

# **Documenting the Technical Quality of Your State's Alternate Assessment System: An Annotated Workbook Volume I: "Nuts and Bolts"**

**Developed in partnership with the  
New Hampshire Enhanced Assessment Initiative (NHEAI), the  
National Alternate Assessment Center (NAAC), and the  
National Center for the Improvement of Educational Assessment (NCIEA)**

**October 2, 2006**

## **CHAPTER 1: OVERVIEW OF THE ASSESSMENT SYSTEM**

This chapter orients the reader to the set of technical documents and to the assessment system. It should describe the full assessment system and how the AA-AAS fits into this system.

### **Introduction to the “Nuts and Bolts” Volume**

This chapter is an advance organizer to let the readers know what a technical manual does/does not do and what they can expect to find throughout the document. This section also alerts readers that the manual is organized around a validity framework.

#### **Rationale for this Content**

This type of advance organizer should be found in almost all documents of this size and nature. It helps orient the reader to a general overview of the document and the purpose of technical documentation.

#### **Data Sources:**

This introductory section should be written as the state is deciding on the best way to organize its technical documentation.

#### **Guiding Questions:**

1. How are this volume and associated documents organized?
2. Why did you organize the technical documentation in this manner?
3. How does this manual and assessment system fit with the larger state assessment system and collection of technical documentation?
4. Who is your target audience?

#### **Notes**

## **Technical Documentation Workbook NHEAI/NAAC**

### **Statement of Core Beliefs and Guiding Philosophy**

This is where the state leaders present their “mission” statement. This chapter contains explicit statements of beliefs and values about the state’s educational system for all children and how the assessment system is intended to support this view of education. This chapter should also include a description of how the alternate assessment system is part of the larger state assessment system. In the case of this manual, state leaders must address how instruction for students with the most significant cognitive disabilities and the alternate assessment interact to support this mission.

### **Rationale for this Content**

This is where the state must be explicit about beliefs, and how state leaders intend to see those beliefs instantiated, in part through the design of the assessment systems. These core beliefs and guiding philosophies should be logically connected to the purposes and uses of the alternate assessment system in the next chapter.

### **Guiding Questions:**

1. What does the state see as the major purposes for its public education system?
2. How do these purposes relate to the education for students with the most significant cognitive disabilities?
3. How does the alternate assessment system fit into the larger state assessment system?
4. How are the core values supporting alternate assessments on alternate achievement standards similar to those supporting the general education assessment and how and why are they different?

### **Notes**

## **Technical Documentation Workbook NHEAI/NAAC**

### **Purposes of the Assessment System**

The state describes, in this chapter, the purposes for developing its AA-AAS system. For example, NCLB accountability and IDEA 1997 & 2004 are often key reasons for developing these systems. There are almost always governing statutes and regulations at the state level and the state may often articulate other purposes of the system (e.g., instructional change).

### **Rationale for this Content**

This chapter provides the reader with the context of the state assessment system in which the alternate assessment functions. Validity can only be evaluated in the context of the purposes of the assessment(s) and how the results are used (next chapter). State leaders should be clear that if a purpose is specified in this section, evidence should be collected to evaluate the validity of the assessment related to this purpose.

### **Data Sources:**

Enabling legislation, design documents, state board minutes, minutes from constituent group meetings (if applicable), and RFP documents are all potential sources of information to document the purposes of the assessment system.

### **Guiding Questions:**

1. Given all the potential purposes, what are the primary purposes of the system?
2. What are the governing statutes providing the legal authority for the system?
3. What do these legal documents require in terms of purposes?
4. How are the purposes of the AA-AAS consistent with the purposes of the entire system?
5. How has the state ensured that its assessment system will provide coherent information for students across grades and subjects? (Peer Review Guidance p. 3).

### **Notes**

## **Technical Documentation Workbook**

### **NHEAI/NAAC**

#### **Uses of the Assessment Information**

This crucial section identifies the intended uses of the inferences drawn from the assessment results. These uses should be described for individual students, schools, and any other levels for which the results will be used.

#### **Rationale for this Content**

As mentioned above, specifying the intended uses of the assessment results is critical for building the validity argument. We only validate assessments for the way in which the results are used and each use needs to have validity evidence to support it.

#### **Data Sources:**

Enabling legislation, design documents, state board minutes, minutes from constituent group meetings (if applicable), and RFP documents are all potential sources of information to document how the results of the assessment system are to be used. Additionally, score reports, interpretative documents, professional development workshops can all provide data to describe the uses of the assessment results.

#### **Guiding Questions:**

1. Do the documents mentioned above describe how the results are to be used?
2. Does the state offer guidance to local educators about how to use the assessment scores?
3. Are there specific requirements for how the scores are to be used?
4. How are the data derived from the assessment system being used (e.g., accountability, program evaluation, instructional feedback)?

#### **Notes**

## **CHAPTER 2: WHO ARE THE STUDENTS?**

This chapter is designed to describe, as completely as possible, the students participating in the AA-AAS. This is crucial for building the validity argument framed around the assessment triangle.

### **Quantitative and Qualitative Description of the Students Participating in the AA-AAS**

This chapter should present the numbers of students participating in the AA-AAS by specific disability and other relevant characteristics. More important than the quantitative information is the information about how these students learn or struggle to learn, how they are taught, and confounding issues such as medical conditions.

### **Rationale for this Content**

In order to build a validity argument, we need to have a good understanding of who is participating in this assessment. This is not meant to limit who participates, but simply to get as accurate of an understanding of the participants as possible.

### **Data Sources:**

- State and federal special education data bases indicating the counts of students participating in the AA-AAS by disability code and any other pertinent information if possible may be used.
- Results from demographic data other than disability label that describe characteristics of assessment population are critical sources of information for this chapter.
- IEP reviews could be a good source of information to gain a better understanding of the learning characteristics of students participating in the AA-AAS.

### **Guiding Questions:**

1. How many students by specific disability category participate in the AA-AAS?
2. What are the characteristics of the learners that differentiate them from students in the general assessment?
3. How congruent is the description of the intended population to the actual assessed population?

### **Notes**

### **CHAPTER 3: WHAT IS THE CONTENT?**

This chapter is designed to have the state describe, as completely as possible, the content expectations for students participating in the AA-AAS. This information is critical for defining the domain that must be instructed and assessed.

#### **Description of ELA and Mathematics Content and Performance Expectations**

States will need to thoroughly describe the content and performance expectations for students participating in the AA-AAS to help define the domain for instruction and assessment.

#### **Rationale for this Content**

This is a necessary first step in the design of any assessment. The content and achievement domain must be defined for both instruction and assessment. Aspects of the validity argument (e.g., content validity, alignment) cannot be evaluated without these definitions (Peer Review Guidance, p. 4).

#### **Data Sources:**

State content and achievement standards, documentation of the processes used to create such standards, and research supporting the design of the standards would all be data sources for this chapter.

#### **Guiding Questions:**

1. Has the state approved/adopted challenging academic standards in reading/language? (Peer Review Guidance, p. 4).
2. Who was involved in writing/articulating the content standards linkages for the alternate assessment? What were the qualifications of the individuals involved in the articulation of the standards? (Peer Review Guidance, p. 4).
3. What research was used to support the inclusion and exclusion of certain content?
4. How were the performance expectations determined? (note: this is covered in more detail in the standard setting chapter)
5. Are the content standard linkages challenging for the population? (Peer Review Guidance, p. 4).
6. Are the content standard linkages uniform for all students or are they a “menu” from which IEP teams and instructors are expected to choose?

#### **Notes**

## **CHAPTER 4: TEST DEVELOPMENT**

This section describes the assessment approach (e.g., portfolio, checklist/rating scale, performance event, multiple choice, or combination) and the procedures used to align/link items to the grade-level content. This chapter is a good place for the state to describe how it approached the issue of standardization and flexibility, particularly the dimensions of the assessment system where the state was willing to permit greater flexibility and where it wanted to standardize specific aspects of the system.

Examples of types of items/tasks and descriptions of the extent to which stakeholders were involved in the development process should also be presented in this chapter. In addition, this chapter includes the results of any field/pilot testing of items. Descriptions of consideration of the elements of universal design would be described here. Bias review procedures would be reported in this section. Bias review should not only include the typical bias information but also must be sensitive to issues under IDEA (e.g., least restrictive environment). Obviously, many aspects of what each state includes in this chapter are contingent upon the particular approach that state has adopted for its alternate assessment.

### **Rationale for this Content**

This section provides the reader with a clear description of the approach and how the decision-making process is linked with the beliefs articulated in Chapter 1. Examples of items/tasks help illustrate the assessment requirements from the teacher and the student perspectives. Documentation of alignment procedures provides evidence that careful consideration has been given to accessing grade-level content.

### **Data Sources:**

- Minutes from stakeholder group meetings, directions for teachers or alignment groups results of alignment studies.
- Examples of items and descriptions of the types of student responses
- Description of alignment procedures
- Results of field /pilot tests
- Results of bias review
- Results of review or consideration of Universal Design

### **Guiding Questions:**

1. What is the approach?
2. How were/are the items/procedures developed?
3. To what extent were/are stakeholders involved in the design process?
4. What is required of the student, teacher?
5. How was alignment/linkage to grade-level content determined?
6. What were the results of the field/pilot test?
7. What were the results of the bias review?
8. How was Universal Design considered?

**Notes space on next page**

**Technical Documentation Workbook  
NHEAI/NAAC**

**Notes**

## **CHAPTER 5: “ITEM” ANALYSIS**

This type of information is almost always presented in traditional technical manuals. The extent to which item analysis makes sense with an AA-AAS is dependent on the specific form of the alternate assessment, but in most cases, some form of systematic item/task examination makes sense.

### **Traditional Item Analyses—e.g., Difficulty and Discrimination**

This chapter should present the results of systematic examination of the difficulty and discrimination of the items, if this type of examination is supported by the form of the assessment.

### **Rationale for this Content**

Although there may be considerably more flexibility associated with an AA-AAS than a regular assessment, some form of a systematic evaluation of the test items/tasks/procedures should be undertaken if possible.

### **Data Sources:**

- Test items: Both common items and a sample of unique items
- Response patterns on common items
- Response patterns by indicator if individual test items are unique to each student
- Demographic information for each student

### **Guiding Questions:**

1. What is the variation in the number of points earned for each item?
2. What is the variability in the average number of points for each indicator, if applicable?
3. What is the relationship of the pattern among the item responses as they relate to other items and the total test score?
4. Are there differences in item and indicator patterns by important subgroups?

### **Notes**

## Technical Documentation Workbook NHEAI/NAAC

### **Examining Bias (in addition to the usual groups, by disability)**

Regular assessments are almost always evaluated using Differential Item Functioning (DIF) methods to screen for potential item bias. DIF procedures, however, require larger samples of students than are usually available on AA-AAS. Further, DIF is typically employed to search for bias against fairly well-defined gender, ethnic, and poverty groups. Nevertheless, the state still needs to examine the items/tasks/procedures for potential bias or other unfairness against certain types of students participating in the AA-AAS.

#### **Rationale for this Content**

These analyses are critical to ensure that the assessment is as fair as possible for as many students as possible. It will be crucial for the state to define groups to be examined. It can be argued that traditional gender and ethnic categories do not make sense in the case of the AA-AAS, but it might make sense to examine items for potential bias against certain types of disabilities or medical conditions.

#### **Data Sources:**

- Test items/tasks
- Patterns of performance
- Significant disabilities experts

#### **Guiding Questions:**

1. Does the state have a process in place to review items/procedures for fairness/bias?
2. Who leads these reviews?
3. What type of training do the reviewers receive?
4. Are there any post-hoc analyses that the state conducts to evaluate fairness?
5. What are the results of these analyses?
6. What actions have the state and/or test developer taken in response to these reviews?

#### **Notes**

## **CHAPTER 6: ALIGNMENT**

Alignment, as a technical criterion, has received perhaps the most attention in standards-based education. Alignment among the various aspects of the system—e.g., content, curriculum, assessments, and instruction—is thought to be a requirement for the educational system to function as intended.

### **Rationale for this Content**

This is a legal requirement, but most importantly it is a critical educational requirement to ensure that all students are instructed and assessed on the essential grade-level content, whether their work is evaluated against regular or alternate achievement standards.

### **Data Sources:**

- Content standards linkages
- AA-AAS test blueprints (if applicable)
- State content standards
- Curriculum documents
- Alternate achievement performance level descriptors

### **Guiding Questions:**

1. What is the relationship between the grade level content standards and the content used to guide the development of the AA-AAS? In other words, is there a difference in the depth, breadth, and/or complexity of the content domain compared to the grade-level standards?
  - a. How was this relationship evaluated—what protocol(s) were used?
  - b. Who was involved in evaluating this relationship?
  - c. What were their qualifications?
2. What is the relationship among the tasks/items that comprise the alternate assessment, the grade-level content standards, and the “expanded” standards?
  - a. How was this relationship evaluated—what protocol(s) were used?
  - b. Who was involved in evaluating this relationship?
  - c. What were their qualifications?
3. How do the alternate achievement standards link the content of the alternate assessment with scores on the assessment?
  - a. Do the performance level descriptors convincingly describe student performance in relation to the grade-level content linkages of the AA-AAS?
  - b. How were the performance level descriptors (referencing grade-level content) used in the process of setting achievement standards?

### **Notes**

**Alignment of Alternate Achievement Standards with**  
**the State Academic Content Standards**

This chapter presents the results of analyses examining the relationship between the state grade-level content standards and the alternate achievement standards.

**Rationale for this Content**

By definition, AA-AAS are based on different achievement standards compared with the regular assessment, yet, alternate achievement standards should be aligned/linked to grade-level content standards.

**Data Sources:**

- State grade-level content standards
- Extended/expanded content standards, if any
- AA-AAS performance level descriptors and achievement standards

**Guiding Questions:**

1. Do the performance level descriptors convincingly describe student performance in relation to the grade-level content linkage reflected in the AA-AAS?
2. Do the alternate achievement standards reflect expectations that calls for reduced depth, complexity, and/or breadth compared with the regular assessment achievement standards?
3. How were the content expectations used when writing the achievement standards?
4. Do both the content expectations and the alternate achievement standards serve to guide test design?
5. Do the alternate achievement standards promote access to the general curriculum?
6. Do the alternate achievement standards reflect professional judgment of the highest achievement standards possible

**Notes**

**Description of Linkage to Different Content Across Grades**  
**that Supports Individual Growth**

This chapter could be folded into separate ELA and math chapters, but it is placed here as a reminder of the importance of attending to this topic.

**Rationale for this Content**

Expecting that content standards across grades are logically (and theoretically) connected is the case for regular assessments and we should expect these connections for AA-AAS content linkages.

**Data Sources:**

Data sources and documents that were used above should provide the evidence necessary to address this chapter.

**Guiding Questions:**

1. Is there a clear and logical connection of the content expectations across grade levels?
2. Is there any evidence that these connections are empirically supported?
3. Does the state have evidence that students with the most significant cognitive disabilities progress through the content expectations as articulated in the standards?

**Notes**

## **CHAPTER 7: ADMINISTRATION & TRAINING**

This section describes the procedures for administering the assessment, the role of the IEP team regarding the assessment decisions (e.g., items, participation criteria), and the training requirements/opportunities for raters/observers. In addition, the monitoring and quality control procedures are described.

### **Rationale for this Content**

The extent to which the assessment program adheres to procedures for administration, beginning with IEP decision-making teams through monitoring of assessment information is necessary for ensuring procedural fidelity as a validity measure.

### **Data Sources:**

- Inter-observer/rater manuals and resources
- Training procedures and materials
- IEP team instructions
- Training agendas, evaluations, and training participation results
- Test security manuals and procedures

### **Guiding Questions:**

1. What are the procedures for administering the assessment?
2. Who is responsible for administration and at what levels?
3. How are raters/observers trained, monitored, and evaluated?
4. What guides the quality of the training?
5. What professional development opportunities are provided in addition to the training?
6. What instructions are provided to IEP teams?
7. How consistently are those instructions applied?
8. What are the administration quality control procedures?
9. How are test coordinators, administrators, and others made aware of test security and ethical requirements associated with administering the alternate assessment?

### **Notes**

## **CHAPTER 8: SCORING**

This section describes the scoring procedures implemented, by particular assessment approach. This will necessarily include scoring rules and criteria. This may require a description of how the level of student independence is considered in scoring. This section may also include: description of the scoring approach, central vs. distributed, and the rationale for that approach; descriptions of range-finding procedures as well as the selection and use of anchor papers; selection, training, and quality control of scorers; and description of scoring accuracy and consistency.

### **Rationale for this Content**

This section allows the reader to understand and judge the quality of the scoring procedures and the rationale for the procedures currently implemented.

### **Data sources:**

- Reports from range-finding events
- Scoring training manuals
- Scoring exemplars
- Scoring rules and procedures
- Selection, training, and monitoring of scorers
- Scoring accuracy data

### **Guiding Questions:**

1. How are scores derived for this particular approach?
2. To what degree are independence and/or adaptive supports considered in scoring?
3. What other dimensions besides independence are scored?
4. How are scoring exemplars selected?
5. How are scoring rules and procedures determined?
6. How are scorers selected, trained & monitored during the scoring process?
7. How accurate is the scoring?
8. How consistent is the scoring?
9. What is the scoring distribution?

### **Notes**

## **CHAPTER 9: CHARACTERIZING ERRORS ASSOCIATED WITH TEST SCORES**

This chapter will present the methods and results of analyses designed to characterize the measurement and sampling errors associated with test scores.

### **Levels of Analysis**

This chapter will have to be logically connected to the chapters addressing the purposes and uses of the assessment system. The state must clarify how they intend to use student- (if at all) and school-level scores.

### **Rationale for this Content**

The level of the educational system for which the scores will be used will dictate the types of reliability/consistency analyses necessary. For example, if the student level results are limited to serving as another source of information about the student, then traditional reliability/measurement error analyses are much less important than school- or district-level decision consistency analyses.

### **Data Sources:**

- Data files of student-by-item responses
- Results of interrater reliability studies (if applicable)
- Documentation of procedural fidelity for the administration and scoring of the assessment
- Results of any studies examining the influence of specific administrators on the performance of students
- Data documenting the variability of performance across occasions
- Business rules for school-level accountability calculations

### **Guiding Questions:**

1. What decisions—and at what level of the educational system—are made as a result of the assessment scores?
2. Has the state determined the reliability of the scores it reports, based on data for the student population? (Peer Review Notes, p. 14).
3. What are the major sources of error that should be accounted for to accurately portray the reliability of the alternate assessment scores, such as different raters, test administrators, occasions of the assessment, and the collection of test items/tasks and/or observations?
4. What data does the state currently have regarding the sources of error identified in question #3 and what data will the state be able to collect about other sources of error?
  - a. Is this state able/willing to conduct any special studies to gather data about such things as administrator and occasions effects?

### **Notes**

**Decision Consistency and Accuracy**

We know that, at a minimum, scores from the alternate assessment must be used for school accountability purposes and the consistency of these decisions should be evaluated. Further, it is likely that the students' proficiency designations will carry some meaning. If this is the case, it will be important to evaluate how consistently and accurately the decisions are made.

**Rationale for this Content**

Traditional reliability analyses are still important, if the statistics can be reasonably computed, but decision consistency and accuracy are much more important because of the focus on classifying students into performance categories.

**Data Sources:**

- Data files of student-by-item responses
- Business rules for school-level accountability calculations

**Guiding Questions:**

1. What are acceptable levels of decision consistency/accuracy for school-level decisions?
2. What are acceptable levels of decision consistency/accuracy for student-level decisions?
3. What are the methods for calculating decision consistency and accuracy?
4. What are the reasons for choosing these particular methodological approaches?

**Notes**

**Classical Reliability Analyses**

While the data generated from alternate assessments do not often lend themselves to traditional reliability analyses, some forms of AA-AAS may yield data that can be analyzed in this manner.

**Rationale for this Content**

It can be argued that many of the required assumptions for using traditional reliability methods are not met with most AA-AAS. However, states still might consider employing such methods, if the data are available, to get a sense of the upper bound of reliability. However, if the traditional reliability coefficient appear spuriously high (e.g., greater than 0.92 or so), it is likely due to fact that traits such as independence or functionality are strongly related to item performance, which therefore violates a major assumption of item independence when computing reliability coefficients such as Cronbach's alpha.

**Data Sources:**

- Student-by-item data files

**Guiding Questions:**

1. Do the data meet the assumptions for the reliability approach employed such as relative independence of item responses and uncorrelated errors?
2. What are the methods used to estimate reliability?
3. Why were these methods selected?
  - a. If generalizability analyses were not conducted, why were these methods chosen instead?
4. What are the results of the analyses?
5. What are the implications of these analyses for the validity of the score inferences and for future decisions about test design?

**Notes**

## Technical Documentation Workbook NHEAI/NAAC

### Generalizability Analyses

These analyses, like traditional reliability analyses, will be constrained by the structure of the data and limited number of students participating in the AA-AAS. But, if the data are available, generalizability is the most appropriate way to conceptualize measurement error.

#### **Rationale for this Content**

Generalizability approaches, if the data support such analyses, are the most appropriate methods for characterizing measurement error in general, and for AA-AAS specifically. Traditional reliability analyses only allow us to distinguish between error, observed, and true variance, but generalizability allows us to quantify the error variance associated with facets such as raters, tasks, occasions, and, perhaps most importantly, administrators.

#### **Data Sources:**

- Student-by-item data files that include data for facets of the system for which we would like to estimate variance, such as
  - Raters,
  - Tasks,
  - Occasions, and
  - Administrators.
- Videos of test administration and scoring, etc
- Data describing any evaluations of procedural fidelity

#### **Guiding Questions:**

1. Are the data appropriate for the generalizability approach employed?
2. What g-study and d-study designs were employed?
3. Why were these designs selected?
4. What evidence of generalizability for all relevant sources has been reported?

#### **Notes**

**CHAPTER 10: COMPARABILITY (SCALING AND EQUATING)**

This chapter presents the methods and results for ensuring comparability among scores from different students within and across years as well as from the same students across occasions.

**Choice of Scale and Rationale for Choice**

Most alternate assessments transform the raw scores into some type of scale to facilitate comparisons across tests. This chapter should describe the choice of scale, the rationale for the choice, and methods for transforming the raw scores into the scale.

**Rationale for this Content**

In almost all cases, raw scores should be transformed onto a measurement scale that conveys more meaning than simple raw scores.

**Data Sources:**

- Test scores and distributions, scale choice, and transformational methods.

**Guiding Questions:**

1. How many raw score points are associated with the AA-AAS?
2. What is the choice of score scale and what is the rationale for this choice?
3. What methods have been used to transform the raw scores to scale scores?
4. Are these scales articulated across grades?

**Notes**

### **Comparability of Scores Across Years**

This chapter is framed in a more general way than a typical equating chapter found in a regular assessment technical manual. Equating is simply a means of ensuring comparability of score inferences across different forms of the test. Equating certainly could be a component of this chapter, but it is likely that other forms of establishing the comparability of score inferences will need to be presented.

#### **Rationale for this Content**

If a single assessment score is a sample of behavior from which we would like to generalize to a larger domain, then we need to have some way of ensuring the comparability of score inferences for multiple tests that are all supposed to be tapping the same domain. This could mean different forms of the test administered to different students whether in the same year or across years.

#### **Data Sources:**

- Item-level information and complete score distributions from multiple forms of the assessment
- In the case of establishing the comparability of scores from a small number of assessments (e.g., the same student over multiple occasions), data should include complete descriptions of the tasks, responses, and judgments.

#### **Guiding Questions:**

1. If equating methods have been employed, what were the equating design, methods, results, and decisions (e.g., eliminating items from the linking pool)?
2. If traditional equating methods are not appropriate or available (because of limited sample size), what rules have been established for judging two or more assessment scores to be comparable (or comparable enough given the purposes and uses specified)?

#### **Notes**

**Linkage Across Grades (measuring growth)**

Many states have an interest in documenting growth in students' knowledge and skills across grades. This chapter should present a description of how the state is doing this and the results of this linkage.

**Rationale for this Content**

While many have an interest in measuring growth through traditional methods like vertical score scales and less traditional approaches such as vertically-articulated achievement standards, it is not as straightforward as people hope even with regular assessments. The challenges—as a result of small numbers of students and greater flexibility in the assessment system—are much more significant for AA-AAS than for regular assessments. If the state intends to make claims about growth, they need to thoroughly describe the methods for establishing cross-grade links and should also provide evidence of the validity of these linkages.

**Data Sources:**

- Item-level information and complete score distributions from multiple forms of the assessment across grades
- In the case of establishing the comparability of scores from a small number of assessments (e.g., the same student over multiple grades), data should include complete descriptions of the tasks, responses, and judgments.

**Guiding Questions:**

1. If vertical scaling methods have been employed, what were the linking designs, what specific methods, and results of such approaches?
2. Have the performance level descriptors and associated cut scores on the assessment been vertically-articulated across grades such that a proficient score, for example, is intended to have the same relative meaning across grades?
3. If neither of these methods have been used because they are not appropriate or available (because of limited sample size), what rules have been established for judging two or more assessment scores to be comparable (or comparable enough given the purposes and uses specified)?

**Notes**

## **CHAPTER 11: STANDARD SETTING**

This chapter should provide a description of the methodology used to set cut scores, the reason for choosing this method(s) and for not selecting other methods.

### **Rationale for this Content**

This is one of the most visible decisions that a state will make and it is crucial to provide a clear rationale for the specific methodology selected. Documentation of the standard setting process is required (Peer Review p. 14) including: the selection of judges, methodology employed, and final results.

### **Data Sources:**

- Student scores and frequency distributions
- Literature reviews to support methodological choice
- Item difficulty and item mapping information (depending on method)
- Student work samples (depending on method)

### **Guiding Questions:**

1. Do the data lend themselves to one type of method over others?
2. Have the performance descriptors been written and do these suggest a particular methodology?
3. How have panelists been selected? What is the nature of their expertise? Do they represent a diverse group of stakeholders?
4. Have the protocols been modified for use with alternate assessments?
5. What type of training did the panelists receive?
6. Were any smoothing procedures used?
7. Was impact data introduced? If so, when in the process?

### **Notes**

### **Standard Setting Results**

This chapter presents the results of all of the decisions and procedures described in the previous chapter. It presents the cut scores in the appropriate metric and the percentages of students scoring in each category.

#### **Rationale for this Content**

All of the decisions affecting the final standard setting results need to be well documented because of the implications for the validity of the score inferences.

#### **Data Sources:**

- Results of panelist evaluations
- Cut scores in the appropriate metric
- Cumulative frequency distribution

#### **Guiding Questions:**

1. What were the unadjusted results?
2. What were the adjusted/smoothed results?
3. What were the policy decisions in view of the unadjusted and adjusted results?
4. Are the standards coherent across grade levels?
5. Are the standards coherent across subject areas?

#### **Notes**

## **CHAPTER 12: REPORTING**

The reporting section descriptions should align with purposes and uses described earlier. Reports may be disseminated at the student, building, LEA and SEA levels. The process for determining the appropriate way to report the results of the assessment should be described including the extent to which stakeholders were involved in the process. The information that schools and parents receive should be provided in this section.

The extent to which reports for various constituencies adhere to the **Standards for Educational and Psychological Testing** (AERA, APA, NCME, 1999) should be described here. Summary scores of students, schools, and other stakeholders should be described.

### **Data Sources:**

- Stakeholder meeting reports
- Reporting procedures from RFP
- Score interpretation guides
- Student, school data
- Parent letters

### **Guiding Questions:**

1. What constituencies receive reports?
2. What is the critical information that should be shared?
3. Who receives reports and interpretation guides?
4. Have the report formats been reviewed by members of the intended audiences?
5. What level of student data may be reported and at what levels may these data be reported?
6. Do the reports comply with the recommendations found in the **Standards for Educational and Psychological Testing** (AERA, APA, NCME, 1999)?

### **Notes**