
Considerations for the Alternate Assessment based on Modified Achievement Standards

Section III

Technical Considerations and Practical Applications

**Chapter 8: Comparability Issues in the Alternate Assessment
Based on Modified Achievement Standards for
Students with Disabilities**

**Chapter 9: Constructing a Validity Argument for Alternate
Assessments Based on Modified Achievement
Standards (AA-MAS)**

Chapter 10: Operational and Accountability Issues

The contents of this publication were developed under cooperative agreement S283B050019 with the U. S. Department of Education. However, the contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

For the full version of this white paper, visit:

http://nycomprehensivecenter.org/initiatives/inits_sp_assessment



University of the
State of New York
State Education
Department

SECTION III

TECHNICAL CONSIDERATIONS AND PRACTICAL APPLICATIONS

This final section incorporates the overarching themes of comparability and validity of these assessments and then focuses on how the AA-MAS will fit into a state accountability system. Its original intent was to provide information on examining the technical adequacy of these assessments as a logical follower of the section on assessment design. However, it soon became clear that specific issues needed to be addressed, far beyond most technical considerations of item analysis, reliability and equating. Thus, a previous chapter (Welch & Dunbar, Chapter 6, this volume) began the discussion of technical adequacy, focusing on item analyses and psychometric characteristics of the test. This section, then, focuses on very specific questions regarding the technical quality of this assessment, not as a standalone assessment, but as it fits into a larger assessment program.

Chapter 8, by Jamal Abedi discusses issues related to the comparability of the AA-MAS from the perspective of ensuring students who take this assessment should have the same opportunities for success and inclusion as students who take the general assessment. Several components of comparability are examined, including content and construct, psychometrics, scale and score, linguistic structure, basic text features, depth of knowledge and accommodations used for students with disabilities based on their IEP.

Chapter 9, by Scott Marion discusses the importance of developing a validity argument for the implementation and use of the AA-MAS. He emphasizes the importance of articulating the theory of action particularly in light of the uncertain conceptual framework supporting this AA-MAS initiative. He then describes methods for evaluating the argument to provide information about how to improve the program and how to determine the value of AA-MAS in terms of the instructional and social benefits given the costs.

Finally, in Chapter 10 by Chris Domaleski, the focus turns to the practical application of these ideas in a state assessment system. He describes issues of fitting this assessment into a pre-existing state assessment and accountability system. He considers the current state context in reviewing operation considerations and discusses ways to estimate reliability, produce informative score reports, and consider options related to diploma eligibility.

The authors in this section received advice and guiding comments from the expert panel members who reviewed these chapters. In particular, comments from Katherine Ryan, Phoebe Winter, Brian Gong, Suzanne Lane, and Howard Everson were valuable and served to inform the final drafts of the chapters.

CHAPTER 8

COMPARABILITY ISSUES IN THE ALTERNATE ASSESSMENT BASED ON MODIFIED ACHIEVEMENT STANDARDS FOR STUDENTS WITH DISABILITIES

Jamal Abedi

The mandate of including students with disabilities in state and national assessments may not produce desirable results if the assessment outcomes for these students are not comparable with the assessment outcomes for mainstream students. Thus, comparability issues for these students must be given careful attention if these students are to be given a fair chance of inclusion in the assessment and accountability system. The principle of comparability and its related issues have long been debated. In this chapter, the concept of comparability is viewed and discussed in broader terms and from different perspectives, including content and construct, psychometrics, scale and score, linguistic structure, basic text features, depth of knowledge, and accommodations used for students with disabilities based on their IEP. It is indicated that comparability is not an “all-or-none” proposition; rather it is a continuum of varying degrees. Recommendations have been provided for the State of New York on how to view and evaluate comparability between alternate assessments based on modified achievement standards (AA-MAS) and general assessments.

Rationale

Recent legislation such as the reauthorization of the Individuals with Disabilities Education Act and the No Child Left Behind Act of 2001 mandates inclusion of students with disabilities in the assessment and accountability systems (Domaleski, Chapter 10, this volume; Gong & Blank, 2002; Lowrey, et al, 2009; Thompson, Lazarus, Clapper & Thurlow, 2006). This mandate is based on the assumption that the same, or at least comparable, assessments are used across groups of students, those with different types of disabilities and those without any apparent disabilities. In the context of assessment, comparability means that the inferences from the scores on one test can be psychometrically related to a score on another “comparable”

test (Marion, 2006). In other words, comparability assumes equivalence between the assessments (Elosua & Lopez-Jhuregui, 2008). While these definitions provide one aspect of comparability, they emphasize the importance of comparability of AA-MAS with the general assessments as the policy of inclusion may not produce valid outcomes if the assessments used for different subgroups of students do not have the same meaning and do not lead to the same interpretation across these subgroups. In this chapter, issues concerning comparability of assessments for students with disabilities taking alternate assessment based on modified achievement standards (AA-MAS) are discussed and methods for examining such comparability are described. The focus on comparability in this chapter centers on the application of AA-MAS assessments for students with disabilities in the State of New York.

The majority of students with disabilities take the general state assessments, with or without accommodations. However, a small group of students with disabilities—who can make significant academic progress but who are not able to achieve grade-level progress—may not be able to show the full range of their knowledge and skills on the general assessments even with accommodations. Therefore, they are offered alternate assessments (Lazarus, Rogers, Cormier & Thurlow, 2008). These alternate assessments have been described as the “ultimate accommodation” for inclusion of students with significant disabilities in the accountability system (Domaleski, Chapter 10, this volume; Roach, 2005).

However, there are major questions and concerns regarding the purpose, design, development, implementation, and interpretation of the outcomes of these assessments. For example, Kettler and Almond (2009) raise many questions regarding these assessments:

“First and foremost, which students should be eligible for an AA-MAS? Second, what are their unique learning characteristics, and how should an assessment be tailored to their needs based on a better understanding of their cognitive processing?” (p. 5)

The authors also raised questions related to item and test development which include:

“(a) What characteristics make an item or test more accessible? (b) How might changes in test delivery and format interact with altered items? (c) At what point does an alteration to an item affect the construct being measured? (d) How is alignment to the content standards affected by item and test alterations? (e) How do proficiency-level descriptions affect the development of AA-MAS? (f) What criteria should be used to judge student success? (g) How do alterations designed to change the complexity and difficulty of items affect the technical quality of AA-MAS as complete tests?” (ibid, p. 5)

There are also major issues with the standard settings for AA-MAS used for students with disabilities for the 2% student group. For example, how comparable should the cut scores be set for different performance level and how should these cut scores be defined? (Olson, Mead, & Payne, 2002). Answers to these questions require substantial efforts in conducting research in the area of alternate assessments for students with disabilities.

Different forms of alternate assessments have been proposed. Among them are: 1) alternate assessments based on grade-level achievement standards (AA-GLAS), 2) alternate assessments based on alternate achievement standards (AA-AAS), which are usually referred to as the 1% group, and 3) alternate assessments based on modified achievement standards (AA-MAS), often referred to as the 2% group (Gong, 2007). Elliot and Roach (2007) underscore the importance of determining effective strategies for including special needs students in the overall accountability for student achievement, stating:

“Alternate assessments are used with a relatively small population of students with disabilities, yet demand a significant amount of time from educators and state assessment professionals to develop, implement, and evaluate. It appears the efforts of these professionals will need to be extended given the vast majority of states’ have not met the USDOE’s requirements for alignment and technical soundness.”
(pp. 330-1)

Different chapters in this volume address some of the issues raised above. For example, Quenemoen (Chapter 2, this volume) discusses eligibility criteria for students taking AA-MAS. She distinguishes between low-performing students who have disabilities and those with no apparent disabilities. The chapter by Welch and Dunbar (Chapter 6, this volume) discusses issues concerning the development of AA-MAS and the advantages and disadvantages of various options for modifications, and this chapter focuses on the comparability aspect of AA-MAS.

Challenges in Evaluating Comparability

Developing alternate assessments for students with disabilities is quite complex and requires special attention and planning. For example, Lowery et al. (2009) suggest “adherence to the requirement to maintain an individualized, meaningful curriculum for students with severe disabilities complicates delivery of an assessment that is created to measure progress of students toward a standardized curriculum” (page 250). The authors indicate that the use of different approaches by states (e.g. such as simplifying general education standards, redefining them as functional skills, or extending them through the use of foundational skills) brings further complications in the process of developing alternate assessments. Roach (2005) discusses and examines four challenges in designing and implementing alternate assessments for students with significant disabilities. These challenges include: 1) deciding who should participate in alternate assessments, 2) determining the content area that alternate assessments should measure, 3) creating reliable and valid alternate assessments, and 4) defining proficient performance on alternate assessments. While some of these challenges (such as challenge 2 which only applies to AA-AAS) may not apply to AA-MAS, but they emphasize the difficulty in developing these assessments and interpreting their scores.

The U.S. Department of Education (USED) announced the proposed regulation for the AA-MAS in 2005 (Kettler & Almond, 2009). Subsequently, USED provided Peer Review

Guidelines for conducting reviews of state assessment systems including alternate assessments based on modified achievement standards (U.S. Department of Education, December 2007). Based on one USED report (U.S. Department of Education, November 2008), eight states have developed AA-MAS for at least one grade level. The report by the National Technical Advisory Council indicated that “seven of these states have submitted evidence to the Department for peer review but none has met all the requirements” (page 4). One of the main reasons that states were not able to provide sufficient evidence on the comparability of assessments for AA-MAS is that some of the students who face the most challenges in their educational careers belong to subgroups that are small in size. It would be extremely difficult for researchers to examine the factors affecting comparability between AA-MAS and general assessments using traditional research/psychometric methodologies due to such small group sizes. In order to do a comparison between students who take general assessments with those taking alternate assessments, large enough samples are needed in order to detect meaningful differences.

In some categories of low incidence disabilities, there are hardly enough subjects in a school, district, or even in most states to allow for meaningful analyses of data to examine comparability issues. In such cases, researchers may be required to combine some of these categories in order to obtain a large enough sample to conduct studies that are methodologically sound. However, research suggests that issues concerning assessment of students with disabilities might vary across different categories of disabilities; therefore, it may not be reasonable to aggregate findings from students in the different subgroups of disabilities (see, for example, Abedi, Leon & Kao, 2008).

The *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) view comparability as a major foundation underlying valid and fair assessments and allocate an entire chapter to issues regarding

comparability (Chapter 4, pp. 49–60). However, the main focus of the *Standards*' Chapter 4 is on *score comparability*, which can be established through approaches such as scaling test scores. The *Standards* state (p. 49): "Scale scores may aid interpretation by indicating how a given score compares to those of other test takers by enhancing the comparability of scores obtained using different forms of a test, or in other ways."

However, the *Standards* acknowledge the limitations on the score comparability, particularly when implemented in terms of cut score⁶ (p.50): "Criterion-referenced interpretations based on cut scores are sometimes criticized on the grounds that there is very rarely a sharp distinction of any kind between those just below versus just above a cut score."

Approaches to Comparability

In order to address the current need and develop strategies to overcome the challenges, the discussion of comparability in this chapter goes beyond the traditional approach including those discussed in the *Standards*. In addition to score comparability, the chapter discusses comparability in several other areas. Specifically, this study proposes comparability in six major areas: (a) content and construct, (b) depth of knowledge, (c) accommodation, (d) psychometrics, (e) linguistic structure, and (f) basic text features. However, we acknowledge the challenging task of establishing comparability in all six areas. Therefore, we group these comparability features into two categories, (1) required comparability (features "a" through "c"), and (2) complementary or desired comparability (features "d" through "f"). To assure comparability, the test developers must present evidence on the first category, and if feasible, with supplemental (preferred) evidence from the second category. This recommendation of two broad categories is based on literature and experts' opinion (see for example, Allen & Yen, 1979; Thorndike, 2005). However, in many cases, states determine where they would like the

⁶ For a detailed description of cut scores, see Perie (Chapter 7, this volume).

tests to be comparable. For example, the AA-MAS may be a better measure of achievement at the lower end of the scale, but states may want the “proficient” level to be comparable across AA-MAS and general assessments.

Under content and construct, issues concerning achievement standards and proficiency judgments are discussed. The psychometric comparability section provides a discussion of the classical measurement approach in examining comparability of assessments. Under this section, a discussion of reliability and validity as well as scale and score comparability is presented. This section also presents a discussion on structural equation modeling and differential item functioning (DIF) approaches in examining comparability. Under linguistic comparability, a description of grammatical complexity, lexical density, and text length is presented. The comparability in the basic text features includes a discussion of comparability in terms of format, tables, charts, graphs, and pictures. The depth of knowledge section provides a description of a theoretical underpinning of depth of knowledge in the context of comparability and suggests ways to compare the level of depth of knowledge across the two assessments (AA-MAS and the general state assessments). Finally, this chapter reviews comparability in terms of accommodations used for students with disabilities based on their IEPs.

Content and Construct Comparability between AA-MAS and General Assessments

The first and most important criterion for examining comparability of different assessments is to establish content and construct comparability. Assessments that measure different content and constructs may not produce comparable outcomes even if they are shown to be comparable in terms of psychometric characteristics. The concept of content and construct comparability has been discussed from different points of view, including expert judgment, moderation, and alignment with the grade-level content standards. Thus, comparability between AA-MAS and general assessments can be established through expert judgment, moderation by

inspection, social moderation (Winter, 2009), and alignment with the grade-level content standards.

The concept of cognitive demand in assessment is related to the discussion of content and construct comparability. The level of cognitive demand of an assessment (or of an item within an assessment) could be determined by different sources some of which are relevant to the assessments and some of them may be due to the impact of nuisance variables or construct irrelevant sources. For example, irrelevant or poorly labeled visuals may increase the cognitive load of perceiving information for students with disabilities. However, cognitive complexity might be a relevant factor in the assessment. Similarly, in reading comprehension, items that are inferential may significantly increase the cognitive load for students with disabilities and, thus, affect students' ability to display their understanding of the passage. Assessing depth-of-knowledge reveals the level of the cognitive demands of the standards and the cognitive demands of the assessment items. Level 4 of depth of knowledge (extended thinking) requires the highest level of cognitive demand in Webb's model. This level demands complex reasoning, planning, developing, and thinking (Webb, et al., 2006).

Expert judgment. Content comparability can be established through experts' judgment (Mislevy, 1992). A team of experts, including content specialists, teachers, and linguistic experts, could judge the comparability of content across the two assessments. For expert judgment a rubric is often developed and validated to help ensure more consistent judgment across a variety of experts with different backgrounds. To estimate interrater reliability, comparability between the two assessments may be examined by more than one person. Interrater reliability indices such as kappa and intra-class correlations can then be computed and can be compared across the two assessments.

Moderation. Moderation refers to the identification of local scoring instances that are overly stringent or overly lenient to "moderate" those scores to bring them more into line (Burton & Linn, 1994). "Moderation" techniques can be grouped into several categories. A commonly

used approach is classified as moderation by inspection or cross-moderation, which is mainly based on judgmental audits. Another moderation approach is based on statistical moderation. Under this approach, moderation is done based on external criteria. The third approach is the enhancement of one of the two approaches mentioned above or a combination of the two approaches (Burton & Linn, 1994; Linn, 1993; Mislavy, 1992).

Alignment with the grade-level content standards. States conduct alignment studies to demonstrate how and to what extent their assessments are aligned with their content standards (see, for example, Moore & O’Neil, 2004). Alignment is conducted to examine the degree of correspondence between a set of educational standards—often referred to as state content standards—and the assessments that are developed to measure what students are expected to learn in relation to those standards (Moore, & O’Neil, 2004; Webb, 1999, 2002). According to Webb, there are several major criteria for alignment. These criteria include: a) categorical concurrence, b) depth-of-knowledge consistency⁷, c) range-of-knowledge correspondence, and d) balance of representation (Webb, 1999). Studies suggest that Webb’s alignment model, used for the alignment of assessment content with the state content standards for regular state assessments, can be meaningfully applied to alternate assessments, which provide states a way to comply with the requirements of IDEA and NCLB (Roach & Elliott, 2004; Gong & Marion, 2006; Tindal, 2005). Tindal (2005) describes procedures for alignment of alternate assessments using the Webb alignment model.

In fact, the report on the peer review results from six states suggests that test blueprints should provide evidence on the alignment between the AA-MAS and grade-level content standards (Filbin, 2008; Kettler & Almond, 2009). These assessments are required to assess the same breadth and depth as the general assessments.

⁷ A more detailed discussion of the depth of knowledge alignment will be presented later in this chapter.

Psychometric Comparability between AA-MAS and General Assessments

Psychometric comparability data can serve as complementary and supportive evidence to the content and construct comparability. In this section, psychometric comparability is discussed in the context of both classical and modern theory of measurement.

Classical Measurement Approach in Examining Comparability of Assessments. Under the classical test theory, assessment outcomes can be considered comparable if they are from parallel or tau-equivalent tests. To consider different forms of assessments as parallel or tau-equivalent, certain assumptions underlying parallel and tau-equivalent tests must be met. The main assumption underlying classical test theory is that the measurement error is randomly distributed and that the correlation between measurement errors of two tests is zero ($\rho_{E_1E_2} = 0$). This implies that the correlation between the true scores of form A of the test with measurement error of form B of the test is zero ($\rho_{T_1E_2} = 0$). Additionally, if two tests have observed score of X and X' that satisfies the assumption of randomly distributed measurement error, and if, for every population of examinees, the true score of test 1 (T) equals the true score of test 2 (T'), and if the variance of measurement error of test 1 ($\sigma^2_{E_1}$) equals the variance of measurement error of test 2 ($\sigma^2_{E_2}$), then the tests are considered parallel tests (Allen & Yen, 1979; Thorndike, 2005).

However, as indicated by the U.S. Department of Education (2007) and in the literature, AA-MAS assessments differ from states' general assessments in many different aspects. Some of these assessments include fewer items with higher p-values (less difficult items), have shorter and fewer reading passages, have less complex linguistic structures, and use fewer distractors in their multiple choice items (Cortiella, 2007; Kettler & Almond, 2009; Lazarus, et al., 2007). Such systematic differences between states' alternate and general assessments create major limitations on the comparability of the two assessments.

One question is whether a shorter version of the test can be considered as parallel (tau-equivalent) to the full version of the test. As indicated above, a test with fewer items, given all

other parallel test assumptions are true (except for an additive constant, C_{12}), can be considered as a tau-equivalent test to the original version. However, in terms of alternate assessments, it is very difficult to assume that the two tests (the state's general assessment and the alternate assessment) meet any conditions of parallel tests. If the shorter version of the test is different than the full version of the test in terms of linguistic structure, item difficulty, or the number of choices (in multiple-choice format), then the shorter version of the test cannot be considered as a tau-equivalent test. For example, Karvonen and Huynh (2007) indicated that the alternate assessment items typically require simple cognitive processes such as recall.

Reliability, Validity, and Standard Error of Measurement. Assessments used by states for accountability purposes are usually developed and field- tested for mainstream students. In the development process, many of the assessment needs of subgroups (e.g., students with disabilities) may not be adequately considered. Therefore, there may be many sources of nuisance variables that can impact the performance of students with disabilities. These sources, which are also referred to as extraneous variables (Linn & Gronlund, 1995), contaminants, or construct-irrelevant (Haladyna & Downing, 2004; Messick, 1984), may differentially impact the reliability and validity of assessments for students with disabilities. Linn and Gronlund (1995) indicated that "During the development of an assessment, an attempt is made to rule out extraneous factors that might distort the meaning of the scores, and follow-up studies are conducted to verify the success of these attempts" (p. 71). Further, Zieky (1988) cautions that a fairness review to identify construct-irrelevant sources is a major effort when constructing impartial tests. Welch and Dunbar (Chapter 6, this volume) address some of the issues concerning the development of AA-MAS by first discussing the best practice in test development and then highlighting the advantages and disadvantages of various options for modifications.

Reliability. The linguistic complexity of assessment and format and structure of test booklets (e.g. font size, complex and irrelevant charts and graphs, crowded text on pages) may cause fatigue and frustration for students with disabilities and may result in a higher level of

measurement error that can substantially reduce the reliability of assessment outcomes for these students. For example, Abedi, Leon and Mirocha (2003) found a gap of over .32 in the internal consistency coefficient in scores of state assessments in math between students with disabilities and students without disabilities. The standard error of measurement was substantially larger in assessment outcomes for students with disabilities.

More importantly, some of these sources of construct-irrelevant variance may bring another dimension to the measurement model and make it multi-dimensional. This multidimensionality issue would then introduce more complexity into the comparability concept. For example, it would be a challenging task, both in terms of content and psychometric properties, to compare assessment outcomes that are unidimensional in nature (i.e. measuring only the construct relevant aspects of assessment) with the outcomes that represent several dimensions or constructs (construct-irrelevant). Multidimensionality of assessment outcomes may directly impact internal consistency measures (such as alpha coefficient), as these measures are extremely sensitive to multidimensionality and severely underestimate reliability of multidimensional assessments when they are supposed to measure a single construct (Cortina, 1993).

Validity. Sources of construct-irrelevant variance discussed above will not only impact the reliability of assessments, but they also directly affect the construct validity of the assessments. Content-based state tests are designed to measure constructs that are the target of the assessments. Therefore, items within a test are often highly correlated when they are used for students without disabilities for whom the assessments were constructed. For students with disabilities, however, different sources of construct irrelevant variance may negatively impact the validity of these assessments. More importantly, it might be difficult to assess the validity of AA-MAS using external criteria since finding valid external criteria for examining the validity of AA-MAS can be a major challenge.

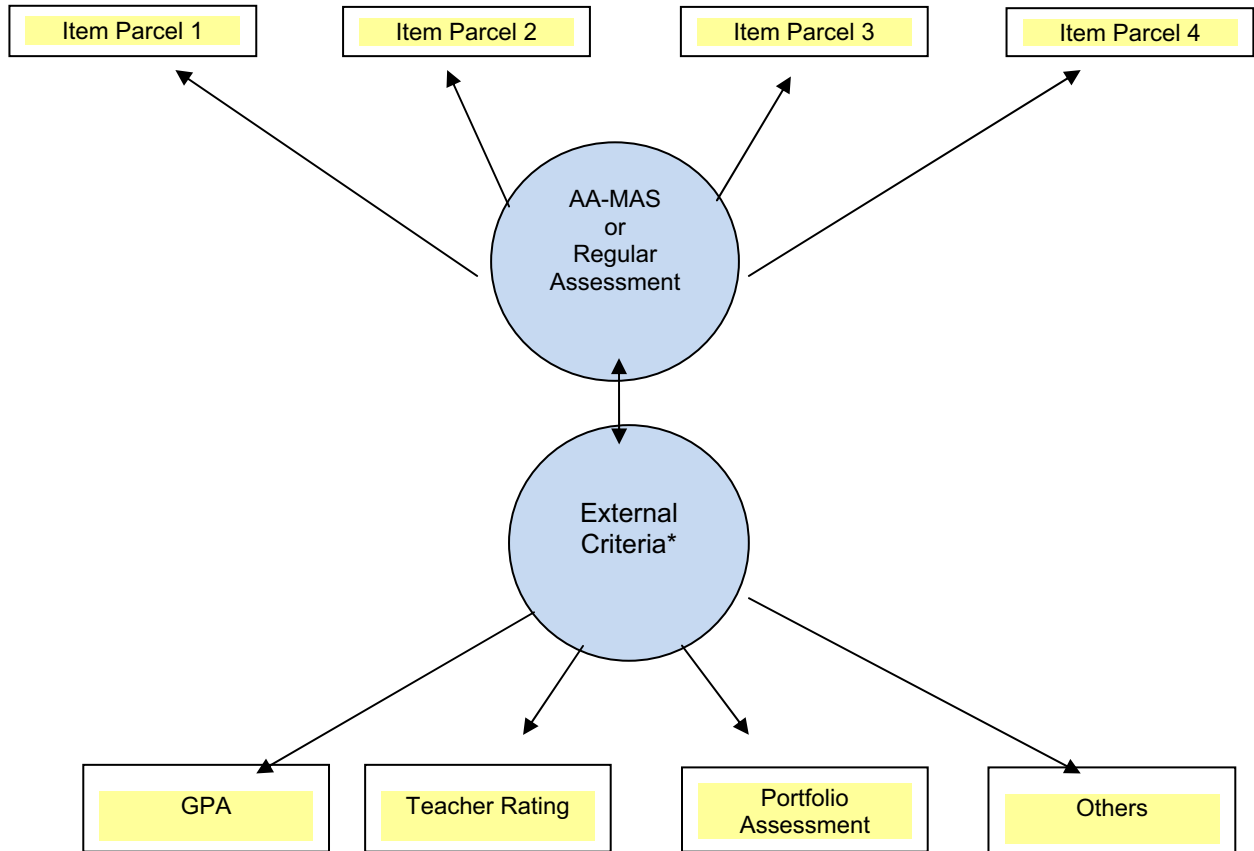
As part of a comprehensive set of studies on score comparability, DePascale (2009) examined the comparability of an AA-MAS (which he called a 2% test). The study addressed validity questions regarding the modified test by examining the relationship between the states' regular test and the modified test. The goals of the study were: "a) to determine that the 2% tests were less difficult than the general tests, and b) to determine that the 2% tests provide more reliable information than the general test in the area of interest for its target population of the 2% test" (p. 11). As one of the major findings of the study, the author indicated that the 2% test provided reliable information at the extreme low end of the scale.

The findings of this study are very informative in terms of psychometric properties of the AA-MAS as compared with those for the regular state assessments. While it is true that the alternate assessments may generally have lower reliability and validity when considering the entire distribution of content knowledge, these assessments do what they set out to accomplish for the lower part of the ability distribution (for a comprehensive presentation of validity of AA-MAS see Marion, Chapter 9, this volume).

Structural Equation Modeling Approach. Comparing the structural relationship between test items, item scores and total test score and between different subscales of the tests across the two assessments (AA-MAS and the general state assessment) using a multiple group confirmatory factor analytic model can provide useful information (Abedi, 2002). Figure 8-1 presents a multiple group confirmatory model that provides comparability evidence. This model includes data from states with two groups of students: 1) students with disabilities taking AA-MAS and, 2) students without disabilities taking general assessments, which can be used in content areas such as math, science, or reading/language arts. A set of item parcels can be constructed based on existing data from each group. Each parcel should include items representing different subscales but items across parcels should be similar. These item parcels can then be used for creating a latent variable for the content-based assessment. A set of performance assessment scores can then be used as external criteria for establishing criterion-

based validity. This set of variables could include student GPA, teacher’s rating of student performance, and a score from the portfolio in the content being assessed. Thus, the model includes two latent variables: one is the test scores, which is computed from the item parcel, and the other is the external criterion.

Figure 8-1. Multi-group Confirmatory Model



*External criteria include assessment outcomes other than state test scores.

In this model, it is not necessary to have an equal number of items across the two assessments; however, the number of parcels across the two groups should be equal. A set of invariance across the two groups of students taking the two different assessments can be tested for significance. These include testing invariance of factor loadings of the item parcels with the overall test latent variables and the overall external criteria and invariance of correlations between the content-assessment’s latent variable with the external criteria latent variable. A

significant outcome on the invariance hypotheses would be an indication of a lack of comparability between the two assessments.

Generalizability approach. A G model can be used to examine comparability between AA-MAS and the regular state assessments. A multi-facet G design can be used to compare the two groups in terms of different sources of measurement error such as variation between items and occasions. The model can be applied separately to each of the two groups. Sources of variability due to items and occasions (and interaction between items and occasions) can be compared across the two groups of subjects taking the two different assessments. The overall G coefficients as well as the percent of variance explained by each of the sources (e.g. subjects, items, occasions, and interaction between items and occasion) can be compared across the two groups. Both relative and absolute decisions for computing G coefficient may be applied and comparisons can be made. For a more detailed discussion of the generalizability concept and instruction on how to conduct a G study see Brennan (2001) and Shavelson & Webb (1991).

The structure and size of the variance components, the significance of the main and interaction effects, and effect sizes across the two assessments can also be compared for any significant differences. For example, if a linguistic complexity facet accounts for 25% of the variance in one assessment but explains less than 5% in the other assessment, then such a difference points to a lack of comparability in terms of the generalizability model.

DIF Approach. Test publishers and states often conduct differential item functioning (DIF) analyses to identify test items that differentially perform across subgroups of students. Different student background variables are used for grouping students. DIF analyses are usually conducted to examine any possible biases due to gender, ethnicity, students' socio-economic status, students' disability status, and students' language background. DIF analyses by students' disability status using states' regular and alternate assessments may shed light on comparability between the two assessments. For example, it would be informative to compare

the trend of results of DIF on a state general assessment in a specific content (e.g., math or science) with the results of DIF on a corresponding AA-MAS assessment by some student background variables such as SES (free/reduced price lunch program) for similarities and differences. One could identify the number of items labeled as “A”, “B”, or “C” DIF across AA-MAS and the general assessments. Similarly, comparing the pattern of uniform and non-uniform DIF across AA-MAS and general assessments could provide useful information. However, there are major limitations in such comparisons when used for comparability purposes.

First, the number of students (sample size) for the AA-MAS is a major challenge since it would be extremely difficult to find a large enough sample to compare with the focal and reference groups. Second, the literature clearly suggests that different procedures for computing DIF may provide quite different outcomes (Abedi, Leon & Kao, 2008). This may be a major problem in using DIF as a criterion for judging comparability of different assessments since different approaches may perform differently across the assessments. Third, and most importantly, test items may perform differentially across students in different subgroups of disabilities. Results of a study on DIF by different subgroups of disabilities found that a substantial number of items were identified as DIF for different disability groups but very few or almost none of the items were identified as DIF across all or several subgroups of disabilities (Abedi, Leon, & Kao, 2008).

Scale and Score Comparability between AA-MAS and General Assessments

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999), scale and score comparability between two assessments is required in order to provide similar interpretation of the outcomes measured by those assessments. Winter (2009) discusses score comparability and indicates that “In general, test scores can be considered comparable if they can be used interchangeably” (page 6). The author argues, however, that comparability depends on the level of scores being used. For example, scores reported at the

scale level or achievement score level can be compared only at that level. Additionally these scores must provide measures of the same set of knowledge and skills, present the same degree of achievement, and have similar technical properties (Winter, 2009).

AA-MAS tests may have major differences from the general assessments, including the number and level of difficulty of test items. Even with such differences some evidences of score comparability can be obtained. “For example, it may be desirable to interpret scores from a shortened (and hence less reliable) form of a test by first converting them to corresponding scores on the full-length version” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, page 52). Converting scores from the AA-MAS and general assessments on the same scale is extremely challenging since such conversions require comparability on many different aspects. As Mislevy (1992) indicated, “No simple statistical machinery can transform the results of two arbitrarily selected assessments so that they provide interchangeable information about all questions about students’ competencies” (p. 91), particularly in the case of AA-MAS and the general assessments, where there are such major and substantial differences. However, despite the limitations, such conversions could provide useful information.

Linguistic Comparability between AA-MAS and General State Assessments

Recent literature on the issues concerning assessments consistently demonstrates the impact of language factors on the assessment outcomes. These factors differentially impact the performance of subgroups of students such as English language learners and students with disabilities, particularly those with learning and reading disabilities. Several linguistic features have been identified in the literature that may have major impacts on the assessment outcomes for these students (Abedi, 2006; Abedi, 2007 [LEP Partnership] Abedi & Lord, 2001; Sato, 2007 [LEP Partnership]). Research literature also suggests that reducing the level of unnecessary linguistic complexity of assessments helps to close the gap between subgroups, such as

students with disabilities and ELL students, and the main group (Abedi, 2006; Abedi, Leon & Mirocha, 2003). The process of reducing the level of unnecessary linguistic complexity of assessments is often referred to as the linguistic modification of assessments. While the linguistic modification process does not affect the performance of native speakers of English at the higher performance level (thus, not affecting the validity of assessments), it helps reduce the performance gap between students with disabilities and students without disabilities.

One approach in examining comparability between AA-MAS and general state assessments is to compare their linguistic structures to see if the level of linguistic complexity is similar across the two assessments. In this comparison, a distinction must be made between linguistic features that are related and those that are not related to the content being measured. To promote comparability with the states' general assessments, the linguistic structure related to the content may not be changed since changing linguistic structures that are content related may alter the construct being measured. Therefore, linguistic modification should only be applied to language that is not related to the construct being measured. The distinction between related and unrelated linguistic features to the content can be made by a team of experts that includes content and linguistic experts.

Literature provides clear guidelines and instructions on how to conduct linguistic modification of assessments and how to distinguish between necessary and unnecessary linguistic complexity in the assessments (Abedi, 2006, 2007; Sato, 2007). These guidelines would help in two important contexts: 1) to identify which linguistic features should be considered in making judgments on comparability between state general and alternate assessments, and 2) to inform the development of alternate assessments when linguistic modification is considered as a factor in the alternate assessment process.

Once again, it is extremely important to distinguish between linguistic structure that is related to the content being measured and unnecessary linguistic complexity that is unrelated to the content. As indicated earlier, some states may choose to remove or reduce unnecessary

linguistic complexity of the AA-MAS to make them more accessible for students with disabilities, which should be considered a reasonable practice. Reducing unnecessary linguistic complexity of assessments makes them also more accessible for students with no disability who are at the lower level of achievement performance distribution.

Assessing the Level of Linguistic Complexity of the AA-MAS and General Assessment Test Items

Outcomes of the studies on the impact of linguistic factors on the assessment of English language learners and students with disabilities have led to identification of 48 linguistic features that make assessment more complex for these students (see, for example, Abedi, 2006; Abedi & Lord, 2001). The first step in examining linguistic comparability is to identify which of the linguistic features are present in the item and the seriousness of their effects. A rating system for evaluating the level of linguistic complexity of test items was developed. The rating system consists of two different rating scales: (1) an analytical scale, and (2) a holistic scale. Test items in the AA-MAS and general assessments may be rated on both scales, and then the ratings can be compared across the AA-MAS and the general assessment. We will elaborate on each of these rating approaches below:

Analytical Rating. Figure 8-2 presents a rubric for rating the level of complexity on each of the 14 features for each test item. The ratings are based on a 5-point Likert scale, with “1” indicating no complexity present with respect for that particular feature and “5” suggesting a high level of linguistic complexity with that feature. Abedi and colleagues combined the 48 linguistic features mentioned above into 14 general categories for ease of rating linguistic complexities (Abedi & Lord, 2001). Ratings were performed on the overall 14 categories. Each test item receives 14 ratings, one for each linguistic feature. For example, with respect to linguistic feature number 1 “Word frequency/familiarity”, if the words used in the item are “very familiar” and “frequently” being used, then the item receives a rating of “1”, “no complexity”. However, if the word is unfamiliar, or being used less frequently, then depending on the level of

unfamiliarity and low frequency, it receives ratings between 2 to 5. Judgments on the familiarity/frequency of the word can be made based on sources such as *The American Heritage Word Frequency Book* (Carroll et al., 1971) and the *Frequency Analysis of English Usage: Lexicon and Grammar* (Francis & Kucera, 1982). The highest rating of 5 in this example would refer to a word that is extremely unfamiliar and rarely occurring.

Figure 8-2. Rubric for Rating Level of Linguistic Complexity

Linguistic Feature	Degree of Complexity				
	Not Complex 1	2	3	4	Most Complex 5
1. Word frequency/ familiarity	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2. Word length	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3. Sentence length	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4. Passive voice constructs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
5. Long noun phrases	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
6. Long question phrases	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
7. Comparative structures	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
8. Prepositional phrases	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
9. Sentence and discourse structure	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
10. Subordinate clauses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
11. Conditional clauses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
12. Relative clauses	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
13. Concrete vs. abstract or impersonal presentations	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
14. Negation	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Holistic Rating. Similar to the ratings that are assigned based on the analytical procedure, this rating is on a 5-point Likert scale, “1” representing items with no or minimal level of linguistic complexity and “5” showing an item with an extremely complex linguistic structure. Figure 8-3 shows the Holistic Rating Rubric. As Figure 8-3 shows, a test item free of linguistic complexity (with a rating of “1”) does not suffer from any of the 14 linguistic complexity threats. For example, the item uses familiar or frequently used words, the words as well as sentences in these items are generally shorter, there are no complex conditional and/or adverbial clauses,

and there are no passive voices or abstract presentations. On the contrary, an item with a severe level of linguistic complexity contains all or many sources of threats.

Figure 8-3. Holistic Item Rating Rubric

	QUALITY
1	<p>EXEMPLARY ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Familiar or frequently used words; word length generally shorter • Short sentences and limited prepositional phrases • Concrete item and a narrative structure • No complex conditional or adverbial clauses • No passive voice or abstract or impersonal presentations
2	<p>ADEQUATE ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Familiar or frequently used words; short to moderate word length • Moderate sentence length with a few prepositional phrases • Concrete item • No subordinate, conditional, or adverbial clauses • No passive voice or abstract or impersonal presentations
3	<p>WEAK ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Relatively unfamiliar or seldom used words • Long sentence(s) • Abstract concept(s) • Complex sentence/conditional tense/adverbial clause • A few passive voice or abstract or impersonal presentations
4	<p>ATTENTION ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Unfamiliar or seldom used words • Long or complex sentence • Abstract item • Difficult subordinate, conditional, or adverbial clause • Passive voice/ abstract or impersonal presentations
5	<p>PROBLEMATIC ITEM</p> <p><i>Sample Features:</i></p> <ul style="list-style-type: none"> • Highly unfamiliar or seldom used words • Very Long or complex sentence • Abstract item • Very difficult subordinate, conditional, or adverbial clause • Many passive voice and abstract or impersonal presentations

Ratings on the linguistic modification (both analytical and holistic) provide diagnostic information on the linguistic barriers present in test items. This information may help item writers or test developers to identify *problem* items. These items can then be corrected for such problems. Since linguistic modification ratings are on a Likert-scale, median ratings can be

computed and can be used for decisions on how the items should be modified. Different patterns of linguistic complexity across the two assessments may lead to the conclusion that the two assessments are not linguistically comparable (for a detailed description of linguistic complexity assessment, see Abedi, 2006).

Basic Text Features

Text Format and Text Features- This feature includes typeface and point size, passage and item placement on page(s), and the relevance and clarity of all visuals within a passage. It is important to consider typeface and point size when determining if a test passage and its items are accessible to students with low visibility. Similarly, pages with excessive blank space, or conversely, with small margins, may unfairly affect students with low visibility. Additionally, it is essential to determine if the visuals (graphs, tables, charts, and pictures) within a passage are relevant, meaning that they are needed to answer the item, and if they are clearly labeled. Visuals that are not relevant or clearly labeled may increase the cognitive load of perceiving information for students with disabilities.

Type of Passage/Item. This feature identifies the genre of the passage (descriptive, narrative, expository, poetry, or persuasive). This feature also determines whether a test item is informational or inferential. Items that are informational can be answered using only slightly paraphrased or verbatim information that is found in the passage, whereas items that are inferential require the student to combine information from the text together with their own background knowledge in order to recognize implicit relationships and outcomes. Therefore, items that are inferential may significantly increase the cognitive load for students with disabilities and, thus, hinder students' ability to accurately display their understanding of the passage.

Comparability in terms of Depth of Knowledge (DOK)

Depth-of-knowledge (DOK) comparability is confirmed if what is elicited from students on the assessments is as cognitively demanding as what students are expected to know and do as stated in the state and national standards. DOK consistency is defined as the level of consistency between the cognitive demands of standards and the cognitive demands of the assessment items. If between 40% and 50% of the assessment items are at or above the DOK levels of the objectives, then the DOK consistency criterion is “weakly” met. Webb (1999) defines four levels of cognitive complexity when comparing the cognitive demands of the standards and assessment items. They are: Level 1 (Recall), Level 2 (Skills and Concepts), Level 3 (Strategic Thinking), and Level 4 (Extended Thinking).

Level 1 (Recall): Level 1 items require students to use simple skills or abilities. Examples include recall of information. Key words that signify Level 1 include *identify, recall, measure, and recognize*.

Level 2 (Skill/Concept): Level 2 items demand a higher level of cognitive complexity compared to Level 1 items. Assessment items at Level 2 require some decision making on how to approach problems or activities. For example, Level 2 keywords for math include terms such as *classify, estimate, compare, and organize*. These actions imply more than one step.

Level 3 (Strategic Thinking): Assessment items at this level require reasoning, planning, and using evidence, which are at a higher level of thinking than the previous two levels. In most instances, at this level students are required to explain their thinking. The cognitive demands at Level 3 are complex and abstract. An activity, however, that has more than one possible answer and requires students to justify the response they give would most likely be at Level 3.

Level 4 (Extended Thinking): Level 4 items require the highest level of cognitive demand in Webb’s model of depth of knowledge. This level demands complex reasoning, planning, developing, and thinking. Assessment items at Level 4 may include activities such as designing

and conducting experiments, analyzing and interpreting results, combining and synthesizing ideas into new concepts, and critiquing experimental designs (Webb, et al., 2006).

Comparability Issues in the Accommodated Assessments for Students with Disabilities

Many different forms of accommodations are being used for students with different types of disabilities. However, some of these accommodations may alter the construct being measured; therefore, the issues concerning comparability of accommodated and non-accommodated assessments are of paramount importance to this chapter since many of the features that are incorporated in AA-MAS are being used as a form of accommodation for students with disabilities.

The most commonly used accommodations for students with disabilities are: using Braille, using computerized assessments, dictation of response to a scribe, extended time, interpreter for instructions, marking answers in test booklets, reading aloud test items, reading or simplifying test directions, and providing test breaks (Thurlow, et al., 2000). We present a summary of studies that have examined the validity of assessments under these accommodations. As can be seen from these summaries, research evidence suggests that some of these accommodations alter the construct being measured. For others, however, there is not much evidence to judge the validity of assessments using those accommodations. Issues concerning validity of accommodations are directly related to comparability of accommodated and non-accommodated assessments. When an accommodation is not valid, i.e., when it alters the construct being measured, then the outcomes of assessments under this accommodation are not comparable with assessments conducted under standard conditions with no accommodations provided.

Braille is used for students with blindness or significant visual impairments. Developing a Braille version of the test may be more difficult for some items than others. It would be challenging to use Braille items with diagrams and special symbols (Bennett, Rock, & Kaplan,

1987; Bennett, Rock, & Novatkoski, 1989; Coleman, 1990). Thurlow & Bolt (2001) recommend using Braille for students with severe visual impairments. Also, Braille is recommended to be paired with extended time (Thurlow & Bolt, 2001).

Computerized assessment can be used for students with physical impairments who have difficulty in responding to items in paper-and-pencil format. Some studies suggest that this accommodation is effective in increasing the performance of students with disabilities (see for example, Russell & Haney, 1997; Russell, 1999; Russell & Plati, 2001). Other studies did not find computerized assessment to be effective (MacArthur & Graham, 1987) or even as effective as traditional assessments (Hollenbeck, Thurlow & Bolt, 2001; Tindal, Stieber & Harniss, 1999; Watkins & Kush, 1988; Varnhagan & Gerber, 1984).

Extended time is one of the most commonly used and most controversial forms of accommodation for students with different types of disabilities (SWD). Some studies found that extended time affects the performance of both SWD and non-SWD students and, therefore, makes the validity of this accommodation suspect. Similarly, Thurlow et al., (2000) expressed concern on the validity of this accommodation. Chiu and Pearson (1999) found extended time to be an effective accommodation for students with disabilities, particularly for those with learning disabilities. Some studies found extended time to help students with disabilities in math (Chiu & Pearson, 1999; Gallina, 1989). However, other studies did not show an effect of extended time on students with disabilities (Fuchs et al., 2000; Marquart, 2000; Munger & Loyd, 1991). Studies on the effect of extended time on language arts did not find this accommodation to be effective (Fuchs et al., 2000; Munger & Loyd, 1991). Thus, research on this particular accommodation produced inconsistent results. More studies are needed to make a firm recommendation regarding the use of this accommodation.

The *interpreter for instructions* accommodation is recommended for students with hearing impairments. Ray (1982) found that adaptations in the directions help deaf children score the same as other students (see also Sullivan, 1982). Thurlow & Bolt (2001) recommend

that using an interpreter for instructions may be beneficial to students with hearing impairments. However, not much information exists on the validity of this accommodation.

The *marking answers in test booklet* accommodation can be used for students with difficulties in mobility coordination. Some studies on the effectiveness of this accommodation did not find a significant difference between those tested under this accommodation and those using separate answer sheets (Rogers, 1983; Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998). However, other studies found lower performance for students using this accommodation (Mick, 1989). Since a majority of studies did not show a performance increase as a result of this accommodation, it would be safe to say that this may not have much of an impact on the construct being measured.

A *read aloud test* is used for students with learning disabilities and students with physical or visual impairments. While some studies found this accommodation to be valid in math assessments (Tindal et al., 1998), there have been concerns over the use of this accommodation in reading and listening comprehension tests (see for example, Burns, 1998; Phillips, 1994), because the construct being measured may be changed and, thus, the validity of the assessment is affected (see also, Bielinski, Thurlow, Ysseldyke, Freidebach, & Freidebach, 2001; Meloy, Deville, & Frisbie, 2000).

The *reading or simplifying test directions* accommodation is appropriate for students with reading/ learning disabilities. A study by Elliot, Kratochwill & McKeivitt (2001) suggests that this accommodation affects the performance of both students with disabilities (63.4%) and students without disabilities (42.9%), thus expressing concerns over the validity of this accommodation.

Test breaks can help students with different forms of disabilities. In a study, DiCerbo, Stanley, Roberts, & Blanchard (2001) found that students receiving test breaks obtained scores significantly higher than those under standard testing conditions, and that middle and low ability readers benefited more from this accommodation than high ability readers. However, another study (Walz, Albus, Thompson, & Thurlow, 2000) found that students with disabilities did not

benefit from a multiple-day test administration, while students without disabilities did benefit. These results show quite the opposite of what is expected of valid accommodations.

The summary of research presented above on some of the commonly used accommodations shows a lack of consensus regarding validity and comparability of accommodated assessments as compared with the general assessments with no accommodations provided. As indicated earlier in this chapter, when accommodations alter the construct being measured, the accommodated assessment outcomes are not comparable with the non-accommodated assessments.

The issues of comparability of accommodated and non-accommodated assessments are relevant to our discussion of comparability between AA-MAS and regular state assessments for two main reasons. First, with AA-MAS some students with disabilities still need accommodations that are recommended by their IEP team. Therefore, knowledge of comparability of accommodated and non-accommodated AA-MAS would help states to provide comparability evidence to peer reviewers and other authorities. Second, the concept of comparability is well defined on accommodation studies. While there may not be sufficient literature on the comparability of all accommodations used by states, we have enough research to help us design comprehensive studies for examining comparability between various assessments.

Recommendations to NYSED for Establishing Comparability between AA-MAS and General Assessments

Comparability of the outcome of AA-MAS with the general state assessments is one the most fundamental aspects of the assessment and accountability system for students with disabilities who are eligible for taking AA-MAS . The comparability issues may seriously impact many aspects of the academic careers of these students (often referred as the 2% group), including their instruction, promotion, and graduation. Literature presents the comparability argument mostly in terms of psychometric and content comparability (see for example, AERA,

APA, NCME Standards, 1999; DePascale, 2009). While content and psychometric comparability provides convincing evidence on the comparability of AA-MAS with the general state assessments, looking at a more comprehensive picture on comparability will provide states with the information they need to present a strong justification for development and use of AA-MAS.

In this chapter we presented different criteria for judging and examining the comparability between AA-MAS and state general assessments. These criteria included comparability with respect to content and construct, alignment with the state content standards, classical measurement concept of comparability, psychometrics, linguistics, depth of knowledge and accommodations. Such discussions and guidelines could help the State of New York in developing and validating AA-MAS assessments in different content areas. Useful information and guidelines are also provided by researchers and practitioners for developing AA-MAS tests. For example, in a report by the Council of Chief State School Officers (CCSSO, 2007), guidelines are provided on strategies for states to prepare for and respond to peer reviewers. Similarly, researchers provided recommendations on the use of cognitive interviews in the design and development of AA-MAS (Almond, et al., in press). Literature also provides a summary of research on item and test alteration focusing on AA-MAS along with guidelines on the nature and implementation of these alterations (Kettler & Almond, 2009). While many test publishers and states provided comparability evidence on a few of these aspects, this chapter may help New York test developers to provide a comprehensive plan for comparability of any future AA-MAS development.

In general, to respond to the mandate of inclusion of students with disabilities, states must be able to present evidence that alternate assessment outcomes are comparable with the outcomes of general assessments. Lack of comparability between the alternate and general assessments jeopardizes the academic career of students with disabilities in many different ways, including the promotion and graduation of these students.

While the proposed criteria for examining comparability in this chapter apply to different content areas, the application of some of these criteria may be slightly different across different content areas. For example, the linguistic comparability concept may apply differently to content areas in which language is the target of measurement (e.g., reading/language arts) in areas where unnecessary linguistic complexity may be considered as construct irrelevant sources (e.g., math and science).

A major application of the comparability discussion may be on the type of credential that is most appropriate for students in the State of New York (Regents, Local diploma, IEP certificate). Some options are available for students with disabilities taking AA-MAS. If the comparability between AA-MAS and the regular state assessments can be established, then it would be reasonable to recommend credentials for students with disabilities (particularly for those eligible for the AA-MAS) that are similar to those recommended for non-disabled students. For example, students can receive a Local diploma if they follow the same academic program as the Regents diploma but at a lower cut score on the exam. Would a lower cut score make the assessment outcome less aligned with the state content standards for a passing grade or graduation? Currently, the IEP certificate is primarily for those students who have significant cognitive disabilities and are taking the AA-AAS. For students with disabilities who are taking AA-MAS, the more reasonable option is the local diploma. A report by NCEO (Wiener, 2006) presents one alternative way to meet diploma requirements using an AA-GLAS rather than an AA-MAS.

Guidelines for Examining Comparability of AA-MAS: How Much Comparability is Necessary?

In this chapter many different approaches to comparability of AA-MAS with the general state assessments were discussed. It would be extremely challenging, if not impossible, to establish comparability between the two assessments in all areas discussed in this chapter. Therefore, the main question is in which areas and to what extent evidence is needed to

suggest the AA-MAS outcome is comparable with the general assessment outcome. To answer this question, we define comparability in two levels. Level 1 includes comparability features that are necessary and are required in order to assume comparability across the two measures, and level 2 includes features that are desired but not absolutely necessary in establishing comparability between the two assessments.

Necessary Features for Establishing Comparability between AA-MAS and General Assessments

The decision on which comparability features are absolutely necessary and which are desired may be more speculative as there is not enough research evidence on which to base a decision. Therefore, based on existing literature and based on the author's own professional judgment, the following features are deemed to be necessary as the minimum requirement to establish comparability between AA-MAS and regular state assessments:

1. **Content and Construct Comparability.** This feature is one of the most important aspects of comparability. This level of comparability can be established by applying a combination of different approaches such as experts' review and alignment to the state content standards. In conducting expert reviews, New York State Education Department (NYSED) may form a team of experts in the targeted content area to judge the level of comparability of AA-MAS with the regular state assessment. The team should include experts in the area of assessment and accommodation of students with disabilities focusing on the 2% population, content area experts, and test item writers. NYSED could develop and validate a rubric for assessing comparability. The rubric validation process should include focus groups and cognitive labs to assure clarity of instruction for rating comparability. The rubric may use a 5-point Likert-Scale for rating comparability. NYSED may then decide on the level of exact or within point agreement between AA-MAS and general assessment ratings.

2. **Scale and Score Comparability.** As recommended by the Joint Standards (AERA, APA, & NCME, 1999, page 52) score comparability can be roughly achieved by converting scores from the AA-MAS and general assessments on a same scale. While such conversion is extremely challenging it could provide useful information.
3. **Depth of Knowledge Comparability.** One could expect a fair level of comparability between two assessments measuring the same content if they measure the same level of depth of knowledge. As in the case of content and construct comparability discussed above, NYSED can form a team of experts to judge the level of depth of knowledge across the two assessments. Following Webb's methodology (Webb, 1999), ratings of the depth of knowledge can be provided and compared. A cut point on the level of consistency between ratings of the depth of knowledge of the two assessments can then be used to judge the comparability.
4. **Accommodation Comparability.** Accommodated assessment outcomes could be invalid if the accommodations alter the construct. Many students with disabilities require accommodations based on their IEPs. However, research on the validity of accommodations for students with disabilities is very limited. NYSED may compare accommodations used under the two assessment conditions and provide evidence based on the literature that there are no validity concerns that could differentially affect validity of accommodated assessments under the two testing conditions.

Desired Features for Establishing Comparability between AA-MAS and General Assessments

1. **Psychometric Comparability.** A comparison between the overall psychometric properties of the two assessments may shed light on the comparability issues. It would be informative to compare the reliability and validity coefficients of the two assessments. For example, a comparison between the internal consistency coefficients (Cronbach alpha) between the two assessments can be done. Similarly, criterion-related validity

coefficients of the two assessments can be compared. For example, the structural relationships between parcels of items and the total test as well as the relationships between test scores and external criteria can be compared using a multiple group confirmatory factor analyses as elaborated in Figure 8-1. Examining a set of invariance between the structural relationships of the two assessments may shed light on the comparability of the two assessments.

2. **Comparability between Linguistic Structure of the Two Assessments.** Different features of linguistic complexities that may impact the validity of assessments were introduced earlier in this chapter. A comparison between analytical ratings (Figure 8-2) and holistic ratings (Figure 8-3) would provide supporting evidence on the comparability of the assessments.
3. **Comparability between the Two Assessments on Basic Text Features.** Comparability between the basic features of the two assessments may provide additional evidence of comparability. It would be helpful if the text features such as the presentation of the assessments (e.g., computer versus paper-and-pencil), formatting, fonts, tables and charts, and pagination of the two assessments are similar. For example, two assessments may not be highly comparable if one uses complex tables and charts or crowded pages and the other uses simple tables and charts with a large point size and less crowded pages.

References

- Abedi, J. (2002). Standardized achievement tests and English language learners: Psychometrics issues. *Educational Assessment, 8*(3), 231-257.
- Abedi, J., Leon, S. and Kao, J. (2008). Examining Differential Item Functioning in Reading Assessments for Students with Disabilities. Los Angeles: University of California, Center for the Study of Evaluation/National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., Leon, S., & Mirocha, J. (2003). *Impact of student language background on content-based performance: Analyses of extant data* (CSE Tech. Rep. No. 603). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing.
- Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education, 14*(3), 219-234.
- Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.
- Almond, P. J., Cameto, R., Johnstone, C. J., Laitusis, C., Lazarus, S., Nagle, K., Parker, C. E., Roach, A. T., & Sato, E. (in press). *White paper: Cognitive interview methods in reading test and item design and development for alternate assessments based on modified academic achievement standards (AA-MAS)*. Dover, NH: Measured Progress and Menlo Park, CA: SRI International.
- Bennett, R.E., Rock, D.A., & Jirele, T. (1987). GRE score level, test completion, and reliability for visually impaired, physically handicapped, and nonhandicapped groups. *The Journal of Special Education, 21* (3), 9-21.
- Bennett, R.E., Rock, D.A., & Kaplan, B.A. (1987). SAT differential item performance for nine handicapped groups. *Journal of Educational Measurement, 24* (1), 44-55.
- Bennett, R.E., Rock, D.A., & Novatkoski, I. (1989). Differential item functioning on the SAT-M Braille Edition. *Journal of Educational Measurement, 26* (1), 67-79.
- Bielinski, J., Thurlow, M., Ysseldyke, J., Freidebach, J., & Freidebach, M. (2001). *Read-aloud accommodation: Effects on multiple-choice reading & math items (Technical Report 31)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Brannon Brennan, R. L. (2001). *Generalizability theory*. New York: Springer-Verlag.
- Burns, E. (1998) *Test accommodations for students with disabilities*. Springfield: Charles C. Thomas, Publisher, LTD.
- Burton and Linn, 1994. E. Burton and R.L. Linn , Comparability across assessments: Lessons from the use of moderation procedures in England. In: *CSE Technical Report 369*, National Center for Research on Evaluation, Standards, and Student Testing (1994).

- Chiu, C. W. T., & Pearson, P. D. (1999). *Synthesizing the effects of test accommodations for special education and limited English proficiency students*. Paper presented at the National Conference on Large Scale Assessment.
- Coleman, P.J. (1990). Exploring visually handicapped children's understanding of length (math concepts). (Doctoral dissertation, The Florida State University, 1990). *Dissertation Abstracts International*, 51, 0071.
- Cortiella, C. (2007). *Learning opportunities for your child through alternate assessments: Alternate assessments based on modified academic achievement standards*. Minneapolis, MN: University of Minnesota, National center on Educational Outcomes.
- Cortina, J. M (1993). What is coefficient alpha? An examination of theory and application. *Journal of Applied Psychology*, 78, 98-104.
- DePascale, C. (2009). *Modified tests for modified achievement standards: Examining the comparability of a 2% test*. Dover, NH: National Center for the Improvement of Educational Assessment.
- DiCerbo, K., Stanley, E., Roberts, M., & Blanchard, J. (April, 2001). Attention and standardized reading test performance: Implications for accommodation. *Paper presented at the annual meeting of the National Association of School Psychologists, Washington, DC, 2001*.
- Eckhout, T., Larsen, A., Plake, B., & Smith, D. (2007). Aligning a state's alternative standards to regular core content standards in reading and mathematics: A case study. *Applied Measurement in Education* 20(1), 79-100.
- Elliott, S., Kratochwill, T., & McKeivitt, B. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities. *Journal of School Psychology*, 31(1), 3-24.
- Elliott, S. N., & Roach, A. T. (2007). Alternate assessments of students with significant disabilities: Alternative approaches, common technical challenges. *Applied Measurement in Education*, 20, 301-333.
- Elosua, P., & Lopez-Jauregui, A. (2008). Equating between Linguistically Different Tests: Consequences for Assessment. *Journal of Experimental Education*, 76(4), 387-402.
- Filbin, J. (2008). Lessons from the initial peer review of alternate assessments based on modified achievement standards. Paper developed for the U.S. Department of Education, Office of Elementary and Secondary Education.
- Francis, W. N., & Kucera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.
- Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., & Karns, K. (2000). Supplementing teacher judgments about test accommodations with objective data sources. *School Psychology Review*, 29 (1), 65-85.

- Fuchs, L.S., Fuchs, D., Eaton, S.B., Hamlett, C., Binkley, E., & Crouch, R. (2000). Using objective data source to enhance teacher judgments about test accommodations. *Exceptional Children*, 67 (1), 67-81.
- Gallina, N.B. (1989). Tourette's syndrome children: Significant achievement and social behavior variables (Tourette's syndrome, attention deficit hyperactivity disorder) (Doctoral dissertation, City University of New York, 1989). *Dissertation Abstracts International*, 50, 0046.
- Gong, B. (1999). *Relationship between student performance on the MCAS (Massachusetts Comprehensive Assessment System) and Other Tests*. National Center for the Improvement of Educational Assessment, Inc.
- Gong, B. (2007). *Considerations in designing a "2% Assessment" (AA-MAS): A beginning framework and examples of conceptual possibilities*. Paper presented at the Special Education Partnership Conference on Alternate Assessments Based on Modified Academic Achievement Standards. Washington, DC July 26, 2007.
- Gong, R., & Blank, R. (2002). *Designing school accountability systems: Towards a framework and process*. Washington, DC: The Council of Chief State School Officers.
- Gong, B. and Marion, S. (2006). *Dealing with flexibility in assessment for students with significant cognitive disabilities*. National Center for the Improvement of Educational Assessment, Inc.
- Haladyna, T. M., & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*, 23(1), 17-27.
- Hollenbeck, K., Tindal, G., Stieber, S., & Harniss, M. (1999). Handwritten vs. word processed statewide compositions: Do judges rate them differently? Eugene, OR: University of Oregon, BRT.
- Karvonen, M., & Huynh, H. (2007). Relationship between IEP characteristics and test scores on alternate assessment for students with significant cognitive disabilities. *Applied Measurement in Education*, 20(3), 273-300.
- Kettler, R., & Almond, P. (2009). *Improving reading measurement for alternate assessment: Suggestions for designing research on item and test alterations*.
- Lazarus, S. S., Rogers, C., Cormier, D., & Thurlow, M. L. (2008). *States' participation guidelines for alternate assessments based on modified academic achievement standards (AA-MAS) in 2008* (Synthesis Report 71). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Lazarus, S.S., Thurlow, M. L., Christensen, L.L., & Cormier, D. (2007). States' alternate assessments based on modified achievement standards (AA-MAS) in 2007 (Synthesis Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Linn, R. (1993). Linking results of distinct assessments. *Applied Measurement in Education*, 6(1), 83-102.

- Linn, R. L., & Gronlund, N. E. (1995). *Measurement and Assessment in Teaching*. (7th ed.). NJ: Merrill.
- Lowrey, K.A., Drasgow, E., Renzaglia, A., & Chezan, L. (2009). Impact of alternate assessment on curricula for students with severe disabilities. *Assessment for Effective Intervention*, 32(4), 244-253.
- MacArthur, C.A., & Graham, S. (1987). Learning disabled students' composing under three methods of text production: Handwriting, word processing, and dictation. *The Journal of Special Education*, 21 (3), 22-42.
- Marion, S. (2006, October 10). *Introduction to Comparability*. Presented at the Seminar on Inclusive Assessment in Denver, CO.
- Marquart, A. (2000). *The use of extended time as an accommodation on a standardized mathematics test: An investigation of effects on scores and perceived consequences for students of various skill levels*. Paper presented at the annual meeting of the Council of Chief State School Officers, Snowbird, UT.
- Meloy, L.L., Deville, C., & Frisbie, C. (2000). The Effect of a Reading Accommodation on Standardized Test Scores of Learning Disabled and Non Learning Disabled Students. Paper presented at the annual meeting of the National Council on Measurement in Education (New Orleans, LA).
- Messick, S. (1984). The psychology of educational measurement. *Journal of Educational Measurement*, 21, 215-237.
- Mick, L.B. (1989). Measurement effects of modifications in minimum competency test formats for exceptional students. *Measurement and Evaluation in Counseling and Development*, 22, 31-36.
- Mislevy, R. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects*. Princeton, NJ: ETS Policy Information Center.
- Moore, A.D. & O'Neal, S. (2004). *A study of the alignment between the New Mexico K-12 Content Standards, Benchmarks, and Performance Standards and the draft state assessment*. (Unpublished research study.) Santa Fe, NM: New Mexico State Department of Education.
- Munger, G.F., & Loyd, B.H. (1991). Effect of speededness on test performance of handicapped and nonhandicapped examinees. *Journal of Educational Research*, 85 (1), 53-57.
- Olson, B., Mead, R., & Payne, D. (2002). *A report of the standard setting method for alternate assessments for students with significant disabilities* (Synthesis Report 47). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Perez, J.V. (1980). Procedural adaptations and format modifications in minimum competency testing of learning disabled students: A clinical investigation (Doctoral dissertation, University of South Florida, 1980). *Dissertation Abstracts International*, 41, 0206.

- Phillips, S.E. (1994). High stakes testing accommodations: Validity vs. disabled rights. *Applied Measurement in Education*, 7 (2), 93-120.
- Rabinowitz, S., & Schroeder, C. (2006). Creating aligned standards and assessment systems. Washington, DC: The Council of Chief State School Officers.
- Ray, S.R. (1982). Adapting the WISC-R for deaf children. *Diagnostique*, 7, 147-157.
- Roach, A. T. (2005). Alternate Assessment as the "Ultimate Accommodation": Four Challenges for Policy and Practice. *Assessment for Effective Intervention*, 31, (1), 73-78.
- Roach, A. T., & Elliott, S.N. (2004). Alignment analysis and standard setting procedures for alternate assessments. WCER Working Papers, No. 2004-1. Available at: http://www.wcer.wisc.edu/publications/workingpaper/abstract/Working_Paper_No_2004_1.asp
- Rogers, W.T. (1983). Use of separate answer sheets with hearing impaired and deaf school age students. *B.C. Journal of Special Education*, 7 (1), 63-72.
- Russell, M. (1999). Testing writing on computers: A follow-up study comparing performance on computer and on paper. *Educational Policy Analysis Archives*, 7.
- Russell, M., & Haney, W. (1997). Testing writing on computers: An experiment comparing student performance on tests conducted via computer and via paper-and-pencil. *Educational Policy Analysis Archives*, 5(3).
- Russell, M., & Plati, T. (2001). Effects of computer versus paper administration of a state-mandated writing assessment. *TCRecord.org*. Retrieved January 23, 2001, from the World Wide Web: <http://www.tcrecord.org/PrintContent.asp?ContentID=10709>.
- Sato, E. (2007). *A Guide to Linguistic Modification: Increasing English Language Learner Access to Academic Content*. Washington, DC: The U.S. Department of Education—LEP Partnership.
- Shavelson, R. J., & Webb, N. M. (1991). *Generalizability, theory:Aprimer*. Newbury Park, CA: Sage Publication.
- Subkoviak, J.J. (1988). A practitioner's guide to computation and interpretation of reliability indices for mastery test. *Journal of Educational Measurement*, 25(1), 47-55.
- Sullivan, P.M. (1982). Administration modifications on the WISC-R Performance Scale with different categories of deaf children. *American Annals of the Deaf*, 127 (6), 780-788.
- Thompson, S., Lazarus, S., Clapper, A., & Thurlow, M. (2006). Adequate yearly progress of students with disabilities: Competencies for teachers. *Teacher Education and Special Education*, 29 (2), 137-147.
- Thorndike, R. M. (2005). *Measurement and Evaluation in Psychology and Education*. New Jersey, Pearson, Merrill.

- Thurlow, M. & Bolt, S. (2001). Empirical support for accommodations most often allowed in state policy. Minnesota: National Center for Educational Outcome. NCEO Synthesis Report #41.
- Thurlow, M., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 Update (Synthesis Report 33)*. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.
- Tindal, G. (2005). *Alignment of alternate assessments using the Webb system*. Washington, DC. Council of Chief State School Officers.
- Tindal, G. Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An empirical study of student response and test administration demands. *Exceptional Children*, 64 (4), 439-450.
- U.S. Department of Education (2007). Modified Academic Achievement Standards, Non-Regulatory Guidance. Washington, DC.
- Varnhagen, S., & Gerber, M.M. (1984). Use of microcomputers for spelling assessment: Reasons to be cautious. *Learning Disability Quarterly*, 7, 266-270.
- Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students (Minnesota Report 34)*. Minneapolis: University of Minnesota, National Center on Educational Outcomes.
- Watkins, M.W., & Kush, J.C. (1988). Assessment of academic skills of learning disabled students with classroom microcomputers. *School Psychology Review*, 17 (1), 81-88.
- Webb, N. L. (1999). *Alignment of science and mathematics standards and assessments in four states* (NISE Research Monograph No. 18). Madison: University of Wisconsin-Madison, National Institute for Science Education. Washington, DC: Council of Chief State School Officers.
- Webb, N.L. (2002). An analysis of the alignment between mathematics standard and assessments for three states. Paper presented at the American Educational Research Association meeting in New Orleans, LA, April 1-5, 2002.
- Webb, N.L., Horton, M., & O'Neal, S. (1999). An analysis of the alignment between language arts standards and assessments for four states. Paper presented at the American Educational Research Association meeting in New Orleans, LA, April 1-5, 2002. 30
- Welch and Dunbar, S. (this volume). *Developing items and assembling test form for the alternate assessment based on modified achievement standards (AA-MAS)*. In NYCC White Paper on the Alternate Assessment Based on Modified Achievement Standards (AA-MAS). New York: New York State Education Department.
- Wiener, D. (2006). *Alternate assessments measured against grade-level achievement standards: The Massachusetts "competency portfolio"* (Synthesis Report 59). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Winter, P. C. (2009). *Comparing apples to apples: Challenges and approaches to establishing the comparability of test variation*. Paper presented at the annual meeting of the National Council of Measurement in Education. San Diego, California.

Wright, N., & Wendler, C. (1994). Establishing timing limits for the new SAT for students with disabilities. Paper presented at the Annual Meeting of the National Council on Measurement in Education (New Orleans, LA, April 4-8, 1994).

CHAPTER 9

CONSTRUCTING A VALIDITY ARGUMENT FOR ALTERNATE ASSESSMENTS BASED ON MODIFIED ACHIEVEMENT STANDARDS (AA-MAS)

Scott Marion

States are facing complex issues as they have begun developing alternate assessments based on modified achievement standards (AA-MAS). While several researchers have been working to improve the validity evaluations of state assessments in recent years, with a more intense focus on alternate assessments based on alternate achievement standards (AA-AAS; Elliott, Compton, & Roach, 2007; Marion & Pellegrino, 2006; Rabinowitz & Sato, 2005; Shafer, 2005), these challenges are just beginning to be addressed for AA-MAS. The AA-MAS requires a more careful validity evaluation than one might undertake for either the AA-AAS or the general assessment. This is due, in part, to the uncertain conceptual framework supporting this assessment initiative as well as the novelty of the enterprise. This does not downplay the need for validity work on the general and other alternate assessments; rather the lack of conceptual grounding in the case of the AA-MAS requires a thorough validity evaluation. This evaluation should provide the state with information about how to improve the program or even to help the state determine if the AA-MAS is “worth it.” That is, do the benefits (instructional, assessment, accountability, and social justice) outweigh the costs, including negative unintended consequences, of implementing an AA-MAS?

Many writers of technical reports for general assessments nominally align their analyses and results with the *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 1999), particularly when there are student or school stakes requiring that the inferences drawn from the assessment be valid, reliable, and fair (AERA, APA, & NCME, 1999). This is an obvious and important first step, but one that is often not fully met. Leading measurement theorists (e.g., Cronbach, Messick), including the

authors of the 1985 and 1999 *Standards* (AERA, APA, & NCME, 1985, 1999) are clear that validity is 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, APA, & NCME, 1999, p.9). In other words, test scores convey interpretations and inferences that must be verified by both empirical evidence and a logical argument.

The challenge, however, has moved from having states and test contractors conduct research/evaluation studies to investigate particular aspects of testing programs to designing systematic validity plans for evaluating the efficacy of comprehensive validity arguments. This approach requires synthesizing the various empirical results against a theory of action and validity argument (Kane, 2006). This chapter, drawing heavily on Kane (2006), outlines a framework for constructing and evaluating a validity argument for a state’s alternate assessment on modified achievement standards (AA-AAS), by first briefly describing Kane’s argument-based approach to validation in general and as applied to alternate assessment specifically and then presenting strategies for organizing and prioritizing validity evaluations. The last part of the chapter summarizes the types of evidence one might collect as part of such an evaluation. Examples are presented throughout the chapter to make some of these ideas more concrete.

Framework

The proposed validity evaluation is based on a unified conception of validity centered on the inferences related to the construct including significant attention to the social consequences of the assessment (Cronbach, 1971, Messick, 1989, Shepard, 1993). Kane’s (2006) argument-based approach serves as the focus because it offers several pragmatic advantages over evaluations based in the construct model, primarily in terms of prioritizing studies and synthesizing the results of the various studies. At its simplest, Kane’s approach asks the evaluator to search for and evaluate all the threats to the validity of the assessment inferences.

If these threats are not substantiated, the inferences drawn from the assessment results may be supported, at least tentatively. Unfortunately, “tentatively” is the best that can be accomplished with these sorts of falsification-based endeavors. The term validity evaluation is used to encompass the interpretative and validity arguments (discussed below), the plan for conducting various validity studies, the studies themselves, and the evaluation of the results.

Why an Argument?

Kane’s (2006) argument-based framework “...assumes that the proposed interpretations and uses will be explicitly stated as an argument, or network of inferences and supporting assumptions, leading from observations to the conclusions and decisions. “Validation involves an appraisal of the coherence of this argument and of the plausibility of its inferences and assumptions” (p. 17). A validity argument serves to organize studies, provides a framework for analysis and synthesis, and forces critical evaluation of claims using a falsification orientation. For example, part of a validity argument for an AA-MAS should relate to the claim that the modified assessment is measuring “grade-level” knowledge and skills. The content-related evidence then should include information that would allow one to challenge this grade-level claim if, in fact, the test was measuring below grade-level content. An argument-based approach requires the user, developer, and/or evaluator to search for reasons why the intended inferences are NOT supported. Obviously, in practice one cannot search for ALL reasons, so there is a need to prioritize studies. There are several approaches for prioritizing the studies, but using the theory of action and classes of evidence, both discussed later, offer useful frames for thinking about how to prioritize the considerable number of potential interesting studies.

Kane’s Argument-Based Framework

Kane proposed using two 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). In other words, the interpretative argument outlines what the user/evaluator thinks should occur (and why it should occur) as a result of the testing and related systemic endeavors, while the validity argument is essentially the conclusions drawn after weighing the available evidence and logic. A major advantage of Kane’s approach 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 the initial assumptions and resulting inferences is somewhat more straightforward than evaluating the validity of an assessment under a construct model. This does not mean that construct validity is not the focus of the validity evaluation. Kane’s approach simply provides a different orientation and more pragmatic approach for evaluating the validity of the score inferences than under a strict construct model. The construct model is based on more of a research approach where one is searching for causal connections, whereas Kane’s argument-based approach works from an evaluation perspective where one is trying to determine whether a program is operating as intended with minimal unintended consequences.

Kane (2006) pushes 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, designing a fully coherent system built on a sound theoretical model of learning and use has been receiving more attention in the last decade, in part as a result of the publication of *Knowing What Students Know* (Pellegrino, Chudowsky, & Glaser, 2001; see also Pellegrino, Chapter 4, this volume). Unfortunately most assessments do not start from an explicit attention to validity in the design phase so many current-day evaluators working with states are put in the position of having to retrofit a validity argument to the existing system. However, in the case of the AA-

MAS, there is no excuse—since the work is so new—for not starting the validity work at the beginning of the design phase. For example, Pellegrino (Chapter 4, this volume) provides an extensive set of examples showing how understanding the ways in which students develop competence in the domain should guide assessment development.

The Interpretative Argument

The interpretative argument is essentially a mini-theory as it 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. Most importantly, theories are falsifiable and making the connection between the interpretative argument and “mini-theory” is intended to emphasize that validation is not a confirmationist exercise. It is helpful to think of the interpretative argument as a series of “if-then” statements, such as, *if the student is appropriately selected to participate in the AA-MAS, then the observed score will more accurately reflect the student’s grade level knowledge and skills.*

Kane (2006) noted two stages of the interpretative argument. The development stage focuses on the development of measurement tools and procedures as well as the corresponding interpretative argument. Kane (2006) suggested that it is appropriate to have a confirmationist bias (a stance that favors evidence and interpretations supporting the current state of the assessment system) in this stage since the developers (state personnel and contractors) are trying to make the program as good as possible. During the appraisal stage Kane argues that there should be more of a focus 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. However, given the uncertain conceptual foundations of the AA-MAS, it will be important to temper Kane’s allowance of a confirmationist bias during any stage and consider adopting a more critical stance throughout the validity evaluation.

One of the most effective challenges to interpretative arguments (or scientific theories, in general) is to propose and substantiate an alternative argument that is equally or more plausible than the proposed proposition (or hypothesis in terms of scientific theory). With AA-MAS, users must seriously consider and challenge themselves with competing alternative explanations for test scores. For example, one might want to propose (and confirm) that *increases in students scoring at the proficient level on the AA-MAS who were not proficient previously on the general assessment reflects the fact that the modifications made on the AA-MAS allowed the student to better show what they know on the same constructs*. However, the evaluator must consider plausible alternative hypotheses such as *increases in students scoring at the proficient level on the AA-MAS who were not proficient previously on the general assessment might be due to developing an easier test so students answered more items correctly but on a reduced range of constructs and difficulty*.

Bringing this back to a more simple and pragmatic level, 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. Using the assertion and alternate hypothesis in the example above, the evaluator should design studies that evaluate the rigor of the test using some form of cognitive interview to judge whether student responses reflect differences in demonstrated knowledge and skills when comparing the general and modified assessments. The evaluator would then analyze these data in light of both the original and alternative hypotheses. In essence, validity evaluators are continually trying to challenge the supportability of the claims put forth about the testing program.

Values and Consequences

Kane and others suggest that the evaluator must attend to values and consequences when evaluating a decision procedure 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 as a result of test scores must be evaluated.

There might be a lingering theoretical debate about whether consequences are integral to construct validity, but most leading validity theorists (e.g., Cronbach, 1971; Lane & Stone, 2002; Linn, Baker, & Dunbar, 1991; Messick, 1989, 1995; Shepard, 1997) have argued convincingly that consequences are as much a part of validity as is content or any other source of evidence. However, whether or not one agrees with this view of validity, alternate assessments are used for important policy decisions and the consequences of these decisions must be considered in validity evaluations. This is especially true when evaluating the validity of an AA-MAS where stakeholders and evaluators must be particularly attentive to unintended negative consequences that may arise from lower expectations or other potential denied/reduced opportunities for grade-level instruction.

Guiding Philosophy, Purposes, and Uses

It has become axiomatic to say that the validity of an assessment (actually the inferences from the assessment scores) can be judged only in the context of specified purposes and uses. Further, the guiding philosophy must be considered when evaluating the validity of the AA-MAS. The term ‘guiding philosophy’ is used here in the same way that Quenemoen (Chapter 2, this volume) used it earlier. It is meant to describe a particular orientation, set of assumptions, and beliefs about a particular program or policy. For example, if state leaders believe that students eligible for the AA-MAS can score at a level comparable to proficient on the general assessment except that their disability interacts with their chances to show what they know and/or they have not yet been well instructed, then that would lead to certain types of assessment designs and validity arguments. On the other hand, if the leaders believe that eligible students would have little chance, even if well instructed, to score at a level comparable to the proficient score on the general assessment, then that would lead to quite a different

assessment design. The discussions in this white paper are much more aligned with the first example than the second, but the point here is that state leaders need to be explicit and honest about the philosophy behind their decision to develop an AA-MAS.

A state's guiding philosophy should help explain what the state envisions for the relationship among the AA-AAS, AA-MAS, and the general assessment. Most states, as well as the USED regulations, place the AA-MAS closer to the general assessment than the AA-AAS, because both are designed to measure grade level standards, but some state policymakers apparently see the AA-MAS as a true intermediary between the AA-AAS and the general assessment. Again, it is important for the state to explicitly articulate these connections.

The purposes should be conceptually coherent with the state's guiding philosophy. For example, if the state is interested in developing the AA-MAS so that the targeted students can "better show what they know", it would lead to one type of argument and theory of action. Whereas, if the state implemented an AA-MAS in order to better align the assessment with the current learning opportunities and beliefs about how eligible students learn, it would lead to another type of validity evaluation. More perversely, some states could be implementing an AA-MAS to ease accountability pressures on schools associated with the performance of students with disabilities. However, it is doubtful that such states will be explicit about these sorts of goals.

Uses follow, in terms of the validity argument, from the state's guiding philosophy and purposes. In New York's case, the results of the AA-MAS will be used to determine students' achievement levels for the accountability system, particularly for AYP determinations. The Board of Regents and the New York State Education Department will have to decide whether and how the results of the AA-MAS will be used for graduation determinations, particularly in terms of eligibility for a Regent's diploma. However, NYSED would like these assessments to have some instructional value as well. These potential uses — discussed in considerable detail in the next chapter (see Domaleski, Chapter 10, this volume) — have significant implications for

the evaluation of the validity of the AA-MAS. If the scores from the AA-MAS are to be treated as comparable for the purposes of a Regent's diploma, then certain types of comparability studies should be incorporated in the validity evaluation (see Abedi, Chapter 8, this volume for an extensive treatment of comparability). On the other hand, if participation in the AA-MAS shuts off the opportunity for a Regent's diploma, an evaluator should consider certain types of studies to examine the unintended negative consequences.

A Theory of Action: The Starting Point for an Interpretative Argument

Katherine Ryan (2002) and others have suggested that having state leaders (or other assessment stakeholders) lay out a more general "theory of action" can be a useful starting point for developing a more complete interpretative argument. This theory of action is really a simplified interpretative argument that requires the explication of the intended components of an assessment and decision system as well as the mechanisms by which a test user could reasonably expect to get from one step to the next. Developing a theory of action for any validation, evaluation, or test development activity is a useful exercise. Given the field's lack of clarity around the AA-MAS, a well developed theory of action is perhaps even more critical than it might be for other validation initiatives. Policymakers, developers, stakeholders, and technicians should have to very explicitly lay out why they think that implementing an AA-MAS will lead to improved educational opportunities for eligible students. In addition to the "why", they should have to describe the "how" or the mechanisms by which they think that these improved learning opportunities will occur. For example, one might postulate that AA-MAS scores will be more accurate depictions of what eligible students know than general assessment scores so that teachers will be able to provide more appropriate learning opportunities for these students. The evaluator and/or user must specify the mechanism by which these score reports will lead to the anticipated changes in teaching practices, such as targeted instruction and/or more appropriate curricular materials.

Based on two example guiding philosophies presented in Chapter 2 (Quenemoen, this volume), two example theories of actions for a modified assessment system were created to illustrate how these differences could play out as different validity arguments. These examples were purposefully created to represent two quite different guiding philosophies and approaches to the AA-MAS.

Example # 1: The AA-MAS allows eligible students to show that what they know may be comparable to similar performance levels on the general assessment.

1. Academic content standards are the same as for the general assessment, and the test blueprint for the AA-MAS is essentially the same as that for the general assessment, but contains some modifications (e.g., fewer passages) to make adjustments for students' disabilities and includes slightly less difficult items than on the general assessment.
2. The achievement standards incorporate recognition of students' disabilities (e.g., need for supports) and while they signal high expectations for eligible students and their teachers, they are slightly lower than the general assessment achievement standards.
3. The assessment is designed to measure grade-level content and high achievement expectations, accurately allowing students to show what they know as well as what they do not know and are able to do.
4. Teachers provide instruction that is aligned with these high academic expectations and ensure that students get the supports necessary allowing them to succeed with grade-level content.
5. The test and achievement descriptors signal and reinforce appropriate instructional and formative assessment strategies for use in classrooms/schools.
6. Student scores on the AA-MAS provide a more accurate estimate of what eligible students know and can do compared with the general assessment.
7. Student performance on the test is used by teachers and school leaders to help them figure out how to provide more appropriate supports and programs.

8. Improved student/school performance on the AA-MAS leads to higher accountability scores.

Example #2: The AA-MAS will better align with current learning opportunities and beliefs about how eligible special education students learn grade-level academic content.

1. Academic content standards are the same as for the general assessment, but the test blueprint for the AA-MAS focuses on fewer and generally easier items tailored to the lower expectations held for these students. The blueprint and test specifications also contain some modifications (e.g., fewer passages) to make adjustments for students' disabilities.
2. The achievement standards incorporate references to students' disabilities (e.g., need for supports) and are designed to describe eligible students' knowledge and skills relative to their current learning opportunities.
3. Teachers provide instruction that is designed to take students from where they are and then helps the students make progress in this curriculum even if it is below grade level.
4. The assessment is designed to provide measurement information about where students are performing, relative to grade-level content, to better show what they know and are able to do.
5. The test and achievement descriptors signal the appropriate levels and types of instructional and formative assessment strategies for use in classrooms/schools.
6. Student performance on the test is used by teachers and school leaders to support (validate) current supports and programs.
7. The AA-MAS scores provide information about students' current performance to the student, parents, and teachers.
8. A test more aligned to students' instructional levels leads to more proficient students with disabilities and higher accountability scores.

9. Students, in part because of these lower expectations, do not make progress on grade-level standards relative to their same grade peers and certain opportunities are shut off from these students by virtue of these missed (or denied) opportunities (e.g., a Regent's Diploma).

Each aspect of the theory of action leads to claims or propositions that are the basis of the interpretative argument. For example, a proposition such as, “*students of teachers using formative assessment strategies aligned with the AA-MAS targets have higher scores than students of teachers using formative assessments not matched with the AA-MAS targets*, could be specified from the general claim found in the first example theory of action presented in Figure 9-1, “*the AA-MAS reinforces appropriate instructional and formative assessment strategies for use in classrooms/schools.*” An interpretative argument will start with one or more of the goals and guiding philosophy discussed above and then trace the claims of the AA-MAS that results in meeting that goal. Specifying a theory of action is a useful first step in creating a more complete interpretative argument. Sample theories of action were developed in the form of pictures shown in Figures 9-1 and 9-2. However, a theory of action, particular when laid out graphically as in the examples here, is of limited utility. It is necessarily quite broad — perhaps superficial—and therefore on its own, cannot guide a comprehensive validity evaluation. Evaluators must “zoom in” on specific components and linkages within the theory of action in order to explicate the propositions/assertions that form the basis of the interpretative argument. Examples of such propositions are presented below in the evidence section. Further, when test users (e.g., states) and developers create theories of action, there is often little emphasis on negative, unintended consequences. Example #2 above was created to illustrate the importance of searching for and trying to uncover negative, unintended consequences, but evaluators should adopt this stance for any interpretative argument and validity evaluation plan.

Figure 9-1. Example #1 Theory of Action

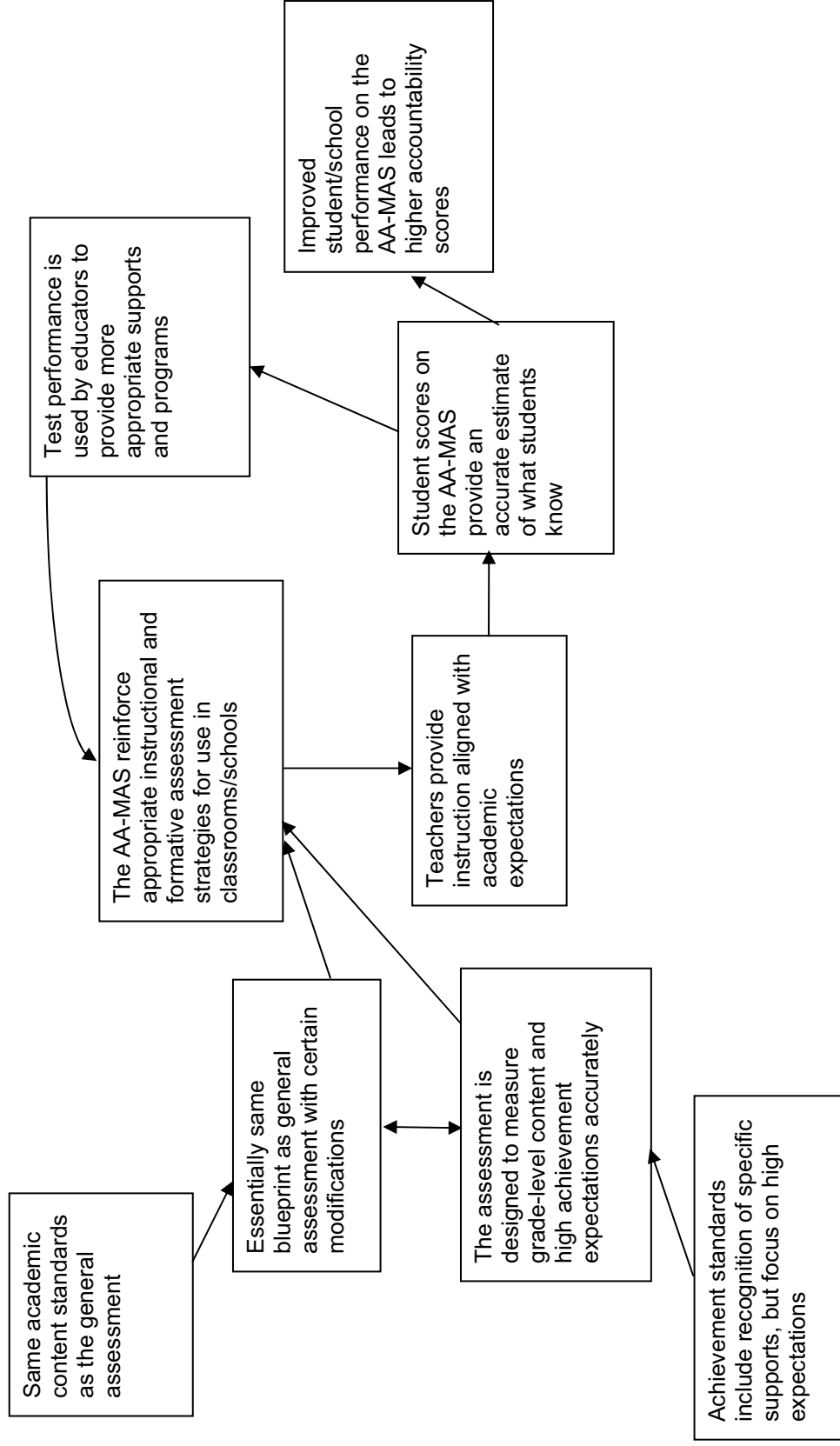
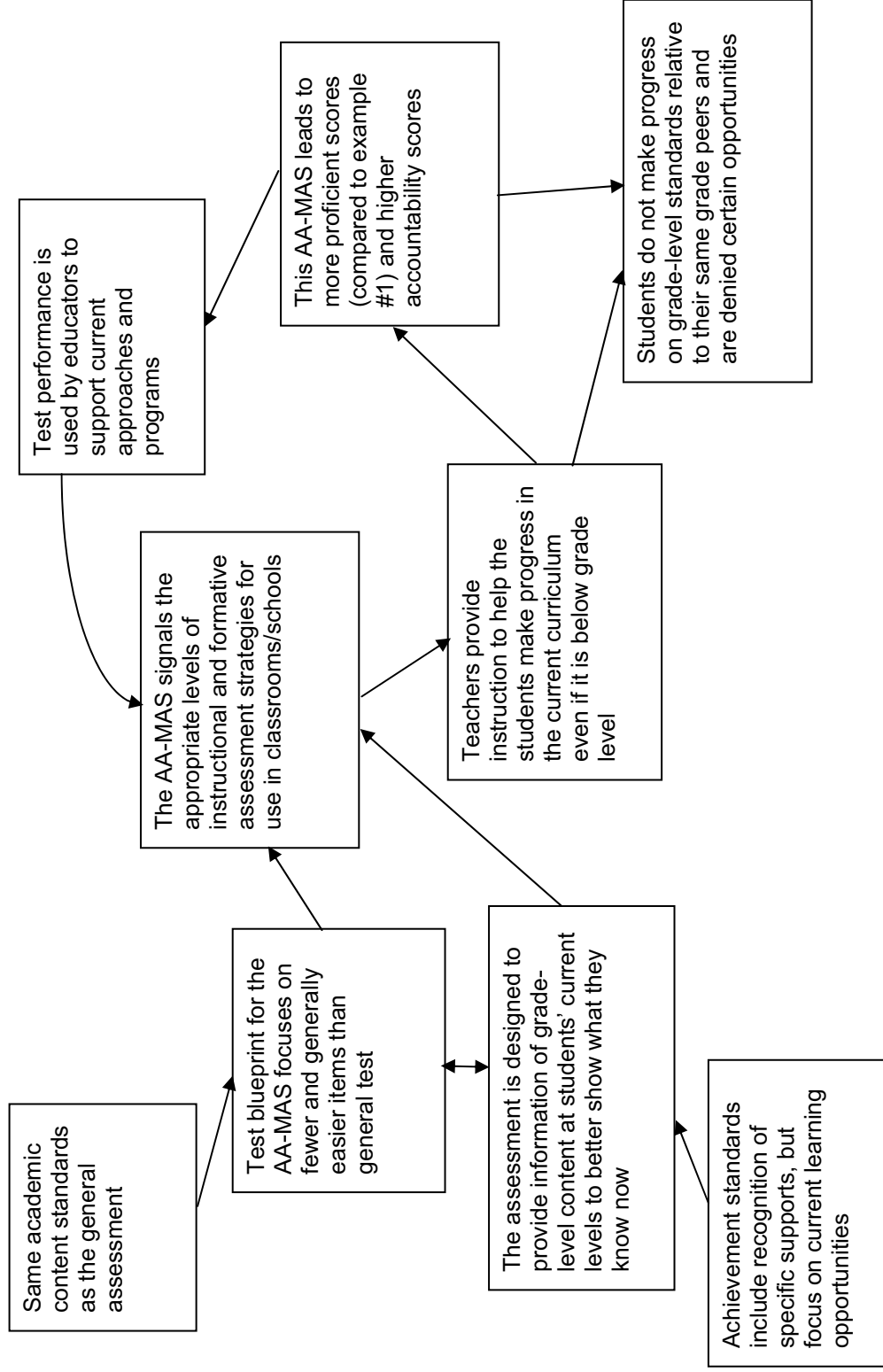


Figure 9-2. Example #2 Theory of Action



Both examples separate out the various claims by the stage of the assessment or accountability process. Both of these theories of action start with the purposes of the assessment, move to content and achievement standards, and then to assessment development (e.g., test blueprint), and end with claims about uses and consequences of the scores. The end result is the goal of increasing student achievement or at least test/accountability scores. An interim goal is to provide information to the teachers to help them improve how they structure learning opportunities for these students. Importantly, these two theories of action lead to different social justice claims, which have implications for the collection and evaluation of consequential evidence.

Prioritizing the Validity Evaluation Questions

The interpretative arguments and the more general theories of action lead to many possible evaluation questions — almost always more than can be addressed in a validity evaluation constrained by time and/or resources. The prioritization should be influenced by the particular guiding philosophy. Following Kane (2006), the state should not select questions and design a validity evaluation to confirm their guiding philosophy. Rather, the validity evaluator should purposefully design studies to contradict the states' beliefs and claims.

While being wary of potential bias, the state can use the guiding philosophy to help prioritize the multitude of possible evaluation questions. A state that adopts a guiding philosophy similar to example #1 should certainly prioritize validity questions addressing comparability of inferences (again, see Abedi, Chapter 8, this volume, for more detail). The evaluator, in this case, should search for proof of concept cases where well-instructed students do in fact perform at levels comparable to students participating on the general assessment. The absence of such cases would be a threat to the guiding philosophy and validity argument found in example #1. On the other hand, a state subscribing to the philosophy articulated in example #2 would have to focus on content validity studies to document that the test actually meets the regulatory

requirements of being on grade level. This evaluation should also collect consequential evidence about students' opportunities to learn meaningful grade level content and skills.

Classes and Sources of Evidence

There are many ways to organize and collect evidence for the validity evaluation. The joint *Standards'* (AERA, et al., 1999) five sources of evidence are the most familiar organizing framework. Earlier work (Marion & Pellegrino, 2006; Marion & Perie, in press) has illustrated how both the assessment triangle (Pellegrino, et al, 2001) and Ryan's (2002) framework could be used to structure validity evaluations. The joint standards are used as the basis here because of both their familiarity and straightforward structure. However, the current (1999) version of the joint standards does not do justice to certain key elements illuminated by the assessment triangle (Pellegrino, et al., 2001), particularly related to the "cognition" vertex of the triangle. Further, the 1999 edition of the joint standards does not fully incorporate recent research making clear the central role of test consequences into validity evaluations (e.g., Lane & Stone, 2002; Shepard, 1997). Therefore, an introductory section was added to this discussion to address "who are the students?" and "how do they acquire proficiency in the domain?" to supplement the joint standards framework. While this type of information should be part of any validity evaluation, it is even more important in alternate assessment and English language learner testing contexts where the specific tested population could vary considerably depending on the selection rules employed. Further, the framework presented here prioritizes the role of test consequences in the evaluation of AA-MAS validity more than the joint standards would suggest. Within each of the following categories, the sources of evidence and types of studies particularly relevant to evaluating the validity of the AA-MAS are described. Several examples are presented throughout the following sections illustrating how specific propositions and study designs might differ depending on the specific guiding philosophies and theories of action.

Who Are the Students and How Do They Learn?

As Quenemoen (Chapter 2, this volume) makes clear, identifying students for participation in the AA-MAS is a complex endeavor. A key eligibility requirement is that students must be instructed in the grade-level curriculum and have an opportunity to learn grade-level content (Quenemoen, Chapter 2, this volume). A major premise associated with implementing an AA-MAS is that students' disabilities interact with their capacity to demonstrate what they know and are able to do and that poor performance is not due to a lack of opportunity to learn. Karvonen (Chapter 3, this volume) discussed methods for documenting the effectiveness of instructional and curriculum strategies. This documentation is crucial evidence to help make the case that students have been appropriately selected to participate in the AA-MAS. Further, IEP teams need to ensure that appropriate supports and strategies are provided so that students have the highest likelihood possible to access the grade-level knowledge and skills.

Pellegrino (Chapter 4, this volume) provides a thorough and excellent discussion about the ways in which students acquire competence in a domain, with a specific focus on mathematics. Pellegrino's exposition is very important for states to keep in mind as they consider developing an AA-MAS, because if state leaders do not have a sense of how eligible students will make progress in the domain, then the rationale for and the validity of the AA-MAS will be suspect. Therefore, a critical aspect of the interpretative argument is the development of propositions related to the way in which students develop domain competence. The theoretical conceptions and the associated evidence—such as the results from tasks specifically designed to measure students' progress along a defined learning continuum—should be evaluated as part of the larger validity investigation for any assessment system, but even more so for the AA-MAS because of the field's limited understanding of the conceptual underpinnings of this assessment.

Evidence Based on Test Content

Important validity evidence can be obtained from an analysis of the relationship between a test's content and the construct it is intended to measure. Test content refers to the themes, wording, and format of the test items, tasks, or questions on a test, as well as the guidelines for procedures regarding administration and scoring (AERA, et al., 1999, p.11).

One of the foundational principles of the AA-MAS is that it is based upon “grade-level” content. Therefore, collecting and evaluating the evidence regarding comparability of the content is critically important to evaluating the validity of the AA-MAS. Many states and evaluators will often use evidence from alignment studies to support claims of content validity. Well done alignment studies can certainly contribute to content related validity evaluations, but alignment studies generally focus on matching test items with content-based standards and objectives. Content-related evidence, especially when one is trying to make claims about “grade levelness” requires evaluating the interaction of both content and process required of the test items and, in the case of the AA-MAS, documenting that the interaction is what is expected for the specific grade level.

In both example theories of action presented earlier, the assessments are based on the state's academic content, but the blueprint described in the second example is based on fewer and easier items than the general assessment. In this case, the evaluator should critically evaluate the assertion that *the test blueprint used in Example #2 accurately represents the construct* even though a purposeful non-representative item sampling (of grade level content) approach is used. Even in Example #1, studies should address the assertion that *the use of certain changes (modifications) to the test blueprint (and items) accurately represents grade-level knowledge and skills as indicated by the content standards.*

Evidence Based on Response Processes

Theoretical and empirical analyses of the response processes of test takers can provide evidence concerning the fit between the construct and the detailed nature of performance or response actually engaged in by examinees (AERA, et al., p 12).

Many validity questions emerge from a state's belief that implementing an AA-MAS will better allow students to show what they know can be grouped under response processes. These types of studies would be applicable for guiding philosophies aligned with either example #1 or #2, but the orientation would be different depending on the underlying beliefs. Several studies can be identified where students' response behaviors are compared between AA-MAS and general assessment items with an attempt to attribute the differences to specific theoretical conceptions outlined in the rationale for the AA-MAS.

Evidence related to the cognition vertex of the assessment triangle can also be considered within the responses process category. Before analyzing evidence on *how* students are responding to specific tasks, it is crucial to describe and analyze *which* students have been nominated to participate in the AA-MAS. There is an implicit assumption in both theories of action that the "right" students are participating in the AA-MAS—an assumption that should be made explicit in a more complete or elaborate theory of action—but this assumption should be evaluated before investigating how students are responding to the items. States should have (and present) a theoretically-grounded rationale as part of the description of the students participating in the AA-MAS.

Another important dimension of the cognition vertex subsumed by the response process category is a description of how students acquire competence (proficiency) in the domain. If there is such a hypothesized progression by which students are expected to develop domain competence, the evaluator/state should describe how students eligible for the AA-MAS are expected to follow the same expected progression or how and why they would develop

differently than their same-age peers. The tasks and associated response processes could then be evaluated against hypothesized learning progressions.

Evidence related to response processes is often collected through the use of cognitive laboratories (“think-alouds”) to get a micro look at how students are interacting with the items and tasks (e.g., Ericsson & Simon, 1980; Johnstone, Bottsford-Miller, & Thompson, 2006). The data derived from well-designed cognitive laboratories can shed light on students’ developing understanding in the grade-level content as a way to ascertain whether the items and tasks on AA-MAS support this developing understanding.

In the case of the AA-MAS, it will be important to determine whether students interact as intended with the modified test items and in ways that differ from the non-modified test items. *Students interact with passages and test items on the AA-MAS in ways that allow them to demonstrate their grade-level knowledge and skills while minimizing construct irrelevant influences* is a proposition that would fit both theories of action. The main difference in how this assertion might be tested in the two examples would play out in the different passages, items, and tasks.

Internal Structure

Analyses of the internal structure of a test can indicate the degree to which the relationships among test items and test components conform to the construct on which the proposed test score interpretations are based (AERA, et al., p. 13).

For states with a guiding philosophy similar to example #1, this section is critical for evaluating the validity of the AA-MAS as an important set of evidence in terms of score comparability. The internal structure of the AA-MAS should be similar to the internal structure of the general assessment or if not, there should be an explicit reason why the internal structures of the two assessments differ. As discussed by Abedi (Chapter 8, this volume), meeting strict comparability criteria (i.e., equating) is generally beyond the reach of almost any AA-MAS

design. Yet, techniques such as confirmatory factor analysis could be used to compare the internal structure of the general and modified assessments to determine if the structure of the modified assessment is “close enough” to the general assessment to argue that both are tapping the same construct. A proposition from the perspective of Example #2 might suggest, the internal structure of the AA-MAS is generally similar to that of the general assessment, while one from the perspective of Example #1 would argue for stronger comparability such as, the same factor structure can be used to explain the variability of the items on both the AA-MAS and general assessment.

Evidence Based on Relations to Other Variables

Analyses of the relationship of test scores to variables external to the test provide another important source of validity evidence. External variables may include measures of some criteria that the test is expected to predict, as well as relationships to other tests hypothesized to measure the same constructs, and tests measuring related or different constructs (AERA, et al., p. 13).

This section is probably less critical for the AA-MAS compared with other sources of evidence, but it can still be important—depending on one’s theory of action—to substantiate claims about the AA-MAS. There are other assessments with documented properties (e.g., grade level or not; difficult or easy, accessible or not) that should be more or less related to scores on the AA-MAS. Since psychometricians are quite good at computing correlations, state leaders and evaluators should articulate the intended relationships *a priori* instead of data snooping for relationships that support one’s conclusions.

Assuming there is an attempt to ensure that the AA-MAS is measuring the same construct as the general assessment, it is difficult to imagine significant differences in the relationship to some external test or other variable. If the state had good longitudinal data on norm-referenced tests or interim assessments, for example, the state might want to put forth a

proposition for Example #2 to gather validity evidence that supports the claim that the AA-MAS is on grade level. Such proposition might state, *the fourth grade AA-MAS is significantly more related to the fourth grade NRT than it is to the 3rd grade NRT*. This proposition could be extended to argue that the correlations between the AA-MAS and the external criterion should be very similar to the correlations between the general assessment and the external test.

Evidence Based on Consequences of Testing

There are a host of intended positive consequences associated with a state's interest in implementing an AA-MAS, but there are some serious potentially unintended negative consequences. As discussed above, states that have an orientation similar to that in example #2 should focus consequential questions on potential lower expectations that could hinder eligible students from achieving at grade level. A state's approach does not have to be as extreme as presented in example #2 for the system to carry unintended negative consequences, therefore state leaders and evaluators need to attend to unintended consequences related to lower expectations for any AA-MAS. On the other hand, states with a philosophy similar to example #1 might address consequential issues related to frustration and/or lack of a meaningful assessment experience from "unrealistically" high expectations. In any case, consequential studies related to the validity of the AA-MAS need to focus on, in large part, searching for and evaluating the potential unintended consequences of an AA-MAS such as lower expectations for students with disabilities.

The AA-MAS was originally conceived as part of the flexibility offered by the U. S. Department of Education under NCLB and ultimately this assessment has been designed to fit into states' accountability systems and contribute to Adequate Yearly Progress (AYP) determinations (see Domaleski, Chapter 10, this volume, for more discussion of AA-MAS accountability issues). The accountability function makes clear that the AA-MAS has been designed, at least as one purpose, as a policy instrument. As Kane (2006) noted, when

assessments are used to support a particular policy, the consequences of such policy actions must be incorporated into the validity evaluation. A range of validity evaluation questions and propositions could be put forth to collect consequential evidence related to the AA-MAS. These questions and propositions will differ depending on the philosophies and goals guiding the development and implementation of the AA-MAS. For instance, a proposition to search for potential unintended negative consequences based on Example #2 might read as follows: *the increase in the percentage of special education students scoring proficient as a result participating in the AA-MAS has not led to an increase in schools falsely meeting AYP targets* (Type II errors).

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. The evidence and logic must be synthesized to evaluate the interpretative argument. 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 unpredictable 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. This evolving understanding will require evaluators to (re)examine evidence in light of these newer understandings.

With the exception of a few states, most AA-MAS are in the very early stages of development. Therefore, initial syntheses could adopt confirmationist biases during the first few years of the program until it gets established. This does not mean that long-term studies,

especially consequential, should not be planned and initial data collected, but the synthesis and evaluation in the early years of the program should focus on substantiating that the development of the AA-MAS has generally occurred as designed and the designs can be theoretically supported.

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 that the state leaders are left only with the option of starting over.

The state should use the guiding principles and purposes of the AA-MAS to determine how to weigh various sources of evidence to arrive at an evaluative judgment. This judgment could take the form of a summative judgment where a state determines that the overwhelming evidence suggests abandoning the AA-MAS or going ahead with it full steam ahead. More likely, however, the state will use the initial validity evaluation in formative ways to improve the AA-MAS.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed., pp. 443–507). Washington, DC: American Council on Education.
- Elliott, S. N., Compton, E., Roach, A. T. (2007). Building validity evidence for scores on a state-wide alternate assessment: A contrasting groups, multimethod approach. *Educational Measurement: Issues and Practice*, 26(2), 30–43.
- Ericsson, K. & Simon, H. (1980). Verbal reports as data. *Psychological Review*, 87, 215–250.
- Gong, B. & Marion, S. F. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities* (Synthesis Report 60). Minneapolis, MN: University of Minnesota, National Center for Educational Outcomes.
<http://education.umn.edu/nceo/OnlinePubs/Synthesis60.html>.
- Haertel, E. H. (1999). Validity arguments for high-stakes testing: In search of the evidence. *Educational Measurement: Issues and Practice*, 18, 4, 5–9.
- Johnstone, C. J., Bottsford-Miller, N. A., & Thompson, S. J. (2006). *Using the think-aloud method (cognitive labs) to evaluate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved [today's date], from the World Wide Web: <http://education.umn.edu/NCEO/OnlinePubs/Tech44/>
- 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. National Alternate Assessment Center, Human Development Institute, University of Kentucky, Lexington.
http://www.naacpartners.org/Products/Files/Research_Focus_LCI.pdf
- Kleinert, H., Browder, D., & Towles-Reeves, E. (2005). *The assessment triangle and students with significant cognitive disabilities: Models of student cognition*. National Alternate Assessment Center, Human Development Institute, University of Kentucky, Lexington.
<http://www.naacpartners.org/Products/Files/NAAC%20Assmt%20Triangle%20White%20Paper%20-%20FINAL%20for%20Website.pdf>
- Lane, S., & Stone, C. A. (2002). Strategies of 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.
- 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.
- Marion, S. F. & Perie, M. (2009). Validity arguments for alternate assessments. In Schafer, W. and Lissitz, R. (eds.) *Alternate assessments based on alternate achievement standards: Policy, practice, and potential* (pp. 115-127).
- 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.
- No Child Left Behind Act of 2001, Pub. L. No.107-110, 115 Stat.1425 (2002).
- 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.
- Rabinowitz, S. & Sato, E. (2005). *The technical adequacy of assessments for alternate student populations*. San Francisco: WestEd.
- Ryan, K. (2002). Assessment validation in the context of high stakes assessments. *Educational Measurement: Issues and Practice*, 21(1), 7–15.
- Schafer, W. D. (2005). Technical documentation for alternate assessments. *Practical Assessment Research & Evaluation*, 10(10).[Available online: <http://pareonline.net/getvn.asp?v=10&n=10>].
- Shepard, L. A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.), *Review of Research in Education*, 19, 405–450.
- Shepard, L. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice*. 16(2), 5–24.

CHAPTER 10

OPERATIONAL AND ACCOUNTABILITY ISSUES

Chris Domaleski

A full examination of the issues and elements related to the design and adoption of a new state assessment program would not be complete without careful consideration of the context in which the program will be situated. It is important to acknowledge that an alternate assessment based on modified achievement standards (AA-MAS) would exist as part of a larger state assessment and accountability system. Therefore, it is essential to understand the interrelationship of the AA-MAS with other assessment programs. Moreover, the potential impact of the AA-MAS on the state accountability system should be carefully explored.

This chapter begins with an overview of the background and context for accountability and then summarizes the key provisions in the United States Department of Education's (USED) regulations that pertain to accountability determinations. This is followed by a discussion of the relationship of the AA-MAS to existing New York State Education Department (NYSED) assessments. Subsequently, specific accountability issues are addressed to include procedures to estimate reliability, and a review of key operational considerations. The chapter ends with a discussion of factors related to student and summary reporting, and a consideration of issues and options related to diploma eligibility.

In exploring these topics, focus will be placed on practical, technical, and policy elements. By so doing, the goal will be to highlight options and provide guidance to assist with implementation and evaluation.

Background and Context for Accountability

Education accountability systems, in some form or another, have been in place for at least the previous three decades. However, earlier accountability systems tended to focus on areas such as regulation compliance and financial management (Fuhrman, 2004). The change

in focus to outputs, chiefly, student performance on standardized assessments, began in earnest in the 1980s. During this time, accountability approaches drawn from business applications gained support from education policy makers (Fuhrman, 2004). This was bolstered by a wave of concern about the perceived decline in quality of education as described in the influential publication *A Nation at Risk* (1983). In subsequent years, accountability systems expanded and focused more on student and school performance.

Another major influence on contemporary education accountability began in the 1990s with increased support for standards-based reform. The guiding idea behind this approach is that expectations for what students know and can do should be clearly established, which will guide all other elements of the educational system, chiefly instruction and assessment (O'Day & Smith, 1993). Advocates argue that such an approach leads to a number of improvements such as clarifying goals, incentivizing improvement, and informing allocation of resources (Darling-Hammond, 2006). This perspective was a guiding factor behind the development and implementation of accountability systems in the 1990s and in the current decade, including the federal *No Child Left Behind* (2001) legislation.

As support increased for standards-based reform, so too did advocacy for students with disabilities. Historically, many educators and stakeholders did not provide students with disabilities access to the general curriculum. With recent reauthorizations of IDEA and NCLB, the view that students should be taught and held accountable for grade-level standards prevailed. This position has not been without opposition, from those that argued that such goals are unreasonable and/or traditional standardized assessment practices are ill-suited for students with disabilities.

Today, a central idea behind contemporary accountability practices is the inclusion of all students, including students with disabilities. This is based on the belief that measuring, reporting, and holding schools explicitly accountable for the performance of students with disabilities is critical to ensuring that educators attend to their needs, provide appropriate

resources, and set high expectations for learning. The extent to which this principle holds rests largely on the integrity of the measures used to gauge student achievement. This is the context that has inspired the state of New York, like many other states, to explore the efficacy of customizing a standards-based assessment for a portion of the population of students with disabilities.

Federal Regulations

Against this backdrop, the United States Department of Education (USED) issued regulations and guidance in April of 2007 that addressed the implementation of modified academic achievement standards and assessments. These regulations were explicitly targeted to a small group of students whose disability precludes them from achieving grade-level proficiency within the year. A more complete overview of the regulations is presented in Chapter 1 (Perie, this volume). The focus of this section will be to review the elements that directly impact accountability determinations.

In terms of accountability, there are two main elements of the policy that merit attention. First, the regulations and guidelines establish that states may count as proficient for the purpose of AYP calculations, the proficient and advanced scores of students with disabilities based on an AA-MAS, provided the number of these scores do not exceed 2% of all students in the grades assessed in language arts and mathematics. In other words, scores on the AA-MAS can be used in adequate yearly progress (AYP) calculations in the same way as scores from the general assessments within the 2% cap. While this seems straightforward, there are a number of caveats and considerations that warrant further examination to fully appreciate the application of this stricture. This will be addressed in a later section of this chapter.

The second major element of the policy with respect to accountability is the expiration of the 'interim-flexibility' policy. Interim-flexibility refers to the practice of allowing states that meet certain criteria to count as proficient for purposes of AYP a portion of the students with

disabilities. This applies at the school or district level if AYP is missed solely because of the achievement of the students with disabilities subgroup. The portion is determined by dividing 2 percent by the percent of students with disabilities in the state. For example, in the State of New York the SWD subgroup is about 12 percent of the student population. Dividing 2 by 12 equals 17 percent. Therefore, 17 percent of the state's SWD population could be counted as proficient for purposes of AYP, where applicable.

The purpose of this flexibility was to forestall the impact of non-proficient classifications based on general assessments for students who may be candidates for an AA-MAS, during the time that new assessments more appropriate for this population, were under development. Importantly, the interim-flexibility, which was initially granted for the 2004–05 academic year is extended through 2008–09 in the regulations; however, it expires beginning in the 2009–10 academic year. Whether an AA-MAS is developed or not, this will have an impact on accountability determinations in the state of New York, which, like many states, has applied the interim-flexibility in AYP computations. The interim-flexibility essentially allows states to count the *maximum* percent of eligible students proficient in AYP computations. Consequently, when this expires, states will likely see an increase in the number of SWD groups that fail to meet their annual measurable objectives (AMOs).

Relationship to Existing Assessments

The state of New York has developed a comprehensive assessment system to measure student achievement of the New York State Learning Standards and to satisfy the accountability provisions of *No Child Left Behind* (NCLB). Assessments used in the accountability system are part of the New York State Testing Program and include English/ Language Arts (ELA) in grades 3-8, mathematics in grades 3-8, and science in grades 4 and 8. At the secondary level the Regents English Comprehensive Exam and the Regents Integrated Algebra Exam are used

for AYP purposes. Moreover, the department has developed the New York State Alternate Assessment (NYSAA) for students with significant cognitive disabilities.

The NYSED follows a development and validation process in keeping with professional standards, partnering with assessment specialists, contractors, educators, and stakeholders. Each item on the assessment is mapped to a performance indicator that is consistent with the state curriculum. The elementary and intermediate ELA assessments consist of multiple-choice and short and/or extended-response items. Some assessments also include an editing paragraph. The Regents Comprehensive English Exam contains multiple-choice items based on passages and stimuli, including a listening portion, as well as a constructed-response writing prompt. The 3–8 and Integrated Algebra assessments also include multiple-choice and constructed-response items, requiring students to generate item responses and show work. While the Regents Examinations are given at various times throughout the year, the 3–8 assessments are typically administered in early spring term. Students take the exams in sections or books over two to three days.

The NYSAA is a datafolio assessment that measures the achievement of students with significant cognitive disabilities. The datafolio is a collection of evidence in response to aligned tasks, evaluated with respect to accuracy and independence, intended to provide information about the student’s achievement. NYSAA tasks are aligned to Alternate Grade Level Indicators (AGLIs) which are entry points to grade-level expectations in the New York State learning standards.

Grades and Content Areas for AA-MAS

An important decision for the NYSED is the determination of the grades and content areas in which to implement an AA-MAS. There is no regulatory requirement to develop or adopt an AA-MAS, so the potential implementation options range from none to all. That is, the NYSED may decide not to proceed with an AA-MAS in any area or to pursue full adoption in all

grades and content areas assessed, regardless of inclusion in NCLB accountability. Naturally, a number of implementation options in between these two extremes are available as well.

A decision about scope of development is, foremost, a policy decision that should be guided by the goals of the NYSED and the purpose for considering an AA-MAS. Assuming it is desirable to implement an AA-MAS as broadly as possible, there are at least three possible perspectives that might guide prioritization of implementation.

First, the extent to which the general assessments are seen as valid and appropriate for students with disabilities could be a guiding principle. By carefully evaluating both the assessment characteristics and student performance, the state might develop priorities for the grades and content areas that should be given primary consideration. For example, one may wish to review blueprints and specifications for the general assessments to determine which are relatively more cognitively complex and/or rigorous. Moreover, one may wish to review the gap between performance of students with disabilities and general education students, and focus on the assessments that have the largest gap. When these two approaches identify the same assessments, a more compelling case for prioritizing these assessments may be made.

Additionally, there may be legal issues to consider. If a general assessment is regarded as not suitable for students with disabilities, the state may be legally compelled to pursue the development of an alternate assessment. This position was supported by *Chapman v. California Department of Education* (2002) in which a federal court ruled that the state of California must provide an alternate assessment if it is determined that students with disabilities are unable to access the general assessment due to their disability.

A second approach may be to allow the consequences or stakes associated with the assessment to guide prioritization of the grades and content areas in which an AA-MAS should be developed. Using this orientation, those areas covered in the state accountability system (ELA and mathematics in 3-8 and high school) may be given higher priority. There may be other stakes, either currently in place or planned, that could guide this decision. These may include

student stakes, such as diploma eligibility, or rewards/consequences at the teacher, school, or system level.

A third lens through which to view this decision is related to practical or operational constraints. Unavoidably, the availability of resources, such as cost and staff capacity, has a significant impact on options that can be considered. Such factors as the format of the assessment, the frequency of administration, or the scope of ongoing development and support, may make some options more feasible than others.

It is important to acknowledge that these three approaches are not mutually exclusive and most likely will interact with each other. For example, assuming resources are limited, the NYSED may get a sense for the scope of implementation which may narrow options down to a specific program or grade span. Thereafter, it may be reasonable to consider the policy implications, then, review the properties and performance of the assessments to further identify the area in which to begin implementation. Another important consideration is whether or not the state would like to use scores on the general assessment to inform placement on the AA-MAS. If so, then it will be important to introduce the AA-MAS at a later grade to acquire score(s) on one or more years of the general assessment.

The state of New York may approach this decision as a cost-benefit analysis. The costs of implementing an AA-MAS are related to finances, operational burden to state and local staff, and possible forfeiture of other programs and initiatives that could be supported by these resources. On the other hand, the benefits may include improved information from assessment and accountability systems and the ability to promote student achievement for students with disabilities.

Although each state likely differs with respect to a number of the factors previously examined, it may be useful to examine the scope of AA-MAS implementation in other states. In 2007 the National Center on Educational Outcomes (NCEO) reviewed the characteristics of the AA-MAS for six states, including the grades and content areas that were addressed. The results

are presented in table 10-1 below, reproduced from that report. The results show that most states implemented the AA-MAS fairly broadly. Each state included reading and mathematics at the elementary level and all but one state (North Carolina) offered an AA-MAS in these areas and at the secondary level. Many states also implemented the assessment in areas not included in the NCLB accountability system, such as Kansas which developed an AA-MAS for writing and social studies. It is important to reiterate, however, that each state's decision is connected to a unique set of policies and priorities. There is not a uniform or best solution for all states.

Table 10-1 AA-MAS Name, Content Areas, and Grade by State

State	Assessment Name	Content Areas/ Grades
Kansas	KAMM (Kansas Assessment of Multiple Measures)	Reading (3-8; once in HS); Math (3-8; once in HS); Writing (5,8, once in HS); History/Gov (6, 8, once in HS); Science (4,7, once in HS)
Louisiana	LAA2 (LEAP Alternate Assessment, Level 2)	English (Grades 4-10); Math (Grades 4-10); Science (Grades 4, 8 and 11); Social Studies (Grades 4,8,11)
Maryland	Mod-MSA (Modified Maryland School Assessment) and Mod-HSA (Modified High School Assessment)	Reading/ELA (3-8, HS); Mathematics (3-8, HS)
North Carolina	NCEXTEND	Reading (Grades 3-8); Math (Grades 3-8); Science (Grades 5 and 8)
North Dakota	North Dakota Alternate Assessment Aligned to North Dakota Content Standards for Students with Persistent Cognitive Disabilities	Reading (3-8,11); Math (3-8,11); Science (4,8,11)
Oklahoma	CARG-M (CARG=Curriculum Access Resource Guide)	ELA/Reading (Grades 3-8, HS); Math (Grades 3-8, HS); Science (Grades 5 and 8)

Table reproduced from Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Participation Options and Evidence

As addressed in Chapter 1 of this volume, there are five alternatives for assessment participation. These are: 1) participation in the general grade-level assessment; 2) participation in the general grade-level assessment with accommodations; 3) participation in an alternate

assessment based on modified academic achievement standards; 4) participation in an alternate assessment based on alternate achievement standards; and 5) participation in an alternate assessment based on grade-level academic achievement standards (AA-GLAS). The fifth option differs from the AA-MAS in that the AA-GLAS performance expectations must be directly related to those on the general assessment.

The regulations further stipulate that a state must establish participation criteria for IEP teams based on evidence that the student's disability has precluded the student from achieving grade-level proficiency and the student's progress suggests the student will not reach grade-level proficiency during the academic year. Therefore, a key issue with the implementation of an AA-MAS will be the development of guidelines to inform participation decisions and the collection of evidence that meets the criteria described. In previous chapters more detailed information was provided about the guiding perspectives and approaches to identify the population of students that are appropriate for the AA-MAS. In this section, the focus is on the specific, objective data sources and methods that may be considered to inform these decisions.

One approach is to analyze extant assessment data from interim, formative, or summative state assessments or other commercially available standardized assessments. The advantage of using state curriculum-based assessments is that the performance level provides direct evidence of student performance with respect to grade-level expectations. Eligibility criteria may be related to persistent low performance (e.g. failure to achieve proficiency in more than one administration) and/or performance that is well below standard (e.g. performance level one.) This approach is bolstered if the state can produce evidence that the probability of achieving on grade level on the general assessment in the current year is low given performance the previous year. For example, if the criterion selected was level one (e.g., Below Basic) performance on the summative state assessment and only a very small percentage of students scoring at level one go on to score at or above level three in the following year, this signals that the expectation is reasonable. Other commercially available standardized

assessments, such as a norm-referenced assessments, may also be candidates for evidence. For example, regression analyses may be employed to produce a predicted score on the state curriculum assessment for various NRT score values. These data can be analyzed to determine a suitable eligibility criterion that indicates that students below the standard are unlikely to perform on grade level.

Another category of evidence to consider is related to the student characteristics. For example, the state may review performance for students based on disability category to determine which are associated with persistent low performance. The Georgia Department of Education conducted one such study that explored many factors including disability type and revealed that students with mild intellectual disabilities were disproportionately represented (Fincher, 2007). While disability category may not be used as a criterion for participation, such analyses can provide information to better identify the group of students who might benefit from participation in an AA-MAS or to evaluate the extent to which schools and systems are making appropriate participation decisions. These analyses involve two basic elements. First, select a condition that identifies students who are consistently below grade level (e.g. below level 3 on state assessment performance in consecutive years). Second, explore these data for patterns that may provide more information about the group. For example, are there strands or domains within content areas where performance is particularly low? Are students who received certain accommodations disproportionately represented compared to the state as a whole?

It is noteworthy that Georgia's study identified many persistently low-performing students who do not receive special education services. This invites serious consideration as to why these students are not meeting academic achievement standards and to what extent these same factors are applicable to students with disabilities. At least part of the answer is likely to be that instructional approaches and supports for these students have been ineffective.

For this reason, evidence should be collected to document the extent to which students received instruction aligned with the curriculum at the appropriate grade level. Moreover, what

supports and interventions have been in place to promote achievement? In reviewing this information, it is worthwhile to consider how approaches are similar or different for low-performing students with disabilities compared to other similarly performing students. This information may help policymakers disentangle which students are the most prominent candidates for an AA-MAS and which students (both with and without disabilities) may benefit from improved instruction and support strategies.

Another important aspect related to participation options is the establishment of guidelines for students to transition from the AA-MAS to the general assessment. It is possible that students may take the AA-MAS in all content areas or take the AA-MAS in selected content areas and the general assessment in others. Given that placement decisions need to be made annually, guidelines for transition should be developed that are informed by appropriate evidence.

One way to accomplish this goal is to establish a policy based on a specific score on the AA-MAS. For example, students scoring at the advanced performance level may automatically move out of the AA-MAS to the general assessment the following year. In Chapter 8 (Abedi, this volume), the topic of establishing comparability between the assessments is explored. The extent to which there is an explicit, quantifiable relationship between the assessments using the techniques discussed will guide the decision. Such evidence should indicate that the AA-MAS can produce a grade-level achievement indicator that is explicitly and demonstrably comparable to proficiency on the general assessment. This should be based on the extent to which both the content and performance expectations are comparable. Examples of evidence might include: comparison of the distribution of content standards addressed, including cognitive complexity, between the general assessment and the AA-MAS at the 'exit' standard; performance level descriptors for the comparable achievement levels are designed to closely match; and/or a review of performance data shows that a reasonable number of students who exit the AA-MAS subsequently achieve proficiency on the general assessment.

The use of multiple indicators will strengthen such decisions. For example, a *profile* approach could be implemented that takes advantage of several data sources. Such an approach may involve establishing a number of categories that indicate various conditions under which eligibility to exit may be supported. Examples of such profiles might include 1) scoring at the advanced level on the AA-MAS; 2) scoring between levels 2 and 3 while also achieving a criterion score on a district assessment; 3) achieving a specific level of course performance in tandem with AA-MAS and/or local assessment scores; 4) recommendation from IEP committee etc. These examples are intended to be illustrative and each profile should be carefully developed and monitored to ensure they are reasonable and appropriate.

Accountability System Background

In addition to considering the role of the AA-MAS in the general assessment system, it is also important to consider how adoption of such an assessment will fit into the NCLB accountability system. New York State's NCLB accountability system is authorized by 8 NYCRR §100.2 which states in part, "Each year...the commissioner shall review the performance of all public schools, charter schools and school districts in the State. For each accountability performance criterion specified...the commissioner, commencing with 2002–2003 school year test administration results, shall determine whether each public school, charter school and school district has achieved adequate yearly progress." The code provides a full description of the system, including how AYP is determined and schools are designed as requiring academic progress.

As described in 8 NYCRR §100.2 and consistent with federal requirements, New York State's accountability system is comprised of three main elements: 1) participation rate; 2) academic achievement; and 3) an additional indicator. The participation criterion requires that 95% of students in all applicable subgroups take part in state assessments annually. Academic achievement is measured by yearly performance on state curriculum assessments in ELA and

mathematics for grades 3–8 and high school. This is operationalized by a performance index system that is evaluated with respect to effective Annual Measurable Objectives (AMOs); these will be discussed in more detail later in this section. Finally, performance on science assessments or attendance serves as the additional indicator in grades 3–8 and graduation rate is the additional indicator for high schools. Meeting the overall AYP standard for schools and LEAs is based on all subgroups meeting all criteria. That is, the criteria are considered conjunctively—if any group fails to meet the standard the school does not make AYP.

An essential component of any examination of accountability practices is to clarify the purpose of the system and the underlying theory of action. Because New York State’s system was designed to be compliant with federal regulations, language from the 2001 NCLB Act (20 U.S.C. § 6301) may serve as a guiding statement of purpose, “to ensure that all children have a fair, equal, and significant opportunity to obtain a high-quality education and reach, at a minimum, proficiency on challenging state academic achievement standards and state academic assessments.” Based on this idea of promoting equity and achievement, a theory of action can be shaped for the accountability system. Drawing on a simplified conceptualization proposed by Marion et al. (2002), the essence of such a theory involves the following: 1) accountability policy provides incentives, such as recognition or sanctions; 2) awareness and expectations regarding school performance are heightened; 3) educators and students benefit from resources and development; 4) these factors contribute to an improvement in student achievement.

Against this backdrop, the impact of introducing an AA-MAS into the New York State accountability can be more appropriately assessed. In the best case, the AA-MAS should provide more trustworthy information about student performance to better guide accountability determinations and allocations of resources. As outlined in Chapter 2 (Quenemoen, this volume) this theory is connected to a guiding philosophy that values improved student outcomes, and promotes systems and structures that effectively and consistently support this

objective. To the extent that this occurs, the validity and reliability of accountability determinations should be augmented.

The validity of the accountability system is strongly tied to the design of the system, as well as the intended use of results. The central validity focus is ensuring that the *assessments* used in the model are trustworthy for classifying selected students with disabilities as proficient or not proficient, which was addressed in Chapter 9 (Marion, this volume). If the assessments provide better information than the general assessments, the validity of the accountability model should be improved. However, if the AA-MAS is poorly suited for this purpose (e.g. is used to lower expectations for students with disabilities rather than provide accurate information with respect to achievement) then the validity of the accountability model is threatened. However, it is assumed that the structure and purpose of the accountability system would remain intact if an AA-MAS were introduced. For this reason, the primary focus will be evaluating the extent to which the system continues to function as it is currently designed in a stable and consistent manner. This is primarily an issue of reliability, which will be the focus in this chapter.

Evaluating the Reliability of Accountability Determinations

There are two primary sources of error that impact the reliability of accountability systems: measurement error and sampling error. Measurement error refers to the extent to which individual assessments in the accountability system produce stable and consistent results. This is influenced by variability in the population of students who take a specific administration of the test. Sampling error, on the other hand, refers to variations in the school population from year to year.

The literature related to evaluating measurement error or reliability is fairly well established. Reliability can be defined in practical terms as the degree to which an examinee's performance on a test is consistent over repeated administrations of the same or alternate forms (Crocker and Algina, 1986.) It is possible to evaluate test score reliability using a number

of approaches to include those based in item response theory, generalizability theory, or classic test theory. Drawing from the latter category, 'test-retest' and/or parallel form methods are well-known. As the name implies, test-retest approaches involve administering the same assessment to a group of examinees on two or more occasions. The correlation of scores yields an indication of the *stability* of the measure. Alternately, one can administer forms designed to be parallel to a group of examinees to produce a measure of *equivalence*. A more robust approach involves combining the two methods by administering different (equivalent) assessments to the same group of examinees at two or more points in time to yield an indication of *stability* and *equivalence*. Because this approach is influenced by error related to time and form differences, a strong correlation bolsters evidence for reliability. Still another approach, used more commonly, is to calculate reliability based on internal consistency. This method is attractive due to the practical advantages of obtaining a reliability measure based on a single administration of a single form. There are a number of methods to implement this, but perhaps the most familiar is Cronbach's coefficient alpha. Finally, there is a family of methods based on inter-rater reliability, suitable for assessments involving responses or evidence that must be evaluated by a human rater.

A full discussion of how to operationalize each of these and other approaches to quantify the reliability of an assessment is beyond the scope of this document. The reader is referred to seminal works such Crocker and Algina (1986) and Haertel (2006). Moreover, the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) provides the 'industry standard' conventions for evaluating and reporting the measurement error associated with test scores.

It is important to acknowledge that the appropriate method for calculating reliability may differ depending on the approach that is selected for the AA-MAS. Moreover, many researchers stress the need for new and flexible approaches that are designed to 'fit' the assessment. Gong and Marion (2006) assert, "evaluating the technical quality of alternate assessment systems

requires drawing on existing psychometric and evaluation techniques as well as modifying existing approaches or inventing new ones.” This could include any number of procedures designed to quantify the precision of scores under various conditions, the consistency of raters, and/or the integrity of the scoring process.

The second factor related to reliability of accountability determinations is sampling error. In fact, Hill and DePascale (2002) emphasize that sampling error, “contributes far more to the volatility of school scores than does measurement error.” Sampling error refers to fluctuations in school scores that can be unrelated to actual school performance. For example, a school may receive a more favorable accountability determination compared to the previous year, because the students enrolled were inherently higher performing, and not because the quality of instruction improved. Naturally, sampling error can work to both advantage or disadvantage reported accountability determinations.

Hill and DePascale (2002) present four approaches to evaluate sampling error by estimating the precision or consistency of accountability classifications. The most straightforward method is termed *split-half* and simply involves dividing the data for each school into randomly equivalent halves and calculating the percentage of times the same decision is made for each half. Another method involves taking *random draws with replacement* by repeatedly producing random samples from the schools to evaluate decision consistency. A *Monte Carlo* approach can also be implemented, which involves simulating the distribution of scores and creating randomly generated samples from which classification consistency can be evaluated. Finally, *direct computation*, involves calculating exact probabilities for correct classification by determining the distribution of errors. For an extended treatment on these methods including details on operationalization, the reader is referred to *Determining the reliability of school scores* (Hill & DePascale, 2002).

Arce-Ferrer, Frisbie, and Kolen (2002) also examined the effect of sampling error on year-to-year changes in achievement expressed as proportions (e.g. percent proficient). They

found that about two thirds of the variability in estimates were related to sampling error and about one third could be broadly attributed to intervention effects, systematic errors, measurement errors, and equating error. The authors evaluated error by comparing observed variability in proportions with expected variability for one and two year changes at different performance ranges and group sizes. Expected variability was determined by calculating the error variance of the difference between proportions under a binomial model. These methods could also be applied to study changes in New York State's accountability determinations.

Perhaps no factor impacts sampling error or classification consistency of an accountability system more than sample size. Simply stated, larger subgroups produce more stable and consistent results. As a matter of practice, confidence intervals are often used in accountability systems to both gauge and mitigate the effects of sampling error due to sample size. Confidence intervals are constructed by: 1) determining the standard error for a proportion, where the proportion is the target percent proficient or AMO; 2) multiplying this by a desired level of precision corresponding to a distribution value (e.g. z score); and 3) subtracting this figure from the target value to achieve a range of performance within which values are regarded as not significantly different.

The state of New York incorporates confidence intervals in the accountability system through *effective AMOs*. Effective AMOs are designed to integrate confidence intervals with the Performance Index (PI) in a straightforward manner. To accomplish this, the NYSED has produced tables that indicate for various group sizes, the smallest observed PI that is not statistically different from the AMO (i.e. within the confidence interval.) New York State uses a 90% confidence interval and a minimum n of 30 for academic achievement.

Operational Considerations for New York State's Accountability System

New York State's accountability system, like those of other states, may be said to be indifferent to the source of proficiency. In other words, the system is designed such that

whatever instrument or process is used to determine a student's performance level, the 'gears' of the system should function to produce an accountability outcome without disruption. Presently, performance levels are input from the NYSTP assessments in grades 3–8, Regents Examinations in high school, and the NYSAA for students with significant cognitive disabilities. Each of these assessments classifies a student into one of four performance levels which are incorporated into the system.

The New York State Accountability model does have a unique feature, however, that governs how proficiency determinations are produced. In lieu of percent proficient measures, New York State uses a Performance Index (PI). The PI system involves computing a ratio such that the students scoring at levels 2, 3, 4 and those scoring at levels 3 and 4 only are divided by all continuously enrolled students. This figure is multiplied by 100 to produce the index. For example, if a school has 200 students and 40 of them scored at level 1, 80 at level 2, 60 at level 3, and 20 at level 4 the index would be calculated as: $((80+60+20+60+20)/200) \times 100$ which is 120. The index can range from 0, if all students are at level 1, to 200, if all students are at level 3 or higher. This approach incentivizes student improvement below proficiency by providing a boost to the index value when a student progresses from level 1 to level 2.

One straightforward approach to incorporating the AA-MAS in the system would be to establish four achievement levels corresponding to those of the existing AYP assessments. By so doing, performance from the AA-MAS can be included in the PI in the same manner. However, design decisions may restrict this possibility. For example, if the assessment is determined to produce limited information such that only three levels can be produced, alternatives for adjusting the PI will need to be considered. This might involve eliminating an advanced designation, which should have no computational impact on the index, or eliminating the basic proficient level (i.e. treat levels 2 and 3 like levels 3 and 4 in the PI) in which case the 'partial-credit' advantage of the PI would be reduced. (Understanding, of course, that the real impact is more connected to the rigor of the standard than the nomenclature of the standard.)

Another operational issue to consider is managing the 2% cap. As previously indicated, the 2% cap refers to the upper limit on the number of proficient and advanced scores that a state or district can count toward proficiency in AYP from the AA-MAS; it does not restrict the number or percent of students who may *participate* in the assessment. The state or system may only exceed the 2% proficiency cap if the percent of students assessed on the NYSAA is below 1%. In this manner the 2% can be thought of as a “soft cap” where the 1% is a “hard cap”. That is, the 2% may be exceeded as long as it does not extend beyond the margin the state or system has under the 1% for the AA-AAS. For example, if 0.7% of all students in New York State’s accountability system are counted as proficient on the NYSAA, then as high as 2.3% of students in the accountability system can be counted as proficient on the AA-MAS.

USED policy further specifies that all proficient scores from an AA-MAS that exceed the 2% limit, must be counted as non-proficient in AYP calculations. These scores must be counted as non-proficient for the state, system, school and for each subgroup in which the student is a member. This compels the state to determine which scores will be deemed non-proficient—a process referred to as ‘redistribution.’

In guidance, USED refers to a paper by Martinez and Olsen (2004) which describes four methods to implement redistribution. The first approach is to *randomly assign* non-proficient scores back to schools where any students tested on the AA-MAS. A second method is termed *proportional*. This involves assigning non-proficient scores back to schools corresponding to either the proportion of tested students or the proportion of proficient students at the school. A *strategic* approach is also described, which involves making decisions for each school that maximize the chance that the school will make AYP (e.g. assigning non-proficient scores back to groups that exceeded AMOs such that the outcome is unchanged.) Finally, the authors propose a *pre-determined school cap approach*. This involves determining a limit or formula for each school based on the expected number of student participating in an AA-MAS.

The decision of which approach to implement should be measured against the department's priorities and the inherent advantages and risks of each. For example, the strategic approach may seem attractive because it will likely produce the fewest number of schools not making AYP. However, not only would this method be difficult to implement in an unbiased manner, it may also enable potentially inappropriate AA-MAS participation practices. The random and proportional methods seem straightforward to implement, however these may penalize sound participation practices and do nothing to account for a school that serves a large number or percentage of students with disabilities. An adapted or hybrid pre-determined method may be the most promising approach. This method would require the state of New York to carefully establish the expected participation rate in the AA-MAS for schools and systems, perhaps based on previous enrollment or assessment practices. Then, the state would apply additional scrutiny to the schools that deviated from expectation by the largest margin. Schools that deviated for defensible reasons would be protected, but others may be required to adjust a selected number of proficient scores.

An additional consideration for the state of New York is to decide which scores should be redistributed and how they should be reassigned within the performance index system. Because level 3 and 4 scores are always fully proficient in the index and level 1 scores are always non-proficient, the primary concern is level 2 scores. Essentially, the index treats these as 'partially' proficient. That is, the current value produced by the index is the midpoint between the values that would have been produced if either all level 2 scores were treated as non-proficient or all level 2 scores were treated as fully proficient. For that reason, it seems appropriate to regard these values as one-half (.5) proficient for purposes of redistribution. In this manner, districts could assign the designated number of level 3 or level 4 scores to level 1 or twice as many of these scores to level 2. For example, if a district had to redistribute 10 proficient scores to non-proficient scores they could either select 10 level 3 scores and make them level 1 scores, or they could select 20 level 3 scores and make them level 2 scores.

Similarly, the district could select 20 level 2 scores and make them level one scores. Mathematically, it is inconsequential, as each approach produces the same PI value.

Evaluating New York State's Accountability Determinations

Earlier it was mentioned that the accountability system is in many ways indifferent to the proficiency *input*. This is intended to convey that from an operational perspective incorporating results from an AA-MAS into New York State's model is, with some exception, straightforward. However, this is not to suggest that the accountability *output* is unaffected by the introduction of an AA-MAS. Indeed, a central question remains: how will mixing the results from three tests into a single accountability outcome affect results?

Addressing this question will require some purposeful analyses to understand the impact. A good starting point would be to explore the distribution of students who may be 'candidates' to take the AA-MAS throughout the state. The information in Chapter 2 (Quenemoen, this volume) may be helpful in indentifying the characteristics of interest — such as students with certain disability types or those who persistently perform at the lowest performance level on general state assessments. Using this information, it will be beneficial to determine if the students are distributed uniformly (i.e., most schools enroll a similar percentage) or if the students are clustered in certain districts or schools (i.e., some enroll a high percentage while others enroll few to none). Moreover, are potential AA-MAS students over represented in other subgroups (e.g. racial ethnic groups, economically disadvantaged, etc.)? Previous research suggests that an expected finding will be that candidates for an AA-MAS are disproportionately distributed in systems, schools, and subgroups. This is likely to have the most impact on accountability determinations for those units or subgroups with the highest representation.

A second category of analyses involves exploring the pattern of accountability determinations for subgroups and schools. This can be accomplished prior to implementing an

AA-MAS by modeling or simulating a hypothesized statewide AYP outcome. One approach to implementing this would be to conjecture that the students who scored in the lowest 2% on the general assessments will take the AA-MAS. Then, 'new' determinations can be produced with extant data by introducing conditions such as: 1) assume none of the students scored proficient on the AA-MAS; 2) assume the top 25% scored proficient; and 3) assume the top 50% scored proficient etc. For example, in the third condition, all students scoring above the median in the distribution of scores for the 2% of lowest performance students on the general assessment would be designated as proficient on a hypothetical AA-MAS. Then, 2008 AYP determinations would be calculated with this change and the results would be compared to the actual outcomes. Of particular interest will be a review of results at system, school, and subgroup level to gauge which areas are likely to have the most substantial impact. In the method described, the performance categories can certainly be modified, but serve to illustrate the proposed approach. This method, while not exact, can provide an indication of expected accountability outcomes (if only 'best' or 'worst case' scenarios) to assist the NYSED in understanding and preparing for fluctuations in accountability determinations.

When an AA-MAS is implemented, the NYSED should continue to carefully monitor the consistency of determinations from year to year. Such monitoring at the district, school, and subgroup level can illuminate components of the accountability system that are most volatile. This may involve simply tracking changes in the PI for schools and subgroups and comparing the numbers and percent of schools and groups that make AYP. For schools that do not make AYP, it will be useful to track both the number and type of subgroups that missed the AMO, as well as the margin by which AMO was not achieved.

As discussed in the previous section, confidence intervals are the primary mechanism for dealing with sample variability in the accountability model. Because the introduction of an AA-MAS can have an impact on the PI for all students and especially the SWD subgroup, the NYSED may find it beneficial to evaluate the effective AMOs. One approach may be to model

results given different same size ranges. Currently, the ranges vary by 5 until a group size of 50 and then increase by units of 10 getting progressively larger. Because the confidence interval stabilizes with large n sizes, it is unlikely that the upper range will be impacted. However, for smaller n sizes, it may be useful to adjust the ranges (perhaps constricting them) and note differences in classification outcomes for schools and subgroups.

The collection of multiple sources of qualitative and quantitative information will strengthen overall findings. For example, if data exist related to outstanding professional development or instructional programs, how do the schools and/or groups recognized for such programs perform on the AA-MAS in particular and the accountability system in general?

Additionally, the NYSED may wish to be intentional about collecting data regarding the opportunity to learn and student characteristics for the population taking the AA-MAS. This may be accomplished through initiatives such as surveying teachers and school leaders on the quality and consistency of instructional opportunities, student engagement, and other indicators (e.g. class work) of student success. Some of these methods are discussed in further detail in Chapter 3 (Karvonen, this volume). By comparing this information with AA-MAS results and accountability determinations, additional evidence about the efficacy of the system may be produced.

Finally, in analyzing findings, it is important to consider both Type I and Type II errors. A Type I error may be said to occur when a school with strong, effective programs does not make AYP and is determined to be in an improvement status. A Type II error describes the situation where a school in need of improvement is erroneously classified as meeting standards. In practice, an increase in Type II error may be the larger threat with the introduction of an AA-MAS. Ideally, if fewer schools are classified as needing improvement, it will be due to more appropriate assessments that accurately reflect a higher level of student achievement previously masked by barriers on the general assessment. However, to the extent that the AA-

MAS is used to lower expectations, Type II error will be elevated and students in need of support services may not be identified.

Reporting

Another important element of an assessment and accountability system is public reporting. Decisions about the design and distribution of performance reports directly impact the theory of action that can promote student and school improvement. Therefore, a plan for effective assessment accountability reporting practices related to the AA-MAS is essential. In general there are three major considerations with respect to reporting: 1) identify the information that should be reported; 2) determine how the information should be presented; and 3) decide how the information will be disseminated.

The United States Department of Education has explicitly defined the information that must be reported in NCLB compliant accountability systems, which is currently incorporated in NYSED's reporting system. Additional requirements from the 2007 regulations stipulate that accountability determinations should include: 1) the number of students with disabilities participating in the general assessments and the number provided accommodations; 2) the number participating in the AA-AAS and the AA-MAS; and 3) performance results for students taking each assessment.

The guiding principle for designing reports is to make the information accessible to stakeholders such that it is actionable. In her 2002 CCSSO publication addressing accountability reporting, Ellen Forte proposes the following criteria for effective reports:

- Accessible to the target audiences, both physically and linguistically;
- Accompanied by adequate interpretive information;
- Supported by evidence that the indicators, other information, and suggested interpretations are valid;
- Coordinated with other reports within the reporting system:

- Across paper and electronic versions of report cards, and
- Across reports cards and assessment reports.

These criteria suggest that the reports should be designed such that they are technically comprehensive, but simple to read and understand by all stakeholders—a nontrivial task.

However, there are a few approaches that may help accomplish this. For example, the NYSED may consider including reader-friendly narratives that describe the knowledge and skills in each performance level on student level reports and/or supporting documents. Moreover, presenting key information in graphical format on both student and summary reports often improves the readability and usefulness of reports. To the extent that it is practicable, reports should follow a standard format across programs, which may reduce confusion for consumers of multiple reports. Finally, reports and supporting documents are often reviewed by broad-based committees to promote the likelihood that the information is presented appropriately.

Moreover, it will be important to support appropriate interpretation and use of the results of the AA-MAS. The *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) address this principle explaining, “interpretations should describe in simple language what the test covers, what scores mean, the precision of the scores, common misinterpretations of test scores, and how scores will be used” (p. 65). This is a vital component of any testing program but is particularly important given the distinctive nature of an alternate assessment and the students assessed. Examples of support initiatives might include distributing an interpretative guide, developing online resources, and/or conducting training workshops with educators.

Diploma Eligibility

One important policy issue for the NYSED is the impact of an AA-MAS on diploma eligibility. Currently, students may exit high school with a local diploma, a Regents diploma, an Advanced Regents diploma, or an IEP certificate. If the AA-MAS is developed for high school

content, a decision regarding whether or not students will be eligible for a Regents or local diploma and, if so, what level performance is required, must be resolved.

USED regulations require states to ensure that students who take an AA-MAS are not precluded from attempting to complete the requirements for a regular high school diploma. This requirement does not compel the state of New York to treat AA-MAS scores as comparable to those from general assessments with respect to diploma eligibility criteria. The regulation is intended to prohibit tracking that might prevent a student from taking a path that leads to a regular diploma. Stated another way, students cannot be denied the option to qualify for a regular diploma (whatever those qualifications are) if they take an AA-MAS at any point.

Therefore, a number of possibilities can be considered to operationalize an AA-MAS in a way that is consistent with federal requirements. One approach would be to establish a level of performance on the AA-MAS that is regarded as an acceptable qualification for a Regents diploma. The methods discussed in Chapters 7 and 8 (Perie and Abedi, this volume) could inform the selection of a cut score that serves this purpose — as might be produced in a linking study. The topic of determining performance standards and cut scores is further addressed in Chapter 6 (Welch and Dunbar, this volume) and Chapter 7 (Perie, this volume).

Another approach would be to continue the policy of using performance on the Regents Examination as the acceptable qualification for a Regents or local diploma. If this approach is selected, the importance of developing clear guidelines and procedures for how students can move from an AA-MAS to the general assessment is elevated. That is, students will need to be clearly informed about the path and requirements necessary to qualify to take a Regents Examination and all students should have an opportunity to pursue that path.

A third option might involve establishing multiple criteria for diploma eligibility for students that take an AA-MAS. This rationale behind this option is that the AA-MAS alone may not provide sufficient evidence that a student has achieved graduation requirements. However, coupled with additional indicators, such a decision can be supported. Examples of indicators

that may provide such evidence might include: recommendation from the students IEP committee, meeting identified course-taking or performance standards, achieving a requisite score on another assessment (e.g. SAT or ACT), or meeting selected vocational or industry certification credentials.

The decision regarding the role of the AA-MAS in diploma eligibility should be based on the values and priorities of the NYSED and the characteristics of the AA-MAS. That is, as a matter of policy the department determines the knowledge, skills, competencies etc. that are required for each diploma type. Then, the extent to which the AA-MAS produces a measure that satisfies these criteria will largely define how it will function with respect to diploma eligibility.

Finally, there are important legal considerations to attend to if a state changes diploma eligibility requirements. As established in the landmark *Debra P. v. Turlington* (1981) the state must provide adequate notice of any changes to assessment requirements related to diploma eligibility and ensure there is a high degree of content validity. Moreover, it is advisable to conduct research (e.g. broad distribution of a survey) to gauge the extent to which students have an opportunity to learn the knowledge and skills covered on the assessment. Such research might include a review of IEPs to ensure learning goals and supports are in line with expectations of the AA-MAS.

Conclusion and Recommendations

The overarching theme of this chapter is that developing and implementing an AA-MAS should not be regarded as an isolated enterprise. A full consideration of the issues and options, should involve a review of many practical and policy issues related to the entire assessment and accountability system.

This process begins with an examination of whether to implement an AA-MAS and, if so, to what extent? As discussed, this question is largely informed by carefully studying the extent to which the current assessment system is appropriate for students with disabilities. The 'stakes'

of the assessment should also be taken into consideration when considering the scale and/or priorities for implementation. Finally, it is unavoidable that availability of resources will influence the capacity to move forward.

Determining eligibility criteria is another key decision. Data sources and approaches to inform this decision were explored in this chapter such as using assessment data to evaluate the likelihood of reaching target performance on future administrations and analyzing the characteristics of persistently low performers. Finally, setting and evaluating participation criteria is bolstered when multiple, corroborating data sources are used.

This task is complicated by the need to disentangle low performance due to disability from that which is due to lack of opportunity to learn. It remains critically important for states to investigate strategies to support all learners by evaluating educational services. Moreover, it is advisable to review the development process and policies related to general assessments to maximize the likelihood that *all* students are afforded the opportunity to demonstrate what they know and can do. This may include such practices as attention to universal design or a review of accommodations options to ensure they are effective and appropriate.

It is also important to explore the impact of the AA-MAS on the state accountability system. There are methods available to evaluate decision consistency, which is impacted by two main sources: measurement error and sample error; the latter of these accounts for most of the variability in accountability determinations. Accordingly, some approaches suggested by Hill and DePascale (2002) and Arce-Ferrer, Frisbie, and Kolen (2002) were presented to evaluate the impact of sample error.

The discussion of impact to accountability systems also included a review of operational considerations. In this section, some features specific to the state of New York (e.g. effective AMOs and the Performance Index) were discussed. Additionally, some approaches suggested by Martinez and Olson (2004) to manage the redistribution of non-proficient scores were presented. The author concludes that a method based on pre-determining thresholds for district

participation rates may be most promising, provided the state applies additional scrutiny to explore and possibly adjust for defensible deviations from these values.

In the following section, a number of specific analyses to evaluate the impact to accountability systems were suggested. Many of these approaches can be conducted annually for ongoing system monitoring, which is certainly advisable. In the discussion, a method to provide advance information about the impact of implementing an AA-MAS was proposed. Because it is likely that fluctuations will be non-uniform, the primary benefit of this approach will be to identify the areas that are likely to have the most substantial impact, which can help the state prepare for implementation.

Certainly, the utility of assessment information is strongly tied to the quality of external reports. For this reason, some succinct recommendations were presented to produce accessible information on student and summary reports and produce well-designed support materials. This may be best accomplished by having broad based groups assist with design or review of materials. Moreover, maintaining some consistency of presentation on the reports will increase the likelihood that the information provided on the reports will be meaningful to stakeholders.

Finally, some considerations related to diploma eligibility policy were presented. The key point is that policies should be established that provide a path for students who take an AA-MAS to be eligible for a regular diploma. Such a policy may identify a specific performance level on the AA-MAS or may involve alternate and/or multiple criteria to meet this standard. In any case, the policy should be clearly articulated and in line with the state's values and priorities for high school graduates.

Ultimately, the NYSED's objective is to ensure the continuance of a coherent, effective assessment and accountability system. This is accomplished by careful planning and systematic evaluation. By so doing, the state is able to design and operationalize a more suitable

assessment and accountability system, which best positions the state of New York, or any other state, to promote student achievement.

References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Arce-Ferrer, A., Frisbie, D.A., & Kolen, M.J. (2002). *Standard errors of proportions used in reporting changes in school performance with achievement levels*. Educational Assessment, 8(1), 59-75.
- Chapman v. California Department of Education, 229 F. Supp. 981 (N.D. Calif., 2002)
- Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Belmont, CA: Wadsworth.
- Cronbach, L.J. (1951). *Coefficient alpha and the internal structure of tests*. Psychometrika, 16, 297-334.
- Darling-Hammond, L. (2006). Standards, Assessments, and Educational Policy: In Pursuit of Genuine Accountability. Eighth Annual William H. Angoff Memorial Lecture. Princeton, NJ: Educational Testing Service.
- Debra P. v. Turlington, 644 F2d 397 (5th Cir. 1981).
- Fincher, M. 2007. "Investigating the Academic Achievement of Persistently Low Performing Students" in the session on *Assessing (and Teaching) Students at Risk for Failure: A Partnership for Success* at the Council of Chief State School Officers Large Scale Assessment Conference, Nashville TN, June 17-20, 2007. Available at: <http://www.ccsso.org/content/PDFs/12%2DMelissa%20Fincher%20Paul%20Ban%20Pam%20Rogers%20Rachel%20Quenemoen.pdf>.
- Forte Fast, E. (2002). *A guide to effective accountability reporting*. Council of Chief State School Officers State Collaborative on Assessment and Student Standards Accountability Systems and Reporting Consortium.
- Fuhrman, S. & Elmore, R. (Eds.). (2004). *Redesigning accountability systems for education*. New York: Teachers College Press.
- Gong, B., & Marion, S. (2006). *Dealing with flexibility in assessments for students with significant cognitive disabilities*. Dover, NH: National Center for the Improvement of Educational Assessment.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65-110). Westport, CT: American Council on Education/Praeger.
- Hill, R.K., & DePascale, C.A. (2002). Determining the reliability of school scores. Portsmouth, NH: The National Center for the Improvement of Educational Assessment Inc.
- Lazarus, S. S., Thurlow, M. L., Christensen, L. L., & Cormier, D. (2007). *States' alternate assessments based on modified achievement standards (AA-MAS) in 2007* (Synthesis

Report 67). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Marion, S., White, C., Carlson, D., Erpenbach, W, Rabinowitz, S., & Sheinker, J. (2002). Making valid and reliable decisions in determining adequate yearly progress. Washington, DC: Council of Chief State School Officers.

Martinez, T., & Olsen, K. (2004). Distribution of proficient scores that exceed the 1% cap: Four possible approaches. Mid-South Regional Resource Center. Available at: http://eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/1b/a3/1f.pdf

No Child Left Behind Act of 2001, 20 U.S.C. § 6301

O'Day, J., & Smith, M. (1993). Systemic school reform and educational opportunity. In S. H. Fuhrman (Ed.), *Designing coherent education policy: Improving the system*. San Francisco: Jossey-Bass.

Perie, M. (2007). *Key elements for educational accountability models*. Washington, DC: Council of Chief State School Officers.

United States Department of Education. *Modified-Academic Achievement Standards: Non-Regulatory Guidance*. April, 2007.

United States. Department of Education. National Commission on Excellence in Education. *A Nation at Risk: The Imperative for Educational Reform*. Washington: GPO, 1983.