VALIDITY FRAMEWORK FOR EVALUATING THE TECHNICAL QUALITY OF ALTERNATE ASSESSMENTS BASED ON ALTERNATE ACHIEVEMENT STANDARDS

Scott F. Marion, National Center for the Improvement of Educational Assessment

Jim Pellegrino, University of Illinois, Chicago

NCME Invited Presentation

April 15, 2009

Complex issues have emerged as states have developed alternate assessments based on alternate achievement standards (AA-AAS) to ensure that all students have access to the benefits of state assessment and accountability systems. The understanding of technical quality for alternate assessments (Marion & Pellegrino, 2006) has improved with a focus on the interactions among the primary considerations in assessment design as articulated in *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001): 1) the population and their competence in academic domains, 2) the appropriateness of the observation techniques employed to assess these students, and 3) the inferences about student performance, teaching, and learning that result from the assessment scores. There is no question that technical documentation of alternate assessments has improved during the past several years, and state leaders have conducted many studies that can legitimately be called *validity studies.* However, very few (if any) states have designed their validity studies to collect data related to articulated validity arguments or have synthesized the results from their multiple studies to evaluate the veracity of the validity arguments and judge the validity of the inferences derived from assessment programs. This paper presents an approach for documenting the technical quality of alternate assessments in ways that allow for the evaluation of a validity argument.

Many writers of technical reports for general assessments attempt to align their analyses and results with the *Standards for Educational and Psychological Testing* (American Educational Research Association (AERA), American Psychological Association, & National Council on Measurement in Education, 1999), particularly when there are student or school stakes requiring that the inferences drawn from the assessment be valid, reliable, and fair (AERA et al., 1999). This is an obvious and important first step, but it often is not fully met. AA-AAS approaches have additional technical quality challenges because many traditional measurement methods may require reconceptualization to "fit" the assessment requirements of students with significant cognitive disabilities. For example, leading measurement theorists (e.g., Cronbach, 1971, Messick, 1989), including the authors of the 1985 and 1999 standards for educational measurement (AERA et al., 1985, 1999), established validity as the most important technical criterion for educational assessment. Validity is defined as the "degree to which evidence and theory support the interpretations of the test scores entailed by proposed uses of the test" (AERA et al., 1999, p. 9).

Some researchers have begun defining how to document the validity and reliability of AA-AASs (Garrett, Towles, Kleinert, & Kearns, 2003; Gong & Marion, 2006; Kleinert & Kearns, 1999; Marion & Pellegrino, 2006). Validity often is discussed regarding 1) the content of the assessment—considering the extent to which the content is representative of the established standards (content evidence), 2) the extent to which the construct indicates or captures the attribute to be measured (construct evidence), and 3) the extent to which the assessment correlates to or predicts another desired state (criterion evidence; Mehrens, 1997). Consequential evidence—the degree to which an assessment produces the desired outcomes—is an area of

particular interest to researchers on AA-AAS. For example, exploring how various approaches to alternate assessment increase access to grade-level curriculum and improve learning outcomes would provide powerful validity data. Linn, Baker, and Dunbar (1991) and Shepard (1993), in their interpretations of Messick (1989), were unambiguous in the legitimacy of consequences as a critical area of validity inquiry and evidence.

The challenge, however, has moved from having states and test contractors conduct research/evaluation studies to investigate a particular aspect of a testing program to designing a systematic validity plan for evaluating the efficacy of a comprehensive validity argument. This approach requires synthesizing the various empirical results against the theory of action and validity argument (Kane, 2006). This paper, drawing heavily on Kane (2006), outlines a framework for using a validity argument to collect and evaluate the technical documentation for a state's AA-AAS.

Validity Framework

The purpose of technical documentation should be to provide data to support or refute the validity of the inferences from the alternate assessments at both the student and program levels. Drawing on the work of Cronbach, 1971, Messick, 1989, Shepard, 1993, and Kane, 2006, our proposed evaluation of technical quality is built around a unified conception of validity centered on the inferences related to the construct and includes significant attention to the social consequences of the assessment.

Kane's Argument Based Validity Framework

Kane indicated that validation is made up of two different types of arguments: an interpretative argument and a validity argument. According to Kane (2006), "an *interpretative argument*

specifies the proposed interpretations and uses of test results by laying out the network of inferences and assumptions leading to the observed performances to the conclusions and decisions based on the performances, [while] the *validity argument* provides an evaluation of the interpretative argument" (p. 17). A major advantage of Kane's perspective is that it provides a more pragmatic approach to validation than the construct model. Explicitly specifying the proposed interpretations and uses of the assessment (system), developing a measurement procedure consistent with these proposed uses, and then critically evaluating the plausibility of these inferences and assumptions can be challenging, but it is somewhat more straightforward than evaluating the validity of an assessment under a construct model.

Kane (2006) pushed for the development of the interpretative argument in the assessment design phase. The notion of specifying purposes and uses up front and then designing an assessment to fit these intentions is certainly not a new idea. However, the concept of designing a fully coherent system built on a sound theoretical model of learning and use began receiving more attention with the publication of *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001). Most assessments do not start with explicit attention to validity in the design phase (e.g., Mislevy, 2006), therefore many evaluators working with states are put in the position of retrofitting validity arguments to existing systems.

The Interpretative Argument

The interpretative argument is essentially a mini-theory in that the interpretative argument provides a framework for interpretation and use of test scores. Like theory, the interpretative argument guides the data collection and methods for conducting the validity analyses. Theories are falsifiable, and making the connection between the interpretative argument and "mini-theory"

is intended to demonstrate that validation is not a confirmationist exercise. Kane (2006) noted two stages of the interpretative argument. The *development stage* focuses on the development of measurement tools and procedures and the corresponding interpretative argument. Kane (2006) suggested that it is appropriate to have a confirmationist bias in this stage because the developers (state and contractors) are trying to make the program as good as possible. During the *appraisal stage,* the focus should be on critical evaluation of the interpretative argument. This should be a more neutral and "arms-length" standpoint to provide a more convincing evaluation of the proposed interpretations and uses. The evaluator should have moved from the confirmationist bias of the development stage to a falsification stance in the appraisal stage, because as Cronbach (1989) noted, "falsification, obviously, is something we prefer to do unto the constructions of others" (p. 153).

Kane (2006) noted the importance of being able to put forth a clear and coherent interpretative argument:

> *Difficulty in specifying an interpretative argument…may indicate a fundamental problem. If it is not possible to come up with a test plan and plausible rationale for a proposed interpretation and use, it is not likely that this interpretation and use will be considered valid (p. 26).*

It is helpful to think of the interpretative argument as a series of *if–then* statements (e.g., if the student performs the task in a certain way, then the observed score should have a certain value). Kane also offered the following criteria for evaluating interpretative arguments: clarity, coherence, and plausibility. A clear argument is one that is clearly stated as a framework for validation with the inferences and that is detailed enough to make proposed claims explicit. A

coherent argument is one that logically ties together or relates the network of inferences from the assessment design, assessment scores, and inferences to the decisions. Finally, the criterion of plausibility focuses on the assumptions underlying the assessment and inferences in terms of all the evidence for and against them.

One of the most effective challenges to interpretative arguments (or scientific theories) is to propose and substantiate an alternative argument that is more plausible. With AA-AAS, it is important to seriously consider challenging ourselves with competing alternative explanations for test scores. For example, evaluators might want to propose (and confirm) that higher scores on a state's AA-AAS reflect greater learning of the content frameworks. However, the evaluators must consider plausible alternative hypotheses: higher scores on a state's AA-AAS might reflect higher levels of student functioning, or the higher scores might reflect greater understanding by teachers about how to gather evidence or administer the test.

Test validation is the process of offering assertions (propositions) about a test or a testing program and then collecting data and posing logical arguments to refute those assertions. In essence, validation requires continually challenging the supportability of the claims put forth about a testing program. If the assertions cannot be refuted, they can be considered tentatively supported, and that is the best that can be done.

Values and Consequences

Kane and others suggested that evaluators must attend to values and consequences when evaluating decision procedures (such as when a testing program is used as a policy instrument, as is the case with essentially all state tests). When conducting such a validity evaluation, the values

inherent in the testing program must be made explicit, and the consequences of the decisions made in response to the test scores must be evaluated. Many have argued (e.g., Marion & Pellegrino, 2006; Shepard, 1997) that consequences must occupy a prominent role in any validity evaluation, but consequences are especially important when the validity of an AA-AAS is being evaluated.

Technical Documentation

Where does all of this leave us in terms of creating prototypical technical documentation? We take the perspective that validity is central to the documentation of alternate assessments, but this does not mean that the technical documents deal only with the validity argument (although we argue that all of the other aspects typically found in technical manuals contribute to the validity argument). The documentation will require sections, among others, describing the students taking the test, item/task development, alignment, administration, and scoring.

Our original intent was to create a single prototypical manual that, unlike typical technical manuals, would weave the various chapters together through the use of the validity argument. In most cases, the different chapters of technical manuals are written to stand alone. As Haertel (1999) reminded us, the individual pieces of evidence (presented in separate chapters) do not make the assessment system valid or not, it is only by weaving these pieces of evidence together into a coherent argument can we judge the validity of the assessment program. Upon further reflection, we have reconsidered our view of a single technical manual as a set of related technical documents. These technical documents will be structured to facilitate the types of validity evaluations we describe.

We have reconceptualized this single technical manual as a set of at least four documents: a somewhat familiar "nuts and bolts" volume that includes chapters that psychometricians are used to seeing in technical manuals; a validity evaluation of the sort that we have been discussing thus far; a stakeholder summary that is drawn from the validity evaluation and the nuts and bolts volume; and a transition document that contains extensive procedural details to aid when programs transition from one contractor to another and/or there is turnover in state DOE personnel. There are clearly different audiences for each of the four volumes and authorship differs across the documents as well. We argue that the complete set is important, but because of space limitations and centrality to the assessment triangle, this paper deals only with the nuts and bolts and the validity evaluation volumes.

Some might argue that it is unrealistic to expect states and contractors to produce this type and amount of technical documentation, especially considering the extensive focus on validity evidence. On the other hand, many would push for all technical manuals to include such analyses. We do not see this as an all or nothing proposition or as something that is a one-time proposition. Given current capacity for web-based publishing, these could be dynamic volumes that evolve as new evidence is brought to bear on the argument. For example, the initial validity evaluation would probably not be more than the state's framework for approaching a validity evaluation, the prioritized validity questions, and initial evidence as gleaned from the nuts and bolts volume. The results of new studies, as they are completed, would be incorporated into the ongoing evaluation and judgments about the validity of the program. However, this does not mean that the validity evaluation should go on endlessly without ever reaching conclusions. States will have to set their own timetables for bringing closure to certain stages of the validity

work while continuing to plan and conduct additional studies. It is critical to keep in mind that unless there is a systematic plan to collect and organize validity evidence, it is unlikely that the state (or its agent) would be able to evaluate the validity of the assessment system. The validity framework for these technical documents is intended to provide this organizing structure.

The proposed table of contents for the "nuts and bolts" volume is found in Appendix A. Most of the chapters listed should appear quite familiar, but applying these to document the technical quality of alternate assessments raises considerable challenges. We address these challenges elsewhere (see Marion & Pellegrino, 2006). The first three chapters focusing on the overview of the system, a description of the content, and a description of the students participating in the alternate assessment are usually not found in manuals describing the technical quality of regular assessments. This volume is typically authored by a test contractor with some sections written by state DOE personnel. The limited audience for this volume usually includes state department personnel, district assessment and special education directors, state technical advisors, and others with technical backgrounds (e.g., university faculty and graduate students). This volume also serves as an empirical foundation for many of the analyses presented in the validity evaluation. In order for this document to serve this purpose, it will require this technical manual to move beyond the typical "data dumps" found in many technical manuals now. We suggest that writers of such documents address the following questions in each chapter or section of the technical documentation:

- What was done?
- Why was this done?
- What were the results (i.e., what did you find)?

- What do these results mean (e.g., how should a reliability coefficient of .85 be interpreted)?
- What will change as a result of what was found?

Addressing these questions will enable this volume to be more interpretable and will also help those charged with writing the validity evaluation to have an inferential base from which to begin. While we originally intended to frame the entire technical manual through a validity lens, issues of authorship and audience were practical constraints that needed to be considered. For example, there are obvious conflicts of interest if a test contractor is to be the validity evaluator for the system. Therefore, it made sense to separate these tasks and volumes. This separation of tasks and volumes, however, does not mean that validity is ignored in the nuts and bolts volume. Rather, adopting a validity frame for the entire set of technical documentation will help ensure that all of the volumes are created with the intent of producing a validity argument. Even though the volumes have been separated, it is crucial that the nuts and bolts volume be written with a focus on contributing empirical information for the validity argument. Unless the writers of the nuts and bolts volume anticipate the use of this information as part of the validity evaluation, they run the risk of having this volume be seen as just another data dump.

The validity evaluation volume is organized to logically synthesize empirical information from the nuts and bolts volume and data from studies focused on specific aspects of the validity evaluation. A proposed table of contents for a well-developed validity evaluation is found in Appendix B. This volume is organized by bridging the framework laid out thus far with the well-established approach to collecting validity evidence outlined in the *Standards for Educational and Psychological Measurement* (AERA, APA, NCME, 1999).

The *Standards* and Technical Documentation of Alternate Assessments

The *Standards* framework for validity evaluations categorizes the sources of evidence as follows:

- Test content
- Response processes
- Internal structure
- Relations to other variables
- Consequences of testing (AERA, et al., 1999, pp.11-17).

Sources of validity evidence for alternate assessments can be derived from each of the categories listed in the *Standards*, but the nature of both the students and the assessments employed creates documentation difficulties. For example, studying response processes through typical methods such as "think-aloud" studies with significantly disabled students—many of whom face noteworthy communication challenges—requires considerably more inference on the part of the researcher when determining students' attention and/or understanding and may have to be gleaned from eye gazes as opposed to the actual "aloud" part of the think-aloud study. Even categories that appear straightforward such as evidence based on test content still present challenges to the evaluator. Content-related evidence can be drawn from alignment studies, examination of test specifications, and task guidelines for how the content represents the intended construct. However, evidence related to construct-irrelevance and construct-under-representation—important sources as noted in the *Standards* and Messick (1989) before the *Standards*—with some alternate assessment designs can be illusive. Several forms of alternate assessments rely on having students complete only one or very few tasks per content area. This would appear to invalidate most of these intended inferences unless the domain was defined very

narrowly, which is rarely the case. We do not point out these examples to argue against collecting validity evidence for alternate assessments, rather our contention is that evaluators should use this framework, but not apply it blindly as noted in the *Standards* document itself.

> *Several cautions are important to avoid misinterpreting the Standards*
>
> *1) Evaluating the acceptability of a test or test application does not rest on the literal satisfaction of every standard in this document, and acceptability cannot be determined by using a checklist...evaluating acceptability involves (a) professional judgment that is based on a knowledge of behavioral science, psychometrics, and the community standards in the professional field to which the tests apply; (b) the degree to which the intent of the standard has been satisfied by the test developer and user; (c) the alternatives that are readily available; and (d) research and experiential evidence regarding feasibility of meeting the standard…*
>
> *4) These standards are concerned with a field that is evolving. Consequently, there is a continuing need to monitor changes in the field and to revise this document as knowledge develops.*
>
> *5) Prescription of the use of specific technical methods is not the intent of the Standards. For example, where specific statistical reporting requirements are mentioned the phrase "or generally accepted equivalent" always should be understood…(AERA, et al., p. 4).*

Point #4 in the quote from the introduction to the *Standards* above is especially important when considering the technical evaluation of alternate assessments. While the *Standards* were published in 1999, the drafts were being produced between 1995 and 1998. The requirements

for alternate assessments, on the other hand, were not encoded in law until the Improving America's Schools Act (IASA) of 1994 and the Individuals with Disabilities Education Act (IDEA) of 1997 and including these results in state accountability systems was not required until the No Child Left Behind Act (NCLB) of 2001 was enacted. This little history lesson is presented to make clear that much of our thinking about alternate assessments on alternate achievement standards has occurred after the *Standards* were written. To drive this point home, the following quote is the only mention of alternate assessments in the Standards.

> ***Using Substitute Tests or Alternate Assessments***
>
> *One additional modification is to replace a test standardized on the general population with a test or alternate assessment that has been specially designed for individuals with disabilities. More valid results may be obtained through the use of a test specifically designed for use with individuals with disabilities. Although a substitute test may represent a desirable accommodation solution, it may be difficult to find an adequate replacement that measures the same construct with comparable technical quality, and for which scores can be placed on the same scale as the original test (AERA, et al., 1999, p. 104).*

The mention of the "same construct" and the "same scale" makes clear how much our understanding has evolved since the *Standards* were written. Very few states are attempting to place the alternate assessment scores on the same scale as the regular assessments and essentially no one would argue that the same construct is being measured.

Revisiting the Assessment Triangle and AA-AAS Technical Documentation

We have argued for organizing validity and technical evaluations around the assessment triangle (Pellegrino et al., 2001) and have presented an approach for organizing the technical documentation of an alternate assessment system. The assessment triangle has proven to be a useful heuristic for these projects to help us focus on the construct validity of the alternate assessments. One shortcoming of the using the triangle is that the inclusion of consequences in the evaluation is not readily apparent. Therefore, in our focus on validity, we had to make sure we incorporated consequential evidence into our validity evaluation, especially given the importance of this type of evidence for the evaluation of the validity of alternate assessment systems.

One other aspect of the assessment triangle that was a key element of the argument in *KWSK* is that each of the three vertices interacts with the other two and it is the reciprocal relationships between and among them that gives meaning to the process of assessment design and deployment for any specific population of students for a given domain of academic performance. For example, how one thinks about the important aspects of knowing something like multicolumn subtraction and how that knowledge is acquired over time via instruction should shape the types of tasks and problems that are presented to students to tap into that understanding and the way in which the resultant data are scored and interpreted should in some way reflect the underlying model of competence. Designing an assessment using the framework of the triangle means multiple iterations between and among the elements at each of the vertices to make sure they are coherent and coordinated. The same argument applies to use of the triangle in constructing a validity argument for any assessment system that has been design for use with a

special population. For example, what gets reported for each of the elements in each of the boxes associated with each vertex of the triangle needs to be consistent with and shaped by what gets reported for elements at the other vertices. Doing otherwise is to persist in a technical reporting process in which there is little coherence among the parts and the overall validity argument remains tacit at best.

Synthesis and Evaluation

Haertel (1999) reinforced the notion that individual pieces of evidence (typically presented in separate chapters of technical documents) do not make an assessment system valid or not; it is only by synthesizing this evidence to evaluate the interpretative argument that the validity of the assessment program can be judged. As Kane (2006) indicated, the evaluative argument provides the structure for evaluating the merits of the interpretative argument. Various types of empirical evidence and logical argument must be integrated and synthesized into an evaluative judgment; this process can be a challenging intellectual activity. In state assessment programs, when new and varied information comes in at sometimes predictable intervals, the challenge is exacerbated. With alternate assessment programs, not only is new evidence being collected along the way, but actual understanding of alternate assessments and the students they serve evolves much more rapidly than in many other programs.

Dynamic Evaluation

In almost all studies that evaluate the validity of state assessment systems, the studies are completed across a long time span. Evaluators rarely have all the evidence in front of them to make conclusive judgments. Therefore, evaluators must engage in ongoing, dynamic evaluations as new evidence is produced. Working in this fashion requires, even more so than in more

predictable evaluations, that each proposition be written to allow judgment of whether the evidence supports a particular claim. As discussed above, this always means exploring the efficacy of alternate hypotheses. However, in the context of states' large assessment systems, evaluators do not have the luxury of concluding, "The system is not working; let's start over." Rather, in such instances, when the evidence does not support the claims and intended inferences, state leaders and test developers must act as if the dynamic results were from a formative evaluation, and they must search for ways to improve the system. Of course, the evidence might be so overwhelmingly stacked against the intended claims—and this has happened in some states—that the state leaders are left only with the option of starting over.

Conclusions and Continued Challenges

This approach to technical documentation is based on the approach to validity articulated in *Knowing What Students Know* (Pellegrino et al., 2001) and on the specific standards put forth in the *Standards for Educational and Psychological Testing* (AERA, et al., 1999). As such, this approach to technical documentation is comprehensive enough to serve as a guide to the technical documentation of the regular assessment system and we urge states to consider their regular and alternate systems comprehensively. This framework goes beyond what is required to meet the USED standard and assessment peer review approval system. That is our intent. There is no question that the USED guidance is important, but we argue that this validity-based approach is more appropriate for long-term technical documentation.

By now, many readers who have been involved in technical documentation for state assessment systems are probably wondering, "who is going to do all this work and who is going to pay for it?" We acknowledge that this approach to technical documentation is more labor and resource

intensive than what has been done in the past. However, that is due, in part, because states and contractors have not devoted adequate resources—for some legitimate reasons—to technical documentation in the past. This is not an all or nothing situation. As noted earlier, the approach to technical documentation advocated herein can provide a means for conducting evaluations of technical quality in a way that recognizes the practical realities of state assessment systems.

The framework for conducting evaluations of the validity of state alternate assessment systems, based on the assessment triangle, provides a means for systematically prioritizing (e.g., Shepard, 1993) and organizing validity questions and analyses. There is no expectation that the full set of documents be produced each year, but it is crucial that there be a plan for systematic data collection and reporting. This validity-based framework requires such a plan. Further, many of these documents can be published on the web, which would allow regular updating as new evidence is collected and analyzed.

This approach to technical documentation was not written from the perspective of any particular form of alternate assessment. Certain sources of evidence and methods of documenting technical quality will differ by the degree of flexibility/standardization (Gong & Marion, 2006) of the state's alternate assessment approach, but this framework is designed to work with any alternate assessment system. Certain sources of evidence and guiding questions apply to particular forms of AA-AAS better than others, but this should not be seen as privileging one form of assessment over another. The project makes no value judgments about specific forms of assessment. Our only position is that every assessment system should have technical documentation to defend the score inferences resulting from its approach.

We have adopted this neutral position regarding appropriate assessment designs because we do not yet have enough validity information to fairly evaluate the quality of inferences derived from one AA-AAS approach compared with another. However, this framework for technical documentation, based on the psychology of achievement testing as articulated in *Knowing What Students Know* (Pellegrino et al., 2001), is designed to allow states to eventually reconsider their alternate assessment system designs if they find that the inferences drawn from the assessment scores are not supported by the evidence. We hope these states would consider the available evidence in light of the assessment triangle framework to redesign their systems (if necessary) to better align their assessment with their understanding of the students and how they develop domain competence as well as the techniques used to interpret the assessment observations.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: Authors.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: Authors.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.

Cronbach, L.J. (1988). Five perspectives on validity argument. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 3–17). Mahwah, NJ: Lawrence Erlbaum Associates.

Cronbach, L.J. (1989). Construct validation after thirty years. In R.L. Linn (ed.). *Intelligence: Measurement, theory, and public policy* (pp. 147-171). Urbana: University of Illinois Press.

Cronbach, L.J. & Meehl, P.E.. (1955). Construct validity in psychological tests. *Pyschological Bulletin, 52,* 281-302.

Garrett, B., Towles, E., Kleinert, H., & Kearns. J. (2003).. Portfolios in large-scale alternate assessment systems: Frameworks for reliability, *Assessment for Effective Intervention, 28, 2,* 17-27

Gong, B., & Marion, S.F. (2006, June). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Rep. No. 60). Retrieved from http://education.umn.edu/nceo/OnlinePubs/Synthesis60.html August 1, 2006.

Haertel, E.H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice, 18*(4), 5–9.

Kane, M.T. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). New York: American Council on Education/Macmillan.

Kearns, J., Towles-Reeves, E., Kleinert, H., & Kleinert, J. (2006). *Learning characteristics inventory (LCI) report.* Lexington: National Alternate Assessment Center, Human Development Institute, University of Kentucky, Lexington.

Kleinert, H., Browder, D., & Towles-Reeves, E. (2005). *The assessment triangle and students with significant cognitive disabilities: Models of student cognition.* Lexington: National Alternate Assessment Center, Human Development Institute, University of Kentucky, Lexington.

Kleinert, H. L. & Kearns, J.F. (1999). A validation study of the performance indicators and learner outcomes of Kentucky's alternate assessment for students with significant disabilities. *The Journal of the Association for Persons with Severe Handicaps, 24, 2*, 100-110.

Lane, S., & Stone, A. (2002). Strategies for examining the consequences of assessment and accountability programs. *Educational Measurement: Issues and Practice, 21*(2), 23–30.

Linn, R.L., Baker, E.L., & Dunbar, S.B. (1991). Complex performance-based assessment: Expectations and validation criteria. *Educational Researcher, 20*(8), 15–21.

Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3* (Monograph Supplement 9), 635-694.

Marion, S.F., & Pellegrino, J.W. (2006). A validity framework for evaluating the technical quality of alternate assessments. *Educational Measurement: Issues and Practice, 25*(4), 47–57.

Mehrens, W. A. (1997). The consequences of consequential validity. *Educational Measurement: Issues and Practices, 16, 2*, 16-18.

Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education, Macmillan Publishing.

Messick, S. (1995). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23.

Mislevy, R. (1996). Test theory reconceived. *Journal of Educational Measurement, 33*(4), 379–416.

No Child Left Behind Act of 2001, PL 107-110, 115 Stat. 1425, 20 U.S.C. §§ 6301 *et seq*.

Perie, M., & Marion, S. (2008). *Developing a validity argument for a state alternate assessment (AA-AAS) system: A guide for states.* Retrieved from http://www.naacpartners.org/projects/valdityGSEG/expertPanel.aspx July 15, 2008.

Pellegrino, J.W., Chudowsky, N.J., & Glaser, R. (Eds.) (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy of Sciences.

Ryan, K. (2002). Assessment validation in the context of high stakes assessments. *Educational Measurement: Issues and Practice, 21*(1), 7–15.

Shepard, L.A. (1993). Evaluating test validity. *Review of Research in Education, 19*(1), 405–450.

Shepard, L.A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–24.

APPENDIX A: THE "NUTS AND BOLTS" OF AA-AAS TECHNICAL DOCUMENTATION

Volume I: The "Nuts and Bolts"
Author: Test contractor or other test developer
Audience: State DOE, district assessment and special education directors, state TAC members, and others with some technical backgrounds. This will also serve as document for legal defensibility.
Notes: The overview of the system and several of the first few chapters will be replicated in the both this volume and the validity evaluation volume.

I. Overview of the Assessment System
- A. Introduction to the technical manual
- B. Statement of core beliefs and guiding philosophy
- C. Purposes of the assessment system
  1. Governing statutes and authority
  2. Stakeholder purposes and intentions
- D. Uses of the assessment information
  1. Inferential target(s)—school, student?
  2. Stakes associated with assessment results
  3. Uses as part of the state accountability system

II. Who are the students?
- A. Quantitative and qualitative description of the students participating in the AA-AAS, including test participation rates

III. What is the content?
- A. Description of ELA content and performance (achievement) expectations
- B. Description of mathematics content and performance expectations
- C. Alignment of AA-AAS content with general content and achievement expectations.
- D. Description of linkage to different content across grades that support individual growth

IV. Test Development
- A. Types of approach—e.g., portfolio, performance task, observation checklist
- B. Creation of test specifications aligned to grade level content
- C. Item/task development process
  1. Types of items/tasks
  2. Item review processes
     a. Use of stakeholders
     b. Role of committees in test development
        i. Content committees
        ii. Bias committees
  3. Item tryouts/field tests
- D. Universal Design

V. Item Analysis and DIF/bias
- A. Traditional item analyses—e.g., difficulty and discrimination
- B. Other forms of items analyses—we're going to have to develop these

C. Examining bias (in addition to the usual groups, by disability)

VI. Alignment

A. Alignment to the grade level content standards
B. Alignment to the alternate achievement standards

VII. Administration & Training

A. Administration procedures and guidelines
    1. Timing
    2. Independence/assistance issues
B. Role of the IEP process and team regarding administration
C. Professional development and training programs
    1. Attendance
    2. Program content
    3. Evaluation of training quality
D. Monitoring and quality control of administration procedures

VIII. Scoring

A. Scoring rules and criteria
B. Scoring process
    1. Distributed/centralized scoring
    2. Use of anchor papers/counting points
C. Selection and training of scorers
D. Scoring quality control
    1. Monitoring of scorers
    2. Scoring accuracy
    3. Scoring consistency

IX. Characterizing errors associated with test scores

A. Uses of test scores and implications for consideration of error
B. Levels of analysis
C. Decision consistency and accuracy
    1. Methodological approaches
D. Traditional reliability analyses
    1. Methodological approaches
E. Generalizability analyses
    1. Methodological approaches

X. Scaling and Equating (year-to-year comparability)

A. Choice of scale and rationale for choice
B. Comparability of scores across years
    1. Equating methods
    2. Other methods for establishing comparability
C. Linkage across grades (measuring growth)

XI. Standard Setting

A. Standard setting methodology, including rationale for selecting this method
    1. Performance descriptors
    2. Panelists
    3. Protocol
B. Standard setting results
    1. Unadjusted results

2. Adjusted/smoothed results
3. Policy decisions
4. Coherence across grade levels
5. Coherence across subject areas

XII. Reporting

A. Report Design Process
   1. Interviews, focus groups, etc.
B. Adherence to Joint Standards
C. Reports for Parents and Students
D. Reports for Classes (teachers), Schools and District
E. Types of scores reported
   1. Summary scores
      a. Students
      b. Schools
   2. Subscores
      a. Students
      b. Schools

APPENDIX B: THE VALIDITY EVALUATION

Volume II: The Validity Evaluation
Author: Independent contractor with considerable input from state DOE
Audience: State policy makers, state DOE, district assessment and special education directors, state TAC members, special education teachers, and other key stakeholders. This will also contribute to the legal defensibility of the system.
Notes: The overview of the system and several of the first few chapters will be replicated in the both this volume and the Nuts and Bolts volume.

I. Overview of the Assessment System
- A. Introduction to the technical manual
- B. Statement of core beliefs and guiding philosophy
- C. Purposes of the assessment system
  - i. What makes it a system (articulation across grade levels/spans)
  - ii. Governing statutes and authority
- D. Uses of the assessment information
  - i. Inferential target(s)—school, student?
  - ii. Stakes associated with assessment results

II. Who are the students?
- A. Quantitative and qualitative description of the students participating in the AA-AAS, including test participation rates
- B. How do they learn? Models of expected domain proficiency.
  - iii. ELA
  - iv. Mathematics

III. What is the content?
- A. Description of ELA content and performance (achievement expectations)
- B. Description of mathematics content and performance expectations
- C. Alignment of AA-AAS content with general content and achievement expectations.
- D. Description of linkage to different content across grades that support individual growth

IV. Introduction of the Validity Framework and Argument
- A. Theoretical approach to validity
  - i. Evaluation in the context of the purposes and uses
- B. Logical connections among the content, learning models, and the assessments
- C. Prioritized validity evaluation questions for this state
  - i. Plans for near-term studies
  - ii. Plans for future studies

V. Empirical Evidence (all should be drawn from "Nuts and Bolts" volume)
- A. Content-related evidence
- B. Internal Structure
- C. Response Processes
- D. Relationship to other variables
- E. Consequential evidence
  - i. Effects on students learning opportunities
    - ➢ Positive and negative

ii. Effects on teacher professional growth
    - Positive and negative
iii. Programmatic effects on schools and districts
    - Positive and negative

VI. The Validity Evaluation

A. Revisiting the validity evaluation questions
B. Logical/theoretical relationships among the content, students, learning, and assessment—revisiting the assessment triangle
C. Synthesizing and weighing the various sources of evidence
    i. Arguments for the validity of the system
    ii. Arguments against the validity of the system
D. An overall judgment of the validity of the AA-AAS system