar to those set by U.S. panelists, this will be an indication that NAEP standards and U.S. student performances meet "world class standards." Higher cutscores set by international panelists would indicate relatively lower performance by U.S. students, and lower cutscores would indicate relatively higher performance by U.S. students.

(9/98)

- ❖ This is a low priority study. Omitted.

### USE OF NAEP SHORT-FORMS AND USE OF DOMAIN SCORES

Purpose: To examine how to use the short forms of NAEP in the ALS process, and to determine whether scores on the short forms of NAEP could be used to improve reporting and public understanding of NAEP ALS results.

Rationale: The study will provide information that will determine the extent to which standards set using the long-form NAEP are similar when compared to those set using the short-form NAEP. If the correspondence between the two sets of standards is similar, then the implication is that performance levels set on the short-form NAEP can adequately represent NAEP achievement levels. This information would be useful in streamlining the ALS process and reporting procedures.

(9/98)

- The future of these studies has not been decided.

### USING DOMAIN SCORES IN REPORTING ACHIEVEMENT LEVELS: MATT SCHULZ

Purpose: To examine how to use domain score estimation procedures for reporting NAEP performance levels related to "market basket" reporting.

Rationale: The study will provide information that will determine how to relate NAGB standards to domain scores as a means for using a percent correct metric for reporting NAEP performance levels. This information would be useful in improving reporting procedures.

(12/99)

- Nichols presented a brief overview of the study, which addressed two challenges common to level-based assessments. The first challenge is to establish a clear definition of the meaning of level scores. The second is to estimate the technical characteristics of level scores.

- Nichols reported that the method described in the study could be used as an alternative standard-setting method for assessments other than NAEP. It could also be used as a means of gaining additional information about the NAEP achievement levels after cutscores have been set. A major drawback of the method is that the quality of the item pool is confounded with the quality of judgments. The procedures described by Schulz would apply more aptly to the bookmark method than to the ACT/NAGB method for setting achievement levels because it uses domains rather than single items. Further, items must be classified according to achievement levels implying that "Basic items" exist, which is in conflict with NAEP principles.

- Mazzeo expressed interest in applying the procedures described by Schulz to the market basket NAEP forms developed by ETS. It would be helpful to have an item classification study conducted to examine the two market basket parallel forms, which are the same length as the NAEP forms. This would be a good way of comparing the level of representation of the items selected for the market baskets with the complete item pool. ETS is interested in reporting market basket results as clusters of achievement at different levels.

- Loomis responded that ACT had planned to compare the cutscores of the market basket items with the cutscores of the complete item pool used during the ALS meeting. Mazzeo remarked that he would find out more details about the items used in the market basket forms and the item rating data for those items.

### EXAMINING STANDARD SETTING

Purpose: To guide NAGB in future standard setting for the redesign of NAEP.

Rationale: This study will review previous research on standard setting methods and develop a conceptual framework that can be used to describe the effects of variations in NAGB standard setting procedures. It will describe how different procedures affect the final standards. Also, it will explain the effects of new variations in standard setting procedures that would have been implemented in research studies proposed for the contract. The information will be useful not only to NAGB in future standard setting efforts, but to all persons and organizations interested in the standard setting process.

(9/98)

- ❖ This study coincides with research interests of TACSS member Mark Reckase, who will conduct the study if NAGB approves the plan.
- ❖ If the Voluntary National Test (VNT) goes forward, linking performance on the VNT to performance on NAEP would be a very useful study. This could occur after the VNT field-testing is completed in the year 2000.
- ❖ Developing achievement levels for TIMSS based on its framework would also be a useful study. It was suggested that TACSS offer fresh thinking regarding these research ideas and readdress this issue at a later date.

(2/99)

- ❖ Reckase reported that he will be writing a comprehensive report about previous research conducted for the standard setting projects for NAEP.

(12/99)

**THE EVOLUTION OF THE ACHIEVEMENT LEVELS-SETTING PROCESS**

A Summary of the Research and Development Efforts Conducted by ACT: Mark Reckase

- ❖ The purposes of this report are to summarize the results of the many standard setting studies that have been conducted under contract to NAGB, and to describe how the results of the studies have lead to changes in the design of the ALS process. TACSS' overall reaction to Reckase's paper was quite positive. It was noted that the study should be very useful when orienting new members of TACSS and NAGB to the history of achievement levels. Kolstad suggested that "holistic" feedback should be de-emphasized and the Reckase Charts should receive more emphasis. Rindone added that the findings from the Booklet Classification Study should be emphasized, particularly that higher cutscores result from implementing the booklet classification method. Zwick commented that the use of examples and a glossary of technical terms would be informative to the reader. Reckase asked that additional suggestions for revisions be sent to him by mid-December.

## DISCUSSION OF RESPONSE PROBABILITY VALUES FOR ALS PROCESS AND NAEP REPORTING

(2/98)

General Discussion

- ❖ Research proposed by R. Zwick related to RP values, but study will not begin until summer, 1998.
- ❖ Group discussed advantages and disadvantages of using a single RP value vs. multiple values.
  - Is there a need for a uniform RP value?
  - NAEP reports have used different values in past for different purposes.
  - Different RP values could lead to confusion when linking information from different sources.
  - RP values directly impact procedures used in bookmarking, item maps, exemplar items.

- ACT has some data on RP values from analysis of NAEP science data.
- ACT has pressing need to proceed with process of deciding RP values.
- Andy Kolstad from NCES is researching topic of RP values.
  - Perhaps he could present information to TACSS on key points to consider for RP values.
  - In addition, he could assist in coordinating any ALS research which might result from ACES recommendations.
  - Perhaps NAGB can examine issue in May.

(3/98)
Andy Kolstad from NCES presented his work on the issues related to the criterion for selecting RP values. The committee members discussed the various uses of RP values during the ALS process. A lower value than 65% corresponds better with what people can do. Higher values give a false sense of individuals not getting as many items right as they, in fact, answer correctly. The primary issue is to identify an RP value (or RP values) to use in reporting NAEP results (or other reports from NCES) that are readily understood and correct interpretations of what students "can do." There was general agreement to continue using the current method for selecting exemplar items with an average RP value across the range of 50%, and to adopt the values used by ETS for other purposes.

(1/99)
- Zwick reported that her Item Mapping Study is underway. She selected data from the 1990 Science NAEP to use for the study because the data met important criteria requirements. Only items in the physical science subscale will be used in her analysis. Achievement levels were not set for the1990 Science NAEP. TACSS discussed whether or not the study, as designed, would be informative to the ALS process since the findings cannot be related to achievement levels. TACSS recommended that the study continue as planned. ACT will conduct the survey element of the Item Mapping Study.

(1/99)
- Zwick explained the design for the study, "An Investigation of Alternative Methods for Scale Anchoring and Item Mapping in the NAEP." The purpose of the study is to determine the best technical approach to item mapping given anchor points or achievement levels. The study also includes a survey by a panel of experts to evaluate the ordering of items by their level of difficulty. Zwick requested from ACT a description of the specific procedures used for selecting exemplar items.
- Zwick chose to use the 1990 NAEP physical science data because they met the criteria developed for the study.
- There was a sufficient number of released items in this single subscale
- The parameter estimates were available
- The data were ready to use. However, there are no achievement levels set on the subscale for physical science in the 1990 Science NAEP. Although she would have preferred to use a scale that already had achievement levels set, those data sets had other aspects that would compromise the study design.
- Ideally the study should use a set of items that exemplify more than simply a rank order of item difficulty. It is essential to have a basis of comparison as part of the process of evaluating the items. The achievement level descriptions (ALDs) would provide this comparison. Without the ALDs, the study is of limited value in what can be learned about the achievement levels-setting process.
- Several points of discussion were heard. Replication is important to support statistical stability. It is a good idea to structure the task for the expert panel, using 3 or 4 clusters of items defined by statistical procedures. Zwick is considering using obvious clusters as well as "pseudo clusters."
- After considerable discussion, TACSS recommended that Zwick continue with the study as it is designed using the dataset for which there are no achievement levels.

(4/99)

- ❖ Zwick reported that she is nearly finished analyzing the NAEP data. The next step in the study is developing the survey, which is scheduled to be mailed to panelists during the summer.

(6/99)

- ❖ TACSS reviewed the survey that is being developed for Rebecca Zwick's study examining exemplar items. Many revisions to the survey were suggested, as summarized in the following "Updates" section.

(12/99)

**An Investigation of Alternative Methods for Scale Anchoring and Item Mapping: Rebecca Zwick**

- ❖ Zwick reviewed the research questions for the study, which had been presented in greater detail at earlier TACSS meetings:
  - Do study results agree with expert judgments?
  - Are results stable over samples?
  - Does the method produce an adequate number of exemplar items?
- ❖ Because data collection has just been completed, the data have not yet been analyzed completely. Preliminary findings indicate, however, that at least 3 items mapped at each level for all methods examined. The major drawback of the study is that no achievement levels were set on the 1990 Grade 8 Physical Science NAEP data, which is the data that was used for the study. To pick the cutpoints, Zwick roughly equated the 1990 Science NAEP data to the 1996 Science NAEP composite data. She put the 1990 cutpoints 2 standard deviations from the mean of the 1996 data.

## TACSS RESEARCH PROPOSALS

(12/97)

### STANDARDS-BASED NAEP SCORE REPORTING: RON HAMBLETON

- ❖ TACSS' suggestions to Hambleton:
  - consult with other research organizations that are studying reporting issues
  - seek information from industries that use similar reporting methods
  - search for existing computer software that produces high quality reporting formats

(12/99)

- ❖ Hambleton reported that the purpose of this study is:
  - To identify the strengths and possible weaknesses in current NAEP score reporting;
  - To design and field test new standards-based score reporting formats;
  - To incorporate the use of international benchmarks and "market baskets" into NAEP.
- ❖ A two-hour focus group consisting of ten persons reviewed the 1996 NAEP science reports. Preliminary findings indicated that people who read the fuller report answered questions about NAEP results more accurately than people who read the highlight report. Readers did not understand the linkage between the achievement levels, sample items, and scoring guides. Readers suggested that data be included that would enable comparisons of student responses that represented the different achievement levels to the same assessment exercise. They disliked having exceptional data (marked with an asterisk) included in the summary tables The final report is due February 2000.

## DISCUSSION OF ACHIEVEMENT LEVELS ISSUES THAT HAVE BEEN RAISED RECENTLY

### SETTING ADDITIONAL ACHIEVEMENT LEVELS BELOW THE PROFICIENT LEVEL

Loomis reported that Education Secretary Riley recommended that NAGB set an additional achievement level to provide more information about performance below the Proficient level. Reckase stated that at the lower range of performance, guessing becomes a major factor when estimating cutscores, causing the cutscores to become unstable. Shakrani stated that Secretary Riley was interested in providing more information about the levels NAGB has defined, not creating another achievement level. The request was for richer text about what students can and can not do, and how parents and teachers can help students improve their performance. Shakrani remarked that more information has been provided for 2% of the students performing at the Advanced level than 40% of students performing at the Below Basic level. Shakrani went on to say that recently Secretary Riley has been referring to NAEP achievement levels using the term "Proficient and above" rather than Proficient and Advanced.

Bourque suggested using items that map at the Below Basic level as a means of identifying examples of what students performing at that level can do. Because the ALDs describe performance across the interval and not at the cutpoint, information could be provided about student performance across the interval. For example, what can students do who are performing at the high end of Basic that students can't do who are performing at the low end of Basic? What are students unable to do who are performing at the high end of Basic that students can do who are performing at the low end of Proficient? The market basket study could address these issues by looking at how many more items a student would need to get right to move up a level.

Although these are interesting questions, they reflect a very naive way of looking at NAEP. Given the limitations of NAEP data, it is unreasonable to make instructional recommendations based on NAEP results. In fact, it would be difficult to give more information that is technically accurate about NAEP data than what has already been given.

### REPORTING ACHIEVEMENT LEVELS WHEN MULTI-DIMENTIONAL SCALING IS USED

Loomis described the problem that occurred when preparing the Reckase Charts using math and reading NAEP data. The charts graphically displayed different cutscores for different subscales. Exemplar items have been selected based on the composite scale score, which has been shown to be highly variable when considering subscales. Perhaps the exemplar items should be selected based on the subscale score rather than the composite score. This has not been an issue for the 1998 NAEP because both civics and writing have used uni-dimensional scaling. It might be an issue, however, when reporting exemplar items for off years when no ALS meeting is held. Further, it will become an issue if the Reckase Charts are used in future ALS meetings when the data are not scaled uni-dimensionally.

Reckase commented that panelists should see the scaling variations on the charts, since this represents what actually happens to the data. Mazzeo remarked that since NAEP is based on a compensatory model, the composite score would seem to be appropriate to use when selecting exemplar items. This issue must be resolved because ETS will use subscales for the next two assessments.

### REPORTING ACHIEVEMENT LEVELS DATA IN PERCENT CORRECT METRIC

NAGB proposed reporting NAEP outcomes as the percentage of total possible points required for performance at each achievement level. Focus groups are underway to determine how this will effect understanding the results. It seems that the public likes being able to relate the score scale to a frame of reference they know. However, reporting the percent correct for NAEP can be very misleading. The

percentages of total possible points required for performance at each achievement level usually are very low, which is inconsistent with the message that Basic is a reasonably demanding level of performance. When panelists use whole booklet data, they tend to make performance at the Advanced level higher than 80%.

### CHANGES IN NAGB POLICY DEFINITIONS FOR ACHIEVEMENT LEVELS

NAGB is questioning how well the achievement levels are understood, and how they can make the policy definitions clearer. Changing the actual terms of the achievement levels is not likely, but improving the policy definitions is possible. Bourque noted that if changes were to be made in the policy definitions, they would apply to new standards, not what has already been done. Loomis mentioned that ACT could ask panelists from the validation studies to assign grades for each achievement level to students who participated in the study. In this way we could determine if Basic performance would be represented by a letter grade of C or D, Proficient performance a B+, and so forth.