An Annotated Workbook for Documenting the Technical Quality of Your State's Alternate Assessment System: Volume II: The Validity Evaluation

Developed in partnership with the New Hampshire Enhanced Assessment Grant (NHEAI), the National Alternate Assessment Center (NAAC), and the Center for Assessment

October 6, 2006

CHAPTER 1: OVERVIEW OF THE ASSESSMENT SYSTEM

This chapter is designed to orient the reader to both the validity evaluation and assessment system. It should describe for readers the full assessment system and how the AA-AAS fits into this system. This chapter will also introduce the validity framework.

Introduction to this Volume

This section of Chapter 1 is essentially an advance organizer to inform the readers how this manual is organized and how if fits with the rest of the technical documentation.

Rationale for this Content

This type of advanced 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 the validity evaluation.

Data Sources:

This introductory section will have to be written as the state is deciding on the best way to organize its presentation of validity evidence.

Guiding Questions:

1. How is this document (and associated documents) organized?
2. Why did you decided to organize the validity evaluation in this manner?
3. How does this volume and assessment system fit with the larger state assessment system and collection of technical documentation?
4. Who is your target audience for this volume?

Notes

Statement of Core Beliefs and Guiding Philosophy

This is where the state leaders must 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 challenges and the alternate assessment system supports this mission.

Rationale for this Content

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

Purposes of the Alternate Assessment System

The state describes, in this chapter, the purposes for developing its AA-AAS system. For example, NCLB accountability and IDEA 1997 & IDIEA 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 system 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. Has the state specified the purposes of the assessments, delineating the types of uses and decisions most appropriate for each? (Peer Review Notes, p. 13; Standards, for Education and Psychological Testing (AREA/APA/NCME, 1999).
2. Given all the potential purposes, what are the primary purposes of the system?
3. What are the governing statutes providing the legal authority for the system?
4. What do these legal documents require in terms of purposes?
5. How are the purposes of the AA-AAS consistent with the purposes of the entire system?
6. 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

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 (or not 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 have the state 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 these students learn or struggle to learn, how they are taught, and confounding issue 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.
- 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

How do they Learn? Models of Domain Proficiency.

Following directly from the previous chapter, this chapter is intended to have states focus more specifically on how students are expected to develop proficiency (i.e., increase their knowledge and skills) in ELA and mathematics.

Rationale for this Content

The construct validity of any achievement test (as opposed to an IQ test or personality measure, etc.) should be dependent on a demonstrated connection between what the test is intended to measure and the instruction and curriculum designed to improve students' knowledge and skills. This does not mean that instruction should focus solely on tested content, but good curriculum and instruction should lead to increases in test scores to show that the test is a measure of achievement and not some instructionally-irrelevant trait. As part of this aspect of the validity evaluation, one must be able to articulate how students develop competence in the particular domain.

Data Sources:

Almost all of the work in this chapter should be drawn from existing sources. In other words, states should not be creating learning theories, but state assessment, curriculum, and special education leaders should have surveyed the existing literature to determine the view of learning for this population on which they will base their assessment system. Ideally, this would have occurred during the assessment design phase, but we are well aware this is not the case for most existing systems. Therefore, we suggest that state leaders still pursue this work to help them better align the three vertices of the assessment triangle for their system. The Kleinert, Browder, and Towles-Reeves (2006) literature review may serve as a useful starting point for states undertaking this work. For states that are not satisfied with the existing learning models, they may pursue conducting cognitive laboratories and collections of student work to build models of domain proficiency aligned with the content expectations in their state.

Guiding Questions:

1. What model of domain proficiency (e.g., reading, math) is guiding the development of the assessment and the interpretation of the results?
2. Has the state provided content expectations in reading and mathematics that are articulated across grades?
3. Is there any research in the state or elsewhere that supports the continuum of developing expertise implied by the content standards?
4. Is there any research supporting the assumption that students progress through the content standards as they are currently articulated?

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 is information is critical for defining the domain that must be instructed and assessed.

Description of ELA Content 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, 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 arts? (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: INTRODUCTION OF THE VALIDITY FRAMEWORK AND ARGUMENT

The validity framework guiding this evaluation is introduced here to frame the upcoming chapters.

Rationale for this Content
We have argued that the technical documentation of assessment systems must be organized around a validity framework in order to properly evaluate whether or not the inferences about students and schools, as a result of the assessment scores, can be supported. This volume is focused on the validity evaluation so the nature of the argument must be presented here to frame the evaluation.

Data Sources:
There are probably no existing state documents or other data sources to help states write this section. Rather, this will require familiarity with validity or having someone conducting a literature view and then having the state adopt a specific standpoint.

Guiding Questions:
1. What is the state's conception of validity and is this conception supported in the theoretical literature?
2. How did the state arrive at this conception?
3. How does the state intend to use the validity evaluation to improve the assessment and education of students with the most significant cognitive disabilities?

Notes

Logical Connections Among the Content, Learning Models, and the Assessments

The assessment triangle is a heuristic method to describe the relationship among learning, assessment and interpretation. Whether the state relies on this approach or another one, it needs to explain the relationship among the assessment, the content of the assessment and instruction, and models of cognition explaining how students come to know this content.

Rationale for this Content

When implementing an assessment system to monitor—among other things—the change in performance of students and schools, we must evaluate the relationship among curriculum, instruction, and assessment. This is critical because we must document that improvements on the assessment can be linked to instruction and curriculum and not to out of school factors (e.g., level of functioning).

Data Sources:

- This is another chapter that must be generated by the state based on a process where theories and assumptions are vetted with key policy makers and others with expertise and an interest in the conceptualization. Ideally, curriculum, instruction, and assessment personnel must collaborate on this chapter.

Guiding Questions:

1. How does the state explain the relationship among curriculum, instruction, and the AA-AAS?

2. Does the state have evidence supporting its view that the assessment is sensitive to good curriculum and instruction?

Notes

Prioritized Validity Evaluation Questions

Evaluators cannot address all possible validity questions; therefore it is crucial for the state to prioritize its evaluation questions. The state will need to consider how to frame the validity evaluation around the prioritized purposes and uses for the AA-AAS. This prioritization should extend over time so that current and future studies are discussed.

Rationale for this Content

The rationale for this chapter is straightforward. The state cannot do everything, so it must decide what is most important to do first in terms of validating the inferences from its assessment system. The validity of the inferences of assessment scores can only be evaluated in the context of the purposes of the assessment and uses of the scores. The purposes and uses can be implied, but it is always more defensible if these are made explicit. The validation priorities should reflect the highest priority purposes.

Data Sources:

Surveys of key constituents, legislative mandates and/or expectations, and other information collected from stakeholders including concern (or hopes) regarding the consequences of the assessment system are all sources of information to help the state prioritize its evaluation agenda. Further, the information presented earlier in the chapters devoted to purposes and uses will help the state with this chapter, but the state will need to argue how the validity evaluation is going to be tied to specific uses and purposes.

Guiding Questions:

1. What are the highest priority consequential, construct, and content questions to pursue for current studies (within the next 2 years)?
2. What are the consequential, construct, and content questions that should be addressed in future studies (3-5 years out)?

Notes

CHAPTER 5: EMPIRICAL EVIDENCE

This chapter is organized in five sections according to the five categories of evidence articulated in the joint AERA, APA, and NCME Standards for Educational and Psychological Testing. In each of the sections of this chapter, the state should summarize the evidence related to the particular category and indicate its plan for subsequent studies.

Content-Related Evidence

The evidence presented in this section will be related to the alignment evidence presented in the first volume and much of the information presented here should be summarized from the alignment evidence.

Rationale for this Content

For this group of students, it can be argued that other than the consequences of the system, the evidence supporting the match between the content of the test items/tasks and the intended targets (content standards) is the most critical piece of the validity evaluation.

Data Sources

- Content standards
- Assessment targets (expanded/extended standards)
- Assessment items/tasks/procedures
- Test blueprints and specifications
- Alignment results
- Notes from development and review meetings

Guiding Questions:

1. What is being aligned with what? For example, are test items or tasks being matched to content standards or, if students complete fairly unique items, are finer grain indicators of the extended standards being aligned to content standards?
2. Is the distribution of items/tasks/procedures in terms of range of representation and depth of knowledge as intended in the design documents?
3. Does the operational balance of representation across the standards reflect the intentions in the design document?
4. Are the items accessible to all students for whom the test is intended?
5. Has an alignment study been conducted for the AA-AAS?
    a. If yes, what do the results indicate?
    b. If no, are there plans for an alignment study?

Notes

Internal Structure

Rationale for this Content
Evidence regarding the internal structure of the test—that is, the extent to which the items and tasks are appropriately representing the intended domain and not other domains—is an important source of information regarding the construct validity of the assessment.

Data Sources
- Reliability/generalizability analyses
- Any dimensionality analyses
- Item correlation matrices
- Description of the intended construct(s)
- Score reports
- Standard setting results
- External information such as classroom work, other test scores, and/or data (interviews, surveys) from educators
- Understanding and impression of standard setting results from policy makers, parents, and other stakeholders

Guiding Questions:
1. What were the results of the dimensionality analyses? Do they support the notion that the test items/tasks/procedural responses are tapping a single domain or does the test represent multiple domains?
2. What were the results of the reliability/generalizability analyses? Do the analyses support the interpretation that there is a single unidimensional construct being assessed? If multiple constructs/sub-domains are being assessed, do the reliability results this interpretation?
3. If subscores are reported, does the item correlation matrix support the intention that items/tasks representing a subscore correlate more highly with other items representing the subscore than they do with items representing other subscores?
4. If there is flexibility in the presentation of specific items to students, are judgmental processes used to determine the relationship of the items to one another such that one could determine that the items appear to be tapping similar domains? If so, what types of judgmental methods are applied?
5. Are the scoring and reporting structures consistent with the sub-domain of the academic content standards (Peer Review Notes, p. 13; Standards, for Education and Psychological Testing (AREA/APA/NCME, 1999)?
6. Do educators and other stakeholders interpret the resulting cut scores in the way that was intended?
7. Are the results of the standard setting confirmed by other data?
8. Are the results of the cut scores supported by the impressions of educators and other stakeholders?
9. Does student performance in relationship to the cut scores lead to the intended actions and those that are justified educationally?

Notes

Response Processes

Rationale for this Content

Evidence regarding the internal structure of the test—that is, the extent to which the items and tasks are appropriately representing the intended domain and not other domains—is an important source of information regarding the construct validity of the assessment. For example, if the test is supposed to assess mathematical reasoning, it is important to collect evidence that students are in fact reasoning and not simply applying algorithms (AERA, et. al, 1999).

Data Sources

- Observation data of students participating in assessment
- Results from any "think aloud" studies if applicable (not always possible with this population)
- Documentation of scorer training and scorer quality control processes
- Description of the intended construct(s)
- Score reports

Guiding Questions:

1. Is there evidence that students are responding the prompts/tasks as intended by the item developers?
2. To what degree are students' responses affected by factors other than what the developers intended? If so, are these factors irrelevant to the intended construct?
3. Have the scoring rubrics been developed in such way as to focus on the key features of the construct and not on irrelevant sources?
4. Is there evidence that the scorers are attending to the features of the task that the item developers intended?

Notes:

Relationship to Other Variables

Rationale for this Content
Other variables may provide evidence of student learning of the intended construct or of what scores on the alternate assessment are intended to predict (AERA, et al., 1999). In regular assessments, this is a very important source of validity evidence for the test interpretations, but for AA-AAS we must approach this category of evidence a bit more cautiously because of the challenges these students face with generalizing their learning to other settings.

Data Sources:
- AA-AAS scores
- Other data about student achievement such as classroom assessments, other district assessments, IEP academic goal results

Guiding Questions:
1. How do the patterns in the AA-AAS scores at the school (if large enough), district, and state levels correlate with patterns of scores for other data that are related to the same general construct?
2. If these correlations or patterns demonstrate a weaker than expected relationship, can this be explained logically?

Notes:

Consequential Evidence

Rationale for this Content
We started this project with the belief that consequential evidence is arguably the most important validity evidence for alternate assessments on alternate achievement standards. We still think this is true especially in the case of large-scale assessment and accountability systems. Further, it is critically important to develop a plan and begin data collection early in a program to make the necessary judgments about the consequences of the system.

Data Sources:
- Surveys and interviews of teachers, administrators (e.g., special education directors), IEP team members, parents, and other stakeholders (e.g., advisory groups)
- Enrollment patterns
- Transition documents
- Compliance monitoring documents
- Test score trends
- Teacher retention data
- Other data sources related to specific consequential research questions

Guiding Questions (these are samples depending on your specific focus):
1. Has the state ascertained whether the assessment system produces intended and unintended consequences? (Peer Review Notes, p. 13; Standards, for Education and Psychological Testing (AREA/APA/NCME, 1999)?
2. What have been the positive and unintended negative effects on students learning opportunities as result of the AA-AAS?
    a. How are test scores used by teachers to adjust instruction for students?
3. What have been the positive and unintended negative effects on teacher recruitment, retention, and professional growth?
4. What have been the positive and unintended negative effects on school and district special education (severe disabilities in particular) programs?
    a. How are test scores used by administrators to adjust programs for students?
    b. Are test scores used to communicate to parents about students' strengths and weaknesses and are the results interpreted appropriately?
    c. Do the test results contribute information that can be used appropriately in the accountability system?

Notes:

CHAPTER 6: THE VALIDITY EVALUATION

This chapter is designed to tie together all the evidence presented thus far and provide an evaluative judgment about the validity of the AA-AAS program.

Revisiting the Validity Evaluation Questions

Rationale for this Content:
This is where the state has the opportunity to remind the reader of the specific evaluation questions and indicate which ones will be discussed in this concluding chapter.

Data Sources:
- Initial evaluation questions

Guiding Questions:
1. What were the prioritized evaluation questions guiding this evaluation?
2. Which evaluation questions will be reported on in this current summary?

Notes:

Logical/Theoretical Relationships Among the Content, Students, Learning, and Assessment: Revisiting the Assessment Triangle

Rationale for this Content:
This is where the state explains how the three vertices of the assessment triangle are connected in its specific system. In other words, why would we expect the scores on the AA-AAS to be related to the particular group of students tested, their instruction, the content on which they are tested, and how they acquire proficiency in the domain?

Data Sources:

- This is not an empirical chapter; rather this chapter is based on the various sources of information presented in the first few chapters of this document as well as key empirical information from Volume I.

Guiding Questions:

1. In the model of learning and assessment guiding the development of your state's AA-AAS, what is the expected (theoretical) relationship among the various components of the system?
2. How should high quality instruction relate to scores on the AA-AAS?
3. How should the scores on the AA-AAS change as students develop increased "proficiency" in the domain?
4. How should your model of domain proficiency for students, the specific AA-AAS test design, and the way that scores are interpreted lead to valid inferences about what students know and are able to do?
5. What is the connection between how students are hypothesized to develop competence in the domain and the structure of the tests and tasks?
6. Is there a logical connection among specific curricular approaches and performance on the items/tasks?
7. What is the relationship between the items/tasks and the intended domain?

Notes:

Synthesizing and Weighing the Various Sources of Evidence

Rationale for this Content:
This is where the state or the evaluator pulls all the evidence from this volume and Volume I together, describes the story that is being told by the data, and puts forth evaluation judgments and recommendations.

Data Sources:

- This is not an empirical chapter; rather this chapter is based on the various sources of information presented in Chapter 5 of this document as well as key empirical information from Volume I.

Guiding Questions:

1. What are the arguments for the validity of the system?
2. What are the arguments against the validity of the system?
3. What are the overall judgments regarding the validity of the AA-AAS system?
4. What are the recommendations to strengthen the system?

Notes: