Briefing Booklet

1998 Civics NAEP
Achievement Levels-Setting
Study

Preface

We developed this booklet to give panelists information regarding the *what*, *why*, *when*, and *how* of the different parts of the Achievement Levels-Setting (ALS) Process. The purpose of providing this booklet is to give you a quick reference source. Please share your suggestions for additions, deletions, and clarifications.

The first part of this booklet is organized chronologically, according to the sequence of activities in the ALS process. Some parts of the process will be repeated during the five days, like receiving particular types of feedback at different stages. These steps are described in detail only once, but they are referenced briefly whenever they occur in the sequence of activities.

The second part of this booklet is a glossary of ALS terms that could be unfamiliar, or that will be used in a way that is unique to the ALS process. These terms are listed in alphabetical order for easy reference.

The final section of this booklet is a brief description of the role that chance plays in student performance. This material will be particularly helpful for estimating student performance on the NAEP.

Please become familiar with all of the information contained within the *Briefing Booklet*. We think you will find it quite helpful in preparing for the task of setting achievement levels for NAEP.

NAEP ALS Project Staff
ACT, Inc.

ALS Activities

Orientation The orientation will be in two parts: the whole group orientation and the grade group orientation. The whole group orientation sessions will provide panelists a common understanding of the purpose of setting achievement levels and the procedures to be followed in setting the levels. There will be time for you to ask questions and raise concerns about information presented by the speakers, and information from the advance material.

During the grade group orientation session, panelists will be informed about their rating groups, ID codes (secret and public), and specific details and concerns related to the ALS process. This is also a time for you to get acquainted with other panelists.

NAEP Exam You will take a form of the 1998 Civics NAEP exam for your grade level. By taking the exam, you will gain experience with an actual test booklet and begin the process of becoming familiar with test items and scoring guides. You will also relive the joys of taking an exam under time constraints. You will review your own test performance using the scoring guides, but your score will not be computed.

Framework Presentation Members of the framework consensus panel (grade level content staff) will present the 1998 NAEP Civics Framework. They will present the achievement levels descriptions (ALDs), along with a detailed description of the conceptualization and philosophical foundations of the assessment framework. Although we expect you will have read the framework document prior to the meeting, we have scheduled this session to help assure that *everyone* has a full understanding of the assessment framework and a clear and common understanding of what students should know and be able to do at each grade level. This is the first step toward reaching a common understanding of the ALDs.

ALDs Grade Group Discussion After you have heard the Framework Presentation in the whole group session, the grade level content

staff will review the ALDs for the grade group. He/she will also help your grade group internalize the descriptions in relation to the framework. You will participate in intense discussions of the ALDs. This session is designed to help you become comfortable and conversant with the ALDs.

ALDs Exercises Two exercises have been planned to help you continue to gain a common understanding of the ALDs. In the first exercise, you will use your understanding of the ALDs to estimate student performance for a block of items. That is, you will determine the level of performance required to respond to each item correctly, based on your understanding of the ALDs. It is important for you to realize that the test items themselves are not Basic, Proficient, or Advanced. Rather, it is expected student performance that is classified as Basic, Proficient, or Advanced.

In the second exercise, you will apply your understanding of the ALDs more holistically, i.e., to a composite or whole set of performances by a student. A sample of ten student booklets will be given to you to review and discuss with respect to your understanding of the ALDs. You will be asked to determine whether the *performance* exhibited in each test booklet (including 30-40 items of different types and difficulty) should be classified as Basic, Proficient, or Advanced.

Performance at the Borderline Performance at the Borderline In order to focus estimates of student performance as precisely as possible, a point is conceived as the cutpoint. The point could be the mid-point of the level, the average, or either the upper boundary or the lower boundary. We use the lower boundary.

Since the main focus in setting achievement levels is on performance at the lower boundary of each achievement level, it is important that you be familiar with and clear about the concept of *borderline performance*. The concept of borderline performance will be introduced to you in a whole group session. An example of rating at the borderline will be presented.

In your grade group, you will write descriptions of borderline student performance for each achievement level.

Your work on developing descriptors of borderline performance will help you continue to develop a common understanding of the ALDs.

You will review borderline descriptors before each round of ratings. Modifications may be made at those times. The sooner you reach agreement on the borderline descriptions, the easier it will be for you to make firm your concept of borderline performance. Borderline descriptors *must* be finalized *before* Round 3 ratings.

Paper Selection Exercise For this exercise, you will be given packets of student papers written in response to constructed response (open ended) items along with scoring rubrics for each item in two blocks of items in your rating pool. These two blocks are included in the rating pool for each panelist in your grade level. We will ask you to select one paper to represent borderline performance at each level for each item. You will use a yellow "flag" to identify your borderline Basic selection; a blue "flag" to identify your borderline Proficient selection; and, a red "flag" to identify your borderline Advanced selection. The scores of the papers will be given to you so that you may know how the papers you selected (and rejected) were scored. Just as a reminder: You are selecting papers to represent borderline performance, regardless of their score. The score, *per se*, does not matter.

This exercise has several purposes. In general it will help you continue developing a common understanding of the ALDs and reinforcing your concept of borderline performance. Specifically, the paper selection exercise will help you: (1) to arrive at a common understanding of borderline performance at each achievement level; (2) to have a "reality check" of student performance relative to the ALDs and your borderline descriptions; (3) to become familiar with the scoring rubrics and how they are used.

Training in Rating Procedures After your grade group arrives at a common understanding of the achievement levels and their respective borderlines; you will next be trained as raters.

Instructions for using the two rating methods will be provided in a whole group session. You will be instructed in the mechanics of rating items of different types. Some aspects of item difficulty and student test-taking behaviors will be described so that you can take these into account when you rate items. It is essential that you understand what the rating task entails and the differences in the methods used to rate different types of items.

The Modified Angoff method will be used to rate dichotomous (multiple choice) items, and the Mean Estimation method will be used to rate polytomous (constructed response) items. At the conclusion of your training as a rater, you must understand the rating methods, how to apply both rating methods, and how to mark the rating forms in order to provide ratings that accurately reflect your judgment of how students at the borderline of each achievement level will perform.

Round 1 Ratings There are three rounds of item ratings, and each rating session will be conducted in grade groups. In each round, you will rate every item in your item rating pool. Dichotomous items will be rated using the Modified Angoff method, and polytomous items will be rated using the Mean Estimation method. During the first round of ratings, you will be asked to think about how you would answer each item, check the scoring guides to determine the correct answer, and then decide on the rating to give each item for each achievement level. You do not necessarily need to write an answer, but you should think carefully about what is required for a student to answer the item correctly. You should always refer back to the agreed-upon performance described for each achievement level and the particular criteria you have identified for borderline performance for each achievement level. You should not discuss your item ratings with other panelists during the first round. Your item ratings will be combined with others in your grade group to

produce the cutpoints for each achievement level. These ratings will also be used to produce the feedback data to inform your second round of ratings.

Feedback Data After the ratings for Round 1 have been entered into a computer data file and analyzed, reports will be prepared for your review. These reports are to inform you about your ratings and to help you in the next round of ratings. The specific types of feedback that will be provided before Round 2 ratings are described below. Aside from the p-values, *all feedback information for Round 2 is based on the ratings that you provided during the first round.*

Cutpoints and Standard Deviations: These numbers let you know where your grade group has set the cutpoint for each achievement level and how the cutpoints compare across achievement levels. The *cutpoints* are the combined ratings over all raters and all items for each achievement level for each grade. Cutpoints are based on the mean (average) of the ratings provided by each panelist in the grade groups. The cutpoints are presented on the ACT NAEP-like scale and take into account statistical information about item difficulty, item discrimination, and chance probabilities.

The *standard deviation* is the indicator of how different the cutpoint for each individual rater is with respect to the overall grade-group cutpoint.

Whole Booklet Exercise and Feedback: Both the whole booklet feedback and exercise offer information to you for the test booklet form that was used when you took the NAEP exam. The whole booklet exercise is another "reality check" in that it provides information on how students performed relative to the achievement levels you set in the first round of ratings. It also provides holistic information about student performance. You will see that students, whose NAEP score is the same as the cutpoint you set, answered some items correctly and

some incorrectly. Not all students with the same score answer the same items correctly!

The whole booklet feedback is illustrated by a pie chart accompanied by a written description. The expected score is diagramed for the set of items in the exam booklet for students whose NAEP score is the same as the cutpoint you set to represent the borderline of each achievement level. For example, a whole booklet feedback report might state: "Based on the average of your group's ratings, students performing at the borderline Basic level are expected to get 49% of the total possible score points for this booklet." Such statements will be provided for each achievement level. This feedback is based on the cutscore you and the other panelists in your grade group set during the previous round of ratings.

The whole booklet exercise is an extension of this feedback. You will be shown copies of booklets with scores around 49% of the total possible score points, for example. You will be asked to examine student responses and determine if they are what you expected from students performing at the borderline of each achievement level. If there is a discrepancy between the performance that you expected and the performance in the booklets, then you should discuss the achievement level descriptions and borderline performance again with other panelists and try to understand the cause for this discrepancy.

Rater Location Data: Based on the item ratings that you provided during the first round of ratings, you will be informed about the location of your ratings on the ACT NAEP-like scale relative to the ratings of the other panelists in your grade group. The rater location data are illustrated graphically, using your secret ID code. Your letter code will be placed on the ACT NAEP-like scale to indicate where the cutpoint would be set had

you been the only panelist setting the achievement levels. Since only you know your secret code, you will see where your cutpoint is on the scale, but no one else will know.

The rater location data offers you a frame of reference when considering your ratings. If your cutpoints are very different from those of others in your grade, you must determine whether your understanding of borderline performance agrees with theirs.

If you think your cutpoint for an achievement level is relatively low and if you want to make it higher, then in the next round(s) you will want to raise your estimates of the percentage of students who would get the correct answer, or raise the mean score you estimate for students. If your cutpoint seems relatively high, then you may lower your estimates in the second round.

The rater location data are provided for your consideration. Changes in your ratings should be made only if you want to raise or lower your cutpoint based on this and other feedback. The decision to change item ratings is that of the individual panelist. You do not need to change your ratings at all if they reflect your understanding of the ALDs and how you have agreed they should be applied—regardless of where your cutpoint is set.

P-Value Data: The p-value feedback is a list reporting overall student performance on each item. The proportion of students who gave the correct answer is the actual "p-value" for each dichotomous item. The mean (average) score is given for each polytomous item. The p-values are listed for items within each block. This information gives you a "reality check" because it shows how students actually performed on each item. The p-values and means for the items indicate how easy or difficult the items are. A higher p-value or mean is associated with easier items

because more students got the item correct or got a high score. A lower p-value or mean is associated with more difficult items because more students got the item wrong or got a low score.

Because there is only one percentage or mean score reported for each item, there is no direct comparison between the p-values or means and the ratings that you gave each item. You must keep in mind that the p-value feedback is based on the performance of all students who answered the items, while your ratings are estimates of how students at the borderline of each of the three achievement levels will perform on the item.

Reckase Chart: For each block of items in your rating pool, you will be given a chart that indicates performance for students scoring at each score point on the ACT NAEP-Like scale. This chart is called the Reckase Chart. For each ACT NAEP-Like score point listed, the chart gives the probability of correct response on each multiple choice (MC) item and the average score on each constructed response (CR) item. Thus, for each cutpoint that you set, you will know the expected probability of correct response for each MC item and the expected (mean) score on each CR item.

You will be instructed in how your Round 1 ratings were marked on the Reckase Chart. You will mark your rater location (cutscore) and the grade level cutscore for each achievement level. You will draw a line to represent your ratings at each level for items in your rating pool. You will then examine the location of your ratings and look for patterns. For each achievement level, you can compare *your* ratings for each item to student performance. The locations and patterns of your round one ratings on the Reckase Chart will inform your ratings for Round 2.

Review Borderline Descriptors Between each rating round, you will have a brief retraining period. Part of this retraining will be a revisit of the ALDs and borderline descriptors. Your concept of borderline performance might change after looking at all the items, actually setting cutpoints based on your understanding of the ALDs and borderline descriptors, and examining responses of students performing at the cutpoints. You may want to discuss the ALDs and borderline descriptors with your grade group and make adjustments to the borderline descriptors. Adjustments may only be made if you all generally agree that they are needed. All adjustments must be made prior to the final round of ratings.

Round 2 Ratings This session is very similar to Round 1. You will still be rating each dichotomous item using the Modified Angoff method and each polytomous item using the Mean Estimation method. You will probably be quite familiar with the items in your item rating pool by this time. Based on the feedback and student performance data provided prior to this round, you may wish to change some, all, or none of your ratings from Round 1. The judgment is yours to make, given the information you have available to you for Round 2 ratings. If you determine that no changes are necessary, then no new ratings need to be recorded on your rating form. You will mark a dash "—" to fill the space for unchanged ratings for Round 2. If you decide to change a rating, you will need to write the new rating in the space provided on the rating form for the Round 2 ratings. Your second round ratings should be based on your understanding of the ALDs and borderline descriptions, the information provided after Round 1, and the locations of your ratings on the Reckase Charts.

Feedback After the ratings for Round 2, feedback reports will be updated. In addition to the feedback you received after Round 1, you will also see consequences prior to Round 3.

These data are all based on your ratings from Round 2. The purpose of this feedback is to inform your ratings for the third and final round.

Cutpoints and standard deviations: (See above.) The cutpoints and standard deviations will be based on Round 2 ratings.

Whole Booklet: (See Above.) The whole booklet feedback will be based on Round 2 ratings.

Rater Location: (See above.) The rater location graphs will be based on Round 2 ratings. They will also include some information regarding the distribution of student performance relative to specific score points on the ACT NAEP-Like score scale.

Reckase Charts: (See above.) The Reckase Charts will have your Round 2 ratings marked on them for your review.

Consequences Data: You will be given information about how student performance is distributed with respect to the cutpoints your grade group has set. The percentage of students performing on the Civics NAEP exam at or above the cutpoints set for each achievement level will be reported for your consideration and evaluation.

You must consider whether these results seem reasonable to you, in light of the achievement levels descriptions (what students *should know and be able to do*) and in light of what you know about student performance in civics (*what students know and can do*). Having seen these data, do you want to adjust your ratings?

<u>Review ALDs and Borderline Descriptions</u> You will be given a last chance to discuss the ALDs and borderline descriptors with others in your grade group. You may make modifications to the Borderline Descriptors, if you find that to be advisable.

Round 3 (Final Round) Ratings This is the last round of rating items. The only difference between the second and third rounds of ratings is that you may discuss your ratings for particularly troublesome items with other panelists in your group during the third round. Note: discussion is not mandatory. If you do not wish to participate in discussion, you do not have to do so.

Feedback After the ratings for Round 3, reports will be prepared once again for your review. The feedback data will be in the same format as that given previously, but you will not have a new set of Reckase Charts. Round 3 ratings will be used to update data for all types of feedback provided throughout the process.

Cutpoints and Standard Deviations: (See above.) The cutpoints and standard deviations will be based on Round 3 ratings.

Rater Location: (See above.) The rater location graphs will be based on Round 3 ratings.

Consequences Data: You will again be given information about how student performance is distributed with respect to the cutpoints you have set. This time, the data will be based on *your* individual cutscores from Round 3. The percentage of students performing on the Civics NAEP exam at or above the cutpoints you set for each achievement level in Round 3 will be reported for your consideration and evaluation. You will be asked to share your opinions regarding these percentages. Having seen these data, do you want to adjust your ratings? If so, you will have the opportunity to do so. A questionnaire has been developed to collect your opinions regarding the "consequences" of your ratings.

The consequences data reported to you are for your own cutpoints. Your recommendations will be for your own cutpoints. All recommended cutpoints will be averaged to compute the final cutpoints for your grade. These final cutpoints will be used for selecting exemplar items

and they will be reported to NAGB as the final cutpoints.

These data are provided to you as a last "reality check" on the relationship between student performance and the achievement levels you have set. The achievement levels descriptions state what students *should* know and be able to do. Your ratings indicate your judgements of how students *would* perform. The "consequences data" report how students *did* perform, relative to the achievement levels. We are interested in knowing whether you feel that the data about student performance is so compelling that you would recommend changes in your cutpoints.

Selection of Exemplar Items You will be asked to recommend exemplar items. Exemplar items are one of the primary outcomes of the ALS process. Exemplar items are used for reporting student performance on the NAEP relative to the achievement levels. Exemplars are items that illustrate knowledge and skills associated with each achievement level.

You will be given lists of items that meet the statistical requirements for consideration as "exemplars" for each achievement level. Specifically, students scoring within the upper and lower boundaries of each achievement level have an average probability of 50% or greater of responding correctly to the items on the list.

The items for consideration as exemplars for each achievement level will be included in either the "Primary" or the "Secondary" list. The items in the Primary list include about 60% of all items under consideration for each achievement level. Each item in the Primary list has a higher degree of *discrimination* than any of the items in the Secondary list. Items in both lists are ordered from the most-discriminating to the least-discriminating. You should select items from the Secondary list only if none in the Primary list can be recommended.

Polytomous items will be listed more than once, depending on the number of score levels. For example, a score of "2" on a given item could have a probability of at least 50% for students scoring within the Basic achievement level, and a score of "3" or "4" on the same item could have less than 50% probability for these students. You will then look in the list of Proficient and Advanced items to find the level(s) at which the probability of correct response for this item reaches at least 50% for scores of "3" or "4."

Process Evaluation Each day, you will be asked to complete an evaluation form covering the day's activities. These evaluations help to improve the process of setting achievement levels. In addition to the "scales" for selecting your responses, you will be given space to comment on any aspect of the ALS process.

Please note your responses will be scanned electronically. Your responses are very important to us, and it is very important that you mark them carefully so we will actually have your evaluation.

These daily evaluations are very important to the ALS process. One purpose is to identify your concerns and determine whether you are experiencing any difficulties with performing the tasks. In addition, we will analyze the evaluation data in conjunction with the rating data. Some questions are new for the 1998 ALS process, some have been included for the 1994 and 1996 ALS processes, and many of these questions have been asked of all NAEP ALS panelists since 1990.

You will also be asked to complete an evaluation form for the process as a whole.

Glossary of Terms[1]

Achievement Levels

Also known as "standards" or "performance standards." Three achievement levels will be set for reporting student performance on the NAEP: Basic, Proficient, and Advanced. Basic denotes partial mastery of prerequisite knowledge and skills that are fundamental for proficient work at each grade; Proficient represents solid academic performance for each grade assessed. Students reaching this level have demonstrated competency over challenging subject matter, including subject-matter knowledge, application of such knowledge to real world situations, and analytical skills appropriate to the subject matter; and Advanced signifies superior performance beyond.

Achievement Levels Descriptions (ALDs)

Statements describing what students should know and be able to do. The achievement level descriptions contain the essential aspects of the assessment framework appropriate to student performance at each level and grade.

Achievement Levels-Setting (ALS) Process

A judgmental process involving broadly representative panels of educators and noneducators. The process includes agreeing upon a common understanding of ALDs, rating items to set cutpoints representing student performance at the borderline of each achievement level, and selecting exemplar items to represent what students should know and be able to do at each level.

ACT

ACT, Inc., formerly known as American College Testing (contractor of NAGB). Responsible for designing and conducting the ALS process.

---

[1] Some terms in this glossary are defined for this specific context. These terms might be used differently outside the context of this achievement levels-setting process. There is a lot of jargon in this process, and we hope that this glossary will help you become familiar with important terminology.

ACT NAEP-Like Score Scale

The score scale used to report results to panelists during the ALS process. This score scale is a linear transformation of the NAEP scale, so there is a one-to-one correspondence between the two scales. This scale has been used to assure that the process is criterion referenced so you will not be unduly influenced by performance NAEP in other subjects.

Assessment

The test.

Block

A group or set of items forming a section of the NAEP exam. Blocks are timed to allow 25-minutes for students to answer the items. Each block contains 15 items for grade 4 and 19 items for grades 8 and 12, with the exception of one block for grade 8 that contains 18 items.

Booklet

The test form or instrument. A booklet is composed of two blocks of test items. The NAEP is administered to a student as one booklet. There are many combinations of blocks to produce many different booklet forms.

Borderline Descriptors

Descriptions of what students performing at the lower borderline of each achievement level should know and be able to do.

Borderline Performance

The level of performance that is minimally acceptable for each achievement level. In other words, the level of performance that just meets the criteria for each level.

***c* Parameter**

Pseudo-chance level parameter associated with each item on the NAEP exam. This number represents the probability that even lowest ability examinees will give the correct answer for the item.

Consequences Data

Information about how student performance is distributed with respect to the cutpoints you have set. The percentage of scores on the Civics NAEP exam at or above the cutpoint you set for each achievement level.

Constructed Response Item

An item requiring examinees to construct or supply a response, such as completion, definition, and essay items; *i.e.*, any item that is not a multiple-choice format.

Cutpoints

The points on the scale that represent or identify the boundaries between adjacent achievement levels. Cutpoints are established through the item rating process. (Same as cutscores.)

Cutscores (See Cutpoints)

Descriptors (See Achievement Levels Descriptions)

Discrimination (See Item Discrimination)

Dichotomous Items

Items that are scored (coded) as either "correct" or "incorrect." All multiple-choice items are scored dichotomously.

ETS

Educational Testing Service (contractor of NCES) responsible for developing the assessment questions according to specifications provided to NCES by NAGB, analyzing results, and working with NCES staff to prepare The Nation's Report Card and other reports on student achievement in various subject areas assessed by NAEP.

Exemplar Items

Items to illustrate knowledge and skills associated with each achievement level. Exemplar items are selected by panelists to use in reporting the NAEP results. They are a primary outcome of the ALS process.

Extended Constructed Response Items

Constructed response items that require longer responses. These items are graded for partial credit with at least four levels; *i.e.*, 1=unacceptable, 2=partial, 3=acceptable, 4=complete.

Feedback

Information provided to panelists between rounds of ratings. Most of the information is based on the ratings provided by panelists during the previous round(s) of ratings. Feedback data are presented to the panelists for their consideration and to inform their ratings in subsequent rounds.

Form (See Booklet)

Framework

The framework defines the aspects of the discipline that are to be assessed. It specifies the relative emphasis in measuring each component within the discipline for each grade level. The framework is the foundation for the assessment and for the ALS process. A consensus panel developed the framework, which guided the development of the assessment.

Grade Group

Panelists are assigned to set achievement levels for a particular grade (*i.e.*, 4th, 8th, or 12th). The group of panelists for each grade level is called a grade group. Tasks are performed in grade group sessions.

Illustrative Items (See Exemplar Items)

Item

A single question or exercise on the assessment.

Item Discrimination

The degree to which a test item differentiates between students performing at different achievement levels. The value is the difference between the average probability of a correct response at an achievement level, say Proficient, and the average probability of a correct response at the next lower level, that is Basic.

Item Pool

All the items for an assessment; *e.g.*, the NAEP item pool. Mostly used in the ALS process to refer to all items on the assessment for one grade level.

Item Rating Group

Panelists in each grade group are divided into two item-rating groups. The two groups are approximately equal in terms of panelist type and demographic characteristics. The item rating groups rate over half of the item pool for their particular grade.

Item Rating Pool

The item pool for each grade is divided into two item rating pools. Each item rating group rates one item rating pool. Item rating pools are approximately equal with respect to item difficulty, item formats, and other item characteristics.

Judges (See Panelist)

Mean

The average. The mean is used in several ways in the ALS process. For example, in the context of ratings, panelists estimate the mean (average) score on each polytomous item. In reporting student performance the mean is the average score of students on each polytomous item. In the context of reporting ratings, the mean is the cutpoint computed from the ratings for a grade group.

Mean Estimation Method

A method used in rating polytomous items. For each polytomous item, panelists will estimate the mean score for examinees performing at the borderline of each achievement level.

Modified Angoff Method

A method used in rating dichotomous items. For each multiple-choice item, panelists estimate the probability of a correct response for examinees performing at the borderline of each achievement level.

NAEP

National Assessment of Educational Progress. The test began in 1969 and is a primary indicator of the level of academic achievement for students in the U.S.

NAGB

National Assessment Governing Board. Created by Congress to formulate policy guidelines for NAEP. Board membership is broadly representative including K-12 classroom teachers, measurement experts, governors, legislators, and interested citizens.

NCES

National Center for Education Statistics. An agency of the U.S. Department of Education responsible for reporting education statistics including NAEP results.

NCS

National Computer Systems. Responsible for printing and scoring the NAEP exam.

Open Ended Item (See Constructed Response Item)

P-Value

The proportion of students who answered a dichotomous item correctly.

P-Value Feedback

Includes the p-value of each dichotomous item, and the mean for each polytomous item—plus the percentage of students scoring at each score level.

Panelists

Teachers, nonteacher educators, and members of the general public selected to participate in the achievement levels-setting process.

Partial

In the scoring guides, the term denotes a response to an open ended item that is *neither unacceptable nor complete*. That is, a code for a response that is only *partially correct*. Do not confuse a partially correct answer with partial mastery. (See Partial Mastery.)

Partial Mastery

Denotes a level of mastery of subject matter that is less than full mastery. Do not confuse with partially correct. (See Partial.)

Polytomous Item

Constructed response items that are scored for partial credit, *i.e.*, items that are ***not*** scored as either "correct" or "incorrect."

Rater Location Feedback

Data graphically presented to ALS panelists to provide information on the location of their ratings from the previous round with respect to the ratings of the other panelists in their grade group. The graphs show the cutscore set by each panelist in the grade group for each achievement level.

Raters

Panelists who are trained to rate items in the ALS process.

Rating Method

The method by which each item is rated in order to set achievement levels. See Modified Angoff Method and Mean Estimation Method.

Rating Process

The process by which items are rated to set achievement levels. Raters estimate student performance for each dichotomous item in their item rating pool using the Modified Angoff Method and each polytomous item using the Mean Estimation Method. The ratings for all the items in each item rating pool and for all panelists in each item rating group are combined to produce a cutpoint. There are three rounds of rating. Feedback data are provided between rounds of ratings.

Reckase Chart

A chart showing the expected percentage of correct responses for each multiple choice item and the expected score for each constructed response item at each score point. Each column reports the expected performance (percent correct or average score) for an item, given different levels of student performance; each row is a score on the ACT NAEP-Like scale. Rows are ordered from the highest to the lowest level of performance. The data in the chart reflect the characteristics of items administered to students taking the NAEP.

Reckase Method

A standard setting method for which item-by-item ratings are transferred to Reckase charts. Panelists adjust their ratings by examining them on the Reckase Charts and selecting the performance scale scores that best represent their concept of borderline performance for each achievement level.

Round of Rating

Part of the process during which every panelist rates each item in the item rating pool.

Scoring Guide

The answer key or reference to scoring rubrics and correct answers. For multiple choice items, the scoring guide is the

scoring key or the list of correct responses. For constructed response items, it is the scoring rubric.

Scoring Rubric

A list of correct responses, acceptable variations, and their corresponding scores (codes). It also includes the rationale for scoring each item and explanations for acceptable answers for each score (code) level.

Secret ID Code

A letter code assigned to each panelist. All panelists will have a "secret" ID code to enable them to see where their ratings are located, relative to ratings of other panelists in their grade group.

Short Constructed Response Item

Constructed response item that does not require an extended response. Most items of this type are scored (coded) as polytomous items with three score points or levels (1=unacceptable; 2=partial; 3=acceptable), but a few are scored as dichotomous ("right/wrong") items.

Standard Deviation

The standard deviation is a statistical measure of the amount of variability in ratings by panelists in your grade group. The bars representing the standard deviation of your cutpoints are one standard deviation above and below the cutpoint for the achievement level. One standard deviation above and below the cutpoint includes 68% of the cutpoints set by panelists in the grade group. The greater the variability in cutpoints set by panelists, the longer the bar will need to be to represent 68% of the scores.

Student Booklet

A student's NAEP exam booklet.

Westat

Contractor responsible for the sampling and test administration for the NAEP.

Whole Group

Composed of all the panelists from all three grade groups.

Whole Group Session

Sessions involving all panelists. These sessions are recommended for consistency in standard setting so that all panelists at all grades hear the same instructions and explanations and see the same examples. Training and instructions are provided in whole group sessions and implemented in grade group sessions.

Information about Chance

The probability of a student correctly answering an item is a function of the difficulty of the item, the ability of the student, discrimination attributes of the item, and chance.

We assume that for items of equal difficulty, the probability of a correct answer increases with increasing ability of the student.

We assume that for students of equal ability, the probability of a correct answer increases with decreasing difficulty of the item.

We assume that laws of probability operate in a test-taking environment just as they do in dice-tossing environment. This means that there is some probability of a correct answer that is "independent" of the difficulty of the item and the ability of the student.

Therefore, there is some probability that even low ability students will correctly answer difficult items. Likewise there is some probability that high ability students will incorrectly answer easy items.

In the case of high ability students incorrectly answering easy items, we attribute that to "random error" in the model.

In the case of low ability students correctly answering items, we have some information about those chances or probabilities. The probability of correctly answering an item with three choices is 1-in-3 or 33%; with four choices, it is 1-in-4 or 25%; and for an item with five choices, the probability is 1-in-5 or 20%.

When you estimate the percentage of students at the borderline of each achievement level who would correctly answer an item, you should remember the "laws of probability," i.e., the probability of giving a correct answer by random guessing. Even students at the lower borderline of the Basic achievement level will have a probability of correctly answering an item that is at least equal to the chance probability. To estimate a lower percentage is illogical or a denial of the laws of probability, or both!

Notes