Assessment and Accountability Comprehensive Center
AACC • A WestEd and CRESST partnership

Guidelines for Ensuring the Technical Quality of Assessments Affecting English Language Learners and Students with Disabilities: Development and Implementation of Regulations

Edynn Sato, Peter Worth, Carole Gallagher, Rachel Lagunoff, and Holly McKeag
Assessment and Accountability Comprehensive Center
September 2007

Table of Contents

Overview and Background

These guidelines, prepared by the Special Populations Strand of the Assessment and Accountability Comprehensive Center (AACC), focus on the technical quality of assessments for English language learners (ELLs) and students with disabilities (SWDs). This document is an evolving document that will periodically be updated to incorporate new information. This document is intended to provide information to Regional Comprehensive Centers (RCCs) and states as they work to comply with the regulations of No Child Left Behind (NCLB) affecting their special student populations (i.e., SWDs, ELLs). These guidelines also are intended to help RCCs and states:

- gauge where a state is with regard to meeting federal requirements relevant to the assessment and accountability of special student populations;
- focus attention on priority issues related to implementing practices and systems that are in compliance with federal regulations; and
- select implementation strategies that have evidence of effectiveness, given the particular needs and conditions of the state.

As mentioned previously, information presented in these guidelines will be updated as new and relevant research, guidance, and strategies become available for consideration and evaluation by the AACC. Additionally, these guidelines will be updated to meet the evolving needs of RCCs and states.

States are at varying stages of implementing federal regulations (NCLB Title I, Title III) affecting the assessment and accountability of their special student populations. According to our initial analysis, which involved an examination of the needs and priorities related to assessment and accountability that were identified in the U.S. Department of Education–organized Regional Advisory Committee (RAC)

> Regions reported that "helping raise the achievement of at-risk, special needs, and ELL students" is a key priority with significant implications for the development of accountability and assessment systems.

WestEd
excellence in research, development, & service

Reports, nine of the ten regions reported that "helping raise the achievement of at-risk, special needs, and ELL students" is a key priority with significant implications for the development of accountability and assessment systems (see Table 1 for an overview of needs across regions). Confirmatory evidence of the most pressing assessment and accountability needs identified by the RAC reports was obtained from a review and analysis of NCLB reports, evaluations, and critiques across the research and political spectrum that ranged from the highly technical (Linn, Baker, & Betebenner, 2002; Gong, 2005; Fast & Hebbler, 2004; NCES, 2003, 2004; Rabinowitz & Ananda, 2002; Rabinowitz, 2004) to the more general (Center on Education Policy, 2005; Uzzell, 2005; *Education*

Table 1. Needs and Priorities Related to Assessment and Accountability Identified in the Regional Advisory Committee Reports[1]

Priorities	Appalachia	Mid-Atlantic	Mid-Continent	North Central	Northeast	Northwest	Pacific	Southeast	Southwest	West
Help raising achievement of at-risk, special needs, and ELL students	X	X		X	X	X	X	X	X	X
Appropriate assessments that are valid and reliable for special and diverse populations (e.g., ELLs, special education, low SES, ethnic minority)	X	X				X	X	X	X	X
Training for teachers in use of assessment data	X			X	X			X		X
Formative and summative assessments	X	X			X		X	X		X
Resources to address needs identified by assessment data	X			X	X		X			
Alignment of standards, instruction, and assessment	X			X	X			X		
Training for administrators in use of assessment data	X				X		X			X
User-friendly and timely dissemination of assessment data	X			X				X		
Locally developed assessments (linguistically and culturally appropriate)			X			X	X			
Dissemination of best practices		X				X				
Assessment-related technology training for teachers								X		
Consistency in benchmarking assessments from LEA to LEA										X
Assessment development training								X		

[1] Based on available reports as of July 2005. Recent conversations with Regional Comprehensive Centers (RCCs) and states have confirmed that the assessment and accountability of English language learners (ELLs) and students with disabilities (SWDs) continue to be areas of need.

Week, 2004, 2005). Therefore, a strand of the AACC's work is dedicated to the assessment and accountability of ELLs and SWDs.

Federal Peer Review comments to states have identified areas where states need assistance vis-à-vis key review criteria and with regard to their special student populations. Such results, along with recent publications and surveys, indicate that states and districts need help with the development and implementation of technically adequate assessment systems for special student populations (Abedi, 2004; Herman & Dietel, 2005; American Diploma Project, 2004), and special attention to the *technical quality* of these assessments is needed to ensure they are *valid* and *accessible* for these students. Various strategies and systems for assessment and accountability of ELLs and SWDs exist; however, they are not aggregated in any methodical fashion so that there is no complete understanding of the (a) quality[2] of these strategies/systems and the (b) context or conditions[3] under which these strategies/systems are being implemented. Additionally, no framework exists to meaningfully organize the information that is available to RCCs and states.

States and districts need help with the development and implementation of technically adequate assessment systems for special student populations, and special attention is needed to ensure these systems are *valid* and *accessible* for these students.

Therefore, this document focuses on the issue of technical quality and presents:

- critical elements from the Federal Peer Review technical quality criteria (Title I) and Title III Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students (OELA) Monitoring Reports, with available examples of acceptable and incomplete evidence;
- a comparison of Federal Peer Review (Title I) critical elements with validated criteria for ensuring the technical adequacy of assessments for special student populations;
- a summary of research and resources relevant to key issues, including: accommodations, standard setting, and Annual Measurable Achievement Objectives (AMAOs) for ELLs; and
- a comparison of Title I and Title III requirements for assessing ELLs that are applicable to the technical quality of those assessments.

Note: At the request of the U.S. Department of Education, the AACC is leading the development of an initial draft framework for English language proficiency standards and assessments. The AACC will provide updates on this framework as they become available.

[2] *Quality* refers to the degree to which the strategies/systems comply with NCLB regulations (Title I, Title III).

[3] *Context and conditions* include: financial, political, historical, and demographic.

Accountability of Special Student Populations: English Language Learners and Students with Disabilities

English language learners (ELLs) are held accountable in two ways under NCLB: as a subgroup, they must meet Adequate Yearly Progress (AYP) under Title I for reading, math, and science; and they must meet Annual Measurable Achievement Objectives (AMAOs) under Title III. Meeting the Title I AYP requirement helps states relate ELL gains in English learning and proficiency to the preparation of this subgroup of students to meet challenging state academic achievement standards. Meeting Title III AMAOs means that states must define annual measurable achievement objectives for the ELLs they serve such that states can show increases in the number and percent of students (a) making progress in learning English and (b) attaining English proficiency.

Students with disabilities (SWDs) also are measured annually per NCLB Title I vis-à-vis challenging academic content standards and academic achievement standards. SWDs are held accountable as a subgroup for meeting or exceeding (or, in some cases, for demonstrating continuous and substantial progress toward) state-specific proficiency targets (AYP) in reading, mathematics, and science. State plans for assisting SWDs in reaching performance goals, including decision-making about supplemental educational services, are developed in coordination with requirements of the Individuals with Disabilities Education Act (IDEA). For students with the most significant cognitive disabilities, states have been granted flexibility in assessing the academic progress of these students, provided that 1) these students continue to be held to appropriate academic content and achievement standards; 2) each student's Individualized Education Program (IEP) team determines the level of participation in state assessments; and 3) these students are not excluded from the state accountability system. All assessments developed by states must allow for reasonable accommodation of SWDs (per Sec. 602[3] of IDEA, 2004) during testing, provide coherent information about student attainment of standards, and be consistent with nationally recognized standards for technical quality. In addition to reporting performance on the state assessment, states must also report the level (percent) of SWD's participation and their performance on a secondary academic indicator (e.g., attendance or graduation rates).

Public Law Title I and Title III

In order to effectively meet requirements for both Title I and Title III for ELLs, states must understand the requirements for valid assessments that are appropriate for both the needs of the states as well as the needs of their special student populations. To assist in this regard, Table 2 presents selected information from the Title I Public Law and Title III Public Law as they relate to the assessment of ELLs.

In order to effectively meet requirements for both Title I and Title III, states must understand the requirements for valid assessments that are appropriate for both their needs as well as the needs of their special student populations.

Table 2. Title I and Title III Requirements for Assessing ELLs

Notes: States can use the same assessment for testing English language proficiency under Title I and Title III. Both Title I and Title III require states to provide reasonable accommodations on state academic content assessments for LEP[4] students (e.g., native language assessments, extra time, linguistic simplifications, etc.).

	Title I	Title III
Who	• Title I mandates the inclusion of LEP subgroup in AYP calculations (school and district). • LEP students who have been in the U.S. for three consecutive years are assessed in reading/language arts in English (except for those residing in Puerto Rico). • For first three years, ELLs may take assessment in student's native language, but the assessment must be aligned with the state content and achievement standards. On a case-by-case basis, districts may continue to administer the assessment in the student's native language for an additional two years.	• Students who receive Title III services must take assessment of English language proficiency; usually the local educational agency (LEA) decides, but the state may have policies that establish parameters for LEA decisions. Therefore, who is tested under Title III could vary by LEA and state.
Excludes	• Newly arrived LEP students are not counted in accountability for either reading/language arts or mathematics for one year, even if they meet the state's definition of full academic year.	• ELLs not receiving Title III services. *Note: Some policies require all ELLs to be tested, but who is counted for Title III accountability is dependent on who receives Title III services.*
What	• Assessment of English language proficiency in four domain areas: reading, writing, speaking, and listening. • All LEP students must take the mathematics assessment with appropriate accommodations. • Starting in school year 2007–2008, LEP students will be required to take state science assessment.	• Assessment of English language proficiency in four domain areas: reading, writing, speaking, and listening. • Must report a separate score for the domain of *comprehension* (can be demonstrated through reading and listening).
When	• Each LEA is required to evaluate their program on an *annual* basis.	• Each LEA is required to evaluate their program on an *annual* basis. In addition, Title III requires LEAs to report on the progress made by LEP students in meeting state academic content and achievement standards for each of the two years after they no longer receive Title III services.
How	• To the extent practicable, assessments written in the native languages should be provided to LEP students until students have achieved English language proficiency.	Title III requires states to: • Conduct an annual, standards-based assessment of English language proficiency • Define annual measurable achievement objectives (AMAOs) for increasing percentage of ELLs progressing toward and attaining English proficiency, and for meeting academic achievement standards: 1. AMAO 1 – annual increases in the number or percentage of children *making progress* in learning English 2. AMAO 2 – annual increases in the number or percentage of children *attaining English language proficiency* by the end of each school year 3. AMAO 3 – *adequate yearly progress* for the ELL subgroup in meeting grade-level academic achievement standards in English language arts and mathematics • Hold LEAs accountable for meeting the AMAOs

[4] The language of NCLB refers to the targeted student population as "limited English proficient." Limited English proficient (LEP) students are a) 3 to 21 years of age, b) enrolled or preparing to enroll in elementary or secondary school, c) either not born in the United States or have a native language other than English, and d) owing to difficulty in speaking, reading, writing, or understanding English, not able to meet the State's proficient level of achievement to successfully achieve in English-only classrooms or not able to participate fully in society (Title IX, Section 9101). We recognize that many researchers and practitioners prefer the term English language learner (ELL). Consistent with this more general, common usage, the remainder of this document will use the term English language learner.

Assessments for Special Student Populations: Technical Quality

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) there are multiple elements that contribute to the technical quality of high-quality assessments. Key elements contributing to technical quality include validity, reliability, and freedom from bias. Each of these key elements is discussed below.

Validity

According to the *Standards*, a primary consideration in determining validity is whether the state has evidence that the assessment results can be interpreted in a manner consistent with the assessment's intended purpose(s). *Construct validity* is the extent to which an assessment measures what it is intended to measure as well as the extent to which inferences and actions made on the basis of test scores are appropriate and accurate.

There are four broad categories of evidence that can be used to support validity (AERA, APA, & NCME, 1999; Kane, 2002; Messick, 1989):

1. **Test content:** the degree to which the standards and the assessment (items and forms) align.
2. **The assessment's relation to other variables:** the relationship between the assessment and other measures known to be accurate indicators of student knowledge/ability.
3. **Student response processes:** the degree to which factors that contribute to assessment ambiguity and inaccuracy have been eliminated or minimized such that assessment results accurately reflect student knowledge/ability vis-à-vis the tested content.
4. **Internal structure:** the degree to which a variety of statistical techniques have been applied to the test to determine its validity and reliability and to ensure a balanced assessment in terms of breadth and depth of knowledge, skills, and content assessed.

Tables 3–8 (pp. 11–16) present examples of evidence that state officials can consider when documenting the validity of their assessments.

Additionally, according to Messick (1989), consideration also must be given to the consequences of the test's interpretations and uses. The validity and accuracy of test interpretation and use are critical because misinterpretation and misuse could result in unintended and negative consequences.

The validity and accuracy of test interpretation and use are critical because misinterpretation and misuse could result in unintended and negative consequences.

State officials should address and document the validity of each of the state's assessments, including alternate assessments, in *all* of the following key areas (based on USED, 2004, Critical Element 4.1):

a. Specify the purposes of the assessments, delineating the types of uses and decisions most appropriate to each.

b. Ascertain that the assessments, including alternate assessments, are measuring the knowledge and skills described in the state's academic content standards and not knowledge, skills, or other characteristics that are not specified in the academic content standards or grade level expectations.

c. Ascertain that the state's assessment items are tapping the intended cognitive processes and that the items and tasks are at the appropriate grade level.

d. Ascertain that the scoring and reporting structures are consistent with the sub-domain structures of its academic content standards (i.e., item interrelationships are consistent with the framework from which the test arises).

e. Ascertain that test and item scores are related to outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and are weakly correlated, if at all, with irrelevant characteristics, such as demographics).

f. Ascertain that decisions that are based on the results of the state's assessments are consistent with the purposes for which the assessments were designed.

g. Determine what are the intended and unintended consequences that result from the state's assessments.

Reliability

Reliability refers to the consistency with which an assessment yields results that are dependable and consistent indicators of particular student knowledge/skills. Such consistency can exist over time, across raters, or across different items/tasks intended to measure the same content. Test reliability has implications for test validity because sources of error that lead to unwanted variation in assessment results may distort the interpretation and use of the results (AERA, APA, & NCME, 1999; Anastasi, 1988; Berkowitz, Wolkowitz, Fitch, & Kopriva, 2000).

There are three major sources of error:

- Factors in the test itself;
- Factors in the students taking the test; and
- Scoring factors.

State officials should address and document the reliability of each of the state's assessments, including alternate assessments, in *all* of the following ways (based on USED, 2004, Critical Element 4.2):

a. Based on data for the state's own student population and each reported subpopulation, determine the reliability of the scores that the state reports.

b. Quantify and report within the technical documentation for the state's assessments the conditional standard errors of measurement and student classification that are consistent at each cut score specified in the state's academic achievement standards.

c. Report evidence of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring.

Tables 3–8 (pp. 11–16) present examples of evidence that state officials can consider when documenting the reliability of their assessments.

Bias

Bias is the presence of information in a test or a condition of the test that unfairly advantages or disadvantages a student (or group of students) such that the student is unable to accurately demonstrate what he or she knows and can do vis-à-vis the tested content. Consequently, test results might underestimate the student's achievement or reflect abilities that are not related to the intended test content (Abedi & Lord, 2001; AERA, APA, & NCME, 1999; Kopriva, 2000).

Sources of bias include:

- Gender;
- Racial/ethnic;
- Cultural;
- Geographic;
- Disability; and
- Linguistic.

Bias can be introduced during various phases of a test's development and use (AERA, APA, & NCME, 1999):

- **Design/development:** The items or tasks do not provide an equal opportunity for all students to fully demonstrate their knowledge and skills.
- **Administration:** The assessments are not administered in ways that ensure fairness.
- **Reporting:** The results are not reported in ways that ensure fairness.
- **Interpretation:** The results are not interpreted or used in ways that lead to equal treatment.

Additionally, bias could be attributed to the insufficient opportunity of students to access and learn the standards.

Therefore, states must ensure that during each stage of their assessments' development and use, potential sources of bias are identified and efforts are made to reduce or eliminate the effects of bias on student performance. For all assessments in the state's assessment system, state officials should ensure that the assessments are fair and accessible to all students, including SWDs and ELLs, in the following manner (based on USED, 2004, Critical Element 4.3):

States must ensure that during each stage of their assessments' development and use, potential sources of bias are identified and efforts are made to reduce or eliminate the effects of bias on student performance.

a. Ensure that the assessments provide an appropriate variety of accommodations for students with disabilities.

b. Ensure that the assessments provide an appropriate variety of linguistic accommodations for students with limited English proficiency.

c. Take steps to ensure fairness in the development of the assessments.

d. Ensure that the use of accommodations and/or alternate assessments yields meaningful scores.

Tables 3–8 (pp.11–16) present examples of evidence that state officials can consider when documenting the manner in which they have controlled for bias in the state's assessments.

Additional factors impacting assessment validity, reliability, and freedom from bias

Aspects of validity, reliability, and bias often are interrelated, and each element is affected by a number of factors. In addition to the factors described above, state officials ought to consider the following (based on USED, 2004, Critical Elements 4.4, 4.5, and 4.6):

Aspects of validity, reliability, and bias often are interrelated, and each is affected by a number of factors.

1. When different test forms or formats are used, state officials must ensure that the meaning and interpretation of results are consistent.
   - *a.* Ensure consistency of test forms over time.
   - *b.* If the state administers both an online and paper-and-pencil test, document the comparability of these two forms of the test.
2. Establish clear criteria for the administration, scoring, analysis, and reporting components of the state's assessment system, including alternate assessment(s), and maintain a system for monitoring and improving the ongoing quality of the state's assessment system.
3. Evaluate the state's use of accommodations.
   - *a.* Ensure that appropriate accommodations are available to students with disabilities and that these accommodations are used in a manner that is consistent with instructional approaches for each student, as determined by the student's IEP or 504 plan.
   - *b.* Determine that scores for students with disabilities that are based on accommodated administration conditions will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration conditions.
   - *c.* Ensure that appropriate accommodations are available to limited English proficient students and that these accommodations are used as necessary to yield accurate and reliable information about what limited English proficient students know and can do.
   - *d.* Determine that scores for limited English proficient students that are based on accommodated administration circumstances will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration circumstances.

Validation efforts should occur during *each phase* of an assessment's development and use, and state officials should carefully gather and document evidence of their assessments' validity, reliability, and freedom from bias.

Tables 3–9 (pp. 11–19) provide relevant information from three key resources in order to assist state officials in their consideration of the evidence that they need to establish the technical quality

Validation efforts should occur during *each phase* of an assessment's development and use, and state officials should carefully gather and document evidence of their assessments' validity, reliability, and freedom from bias.

of their assessments. The three main sources for these tables are:

- *Standards and Assessments Peer Review Guidance* (USED, 2004)
  In response to NCLB legislation (Sec. 111[b][3]) and regulations (Sec. 200.2), the U.S. Department of Education (USED) has provided states with guidance regarding the evidence that can be used to demonstrate state compliance with NCLB requirements. See Tables 3a–8a for examples of acceptable and incomplete evidence of technical quality.
- *Title III OELA Monitoring Reports* (OELA, 2006)
  The Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students (OELA) has issued guidance for its grantees to use in preparing annual reports. This guidance includes descriptions of critical elements for English Language Proficiency standards and assessments as well as acceptable evidence for these elements. Many of the elements and evidence presented in this OELA document are similar to those in the USED's *Standards and Assessment Peer Review Guidance*. Therefore, Tables 3b–8b also present information from the OELA document that is related to the critical elements identified by the Federal Peer Review.
- *Evaluation of the Technical Evidence of Assessments for Special Student Populations* (AACC, 2007)
  The AACC offers a comprehensive set of criteria validated by a team with expertise in assessment, linguistics, and English language development, based on those developed by Rabinowitz and Sato (2005, 2006) to evaluate the technical evidence associated with assessments for ELLs in particular and special student populations in general. These technical criteria are sensitive to the unique characteristics of the student population, the particular purposes of the assessments, and the stage of development and maturity of the assessments. Technical criteria can be found in the document titled *Evaluation of the Technical Evidence of Assessments for Special Student Populations* at www.aacompcenter.org (see Special Populations page).

  See Table 9 for a crosswalk between these technical criteria and the critical elements for technical quality identified in the USED's *Standards and Assessment Peer Review Guidance*.

Table 3a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.1 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.1 For each assessment, including alternate assessment(s), has the State documented the issue of **validity** (in addition to the alignment of the assessment with the content standards), as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories: **(a)** Has the State specified the purposes of the assessments, delineating the types of uses and decisions most appropriate to each? *and* **(b)** Has the State ascertained that the assessments, including alternate assessments, are measuring the knowledge and skills described in its academic content standards and not knowledge, skills, or other characteristics that are not specified in the academic content standards or grade level expectations? *and* **(c)** Has the State ascertained that its assessment items are tapping the intended cognitive processes and that the items and tasks are at the appropriate grade level? *and* **(d)** Has the State ascertained that the scoring and reporting structures are consistent with the sub-domain structures of its academic content standards (i.e., are item interrelationships consistent with the framework from which the test arises)? *and* **(e)** Has the State ascertained that test and item scores are related to outside variables as intended (e.g., scores are correlated strongly with relevant measures of academic achievement and are weakly correlated, if at all, with irrelevant characteristics, such as demographics)? *and* **(f)** Has the State ascertained that the decisions based on the results of its assessments are consistent with the purposes for which the assessments were designed? *and* **(g)** Has the State ascertained whether the assessment produces intended and unintended consequences?	For each assessment, including alternate assessment(s), the State has documented the existing validity evidence in each of the categories and has taken steps to address any deficiencies either in validity or in its approach to establishing and documenting validity evidence. *Possible Evidence* • For category **(a)**, existing written documentation, such as minutes or policies of the State Board of Education or state legislative code, that defines the purpose(s) of the State's assessment system. • For each of the categories **(b) – (g)**, documentation of the studies that provide evidence in support of the validity of using results from State's assessment system for their stated purpose(s).	The State has not provided evidence in all categories **(a) – (g)** or has not taken steps to address any deficiencies either in validity or in its approach to establishing and documenting validity evidence.

Table 3b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Validity

Critical Element	Examples of Acceptable Evidence
3.1 (c) ELP standards are linked to State content and achievement standards in reading/language arts, math, and science (science in 2005–2006)	Acceptable evidence includes a process and documentation for linkage and alignment, findings from linkage and alignment studies, and state responses to findings.
3.2 (c) ELP assessments are aligned to ELP standards	
3.2 (d) ELP assessments are of high technical quality, including being valid, reliable, and fair	Acceptable evidence includes technical manuals for ELP assessment(s), including scoring guides, and other documents that describe the ELP assessment(s).

Table 4a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.2 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.2 For each assessment, including alternate assessment(s), has the State considered the issue of **reliability,** as described in the *Standards for Educational and Psychological Testing* (AERA/APA/NCME, 1999), with respect to *all* of the following categories: **(a)** Has the State determined the reliability of the scores it reports, based on data for its own student population and each reported subpopulation? *and* **(b)** Has the State quantified and reported within the technical documentation for its assessments the conditional standard error of measurement and student classification that are consistent at each cut score specified in its academic achievement standards? *and* **(c)** Has the State reported evidence of generalizability for all relevant sources, such as variability of groups, internal consistency of item responses, variability among schools, consistency from form to form of the test, and inter-rater consistency in scoring?	For each assessment, including alternate assessment(s), the State has documented reliability evidence in each of the categories and has taken steps to address any deficiencies either in reliability or in the State's approach to establishing and documenting reliability evidence. *Possible Evidence* • For each of the categories **(a) – (c)**, documentation of the studies that support the reliability of each of the State's assessments with the State's own student population. • Documentation of the precision of the assessments at cut scores and evidence of a systematic process for addressing any deficiencies identified in these studies. • Documentation of consistency of student level classification and evidence of a systematic process for addressing any deficiencies identified in these studies.	The State has not provided evidence in all categories **(a) – (c)** or has not taken steps to address any deficiencies either in reliability or in the State's approach to establishing and documenting reliability evidence.

Table 4b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Reliability

Critical Element	Examples of Acceptable Evidence
3.2 (d) ELP assessments are of high technical quality, including being valid, reliable, and fair	Acceptable evidence includes technical manuals for ELP assessment(s), including scoring guides, and other documents that describe the ELP assessment(s).

Table 5a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.3 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.3 Has the State ensured that its assessment system is fair and accessible to all students, including students with disabilities and students with limited English proficiency, with respect to each of the following issues: **(a)** Has the State ensured that the assessments provide an appropriate variety of accommodations for students with disabilities? *and* **(b)** Has the State ensured that the assessments provide an appropriate variety of linguistic accommodations for students with limited English proficiency? *and* **(c)** Has the State taken steps to ensure fairness in the development of the assessments? *and* **(d)** Does the use of accommodations and/or alternate assessments yield meaningful scores?	The State has taken appropriate judgmental (e.g., committee review) and data-based (e.g., bias studies) steps to ensure that its assessment system is fair and accessible to all students. Review committees have included representation of identified subgroups. The State assessment system must be designed to be valid and accessible for use by the widest possible range of students. The State is conducting studies to determine the appropriateness of accommodations and the impact on test scores. *Possible Evidence* • Existing written documents describe how the principles of universal design and/or appropriate language simplification were incorporated into each of the State's assessments. • Evidence that students with disabilities were included in the test development process. • Existing written documentation of the State's policies and procedures for the selection and use of accommodations and alternate assessments, including evidence of training for educators who administer these assessments.	The State has conducted data-based bias studies but has not convened committees of stakeholders to review its assessment items. The State has convened committees of stakeholders to review its assessment items but these committees have not included representation of identified subgroups. The State assessment system is not designed to be valid and accessible for use by the widest possible range of students. The State does not have a policy on the appropriate selection and use of accommodations and alternate assessments. The State does not train or monitor personnel at the school, LEA, and State levels with regard to the appropriate selection and use of accommodations and alternate assessments. There are no appropriate accommodations for students with particular disabilities (e.g., no allowable accommodations on the regular assessment or alternate assessments for students who are visually impaired and need large print or Braille or for students who are significantly physically impaired and need assistive technology.)

Table 5b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Fairness

Critical Element	Examples of Acceptable Evidence
3.2 (d) ELP assessments are of high technical quality, including being valid, reliable, and fair	Acceptable evidence includes technical manuals for ELP assessment(s), including scoring guides, and other documents that describe the ELP assessment(s).

Table 6a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.4 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.4 When different test forms or formats are used, the State must ensure that the meaning and interpretation of results are consistent. **(a)** Has the State taken steps to ensure consistency of test forms over time? **(b)** If the State administers both an online and paper and pencil test, has the State documented the comparability of the electronic and paper forms of the test?	The State has conducted appropriate equating or linking studies and has presented data that support the success of the equating or linking. *Possible Evidence* • Documentation describing the State's approach to ensuring comparability of assessments and assessment results across groups and time. • Documentation of equating studies that confirm the comparability of the State's assessments and assessment results across groups and across time, as well as follow-up documentation describing how the State has addressed any deficiencies.	The State has not conducted or documented equating studies to establish whether test forms are comparable across time.

Table 6b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Comparability

Critical Element	Examples of Acceptable Evidence
3.4 (b) If State plans to transition to a new ELP assessment, plan for doing so, including: How State plans to address "comparability" (relationship between old and new ELP assessment (i.e., use of double-testing, bridge studies, judgment procedures, data analysis, or other method).	Acceptable evidence includes plan for establishing comparability (e.g., use of double-testing, bridge studies, judgment procedures, data analysis, or other method), results if available, and plan for developing new AMAOs, if applicable.

Table 7a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.5 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.5 Has the State established clear criteria for the administration, scoring, analysis, and reporting components of its assessment system, including alternate assessment(s) and does the State have a system for monitoring and improving the on-going quality of its assessment system?	The State developed a set of management controls or standards for each of these components and has communicated these criteria to its contractor(s), LEAs, and schools. It requires its contractor(s) to provide specific information on the degree to which each criterion is met. The State uses an extensive system of training and monitoring to ensure that each person who is responsible for handling or administering any portion of its assessments does so in a way that protects the security of the assessments and maintains equivalence of administration conditions across students and schools. *Possible Evidence* • The State's criteria for administration, scoring, analysis, and reporting are communicated to its contractor(s). • The State's test security policy and consequences for violation are communicated to the public and to local educators. • Existing written documentation of the State's plan for training and monitoring assessment administration conditions across the State, even when its assessment system is comprised of only local assessments. • Documentation that the tests clearly delineate which accommodations may be used for specific sections of the test (e.g., specify the items/sections for which a calculator may be used without invalidating the test).	The State does not have a test security policy. The State does not train or monitor personnel at the school, LEA, and State levels with regard to its test administration procedures and security policy. The State provides no criteria to its contractor(s) regarding the quality control and security measures it requires for its assessment system. The State provides no criteria to its contractor(s) to ensure that the procedures for scoring of open-ended tasks meet industry standards for accuracy.

Table 7b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Test Administration, Scoring, and Reporting

Critical Element	Examples of Acceptable Evidence
3.2 (e) If multiple ELP assessments are being used, data can be aggregated for comparison and reporting purposes	Acceptable evidence includes description of how the State ensures that data can be aggregated for comparison and reporting purposes.
3.3 (a) (b) (c) Has the state established and implemented clear criteria for the administration, scoring, analysis, and reporting components of its ELP assessments, and does the State have a system for monitoring and improving the ongoing quality of its assessment systems? (Critical Element 3.3) **(a)** ELP assessments are administered in a uniform manner statewide. **(b)** Methods for administration, scoring, analysis, and reporting have been established. **(c)** The state monitors ELP assessment administration practices.	Acceptable evidence includes: • Test administration manuals; • Evidence of training on test administration, scoring guides, or other documentation that ELP assessments are administered in a uniform manner Statewide; • If accommodations were provided on the ELP assessment to students with disabilities, which accommodations, method for determining accommodations, and number and percentage of students receiving such accommodations; • Procedure used by State to ensure that criteria for administration, scoring, analysis, and reporting have been communicated to LEAs; • Evidence that the State monitors LEA/school administration of ELP assessments, including process for monitoring assessment administration; and • Documentation of the State's plan for training and monitoring assessment administration conditions.

Table 8a. Standards and Assessment Peer Review Guidance Section 4: Technical Quality—Critical Element 4.6 (USED, 2004)

Critical Element	Examples of Acceptable Evidence	Examples of Incomplete Evidence
4.6 Has the State evaluated its use of accommodations? **(a)** How has the State ensured that appropriate accommodations are available to students with disabilities and that these accommodations are used in a manner that is consistent with instructional approaches for each student, as determined by a student's IEP or 504 plan? **(b)** How has the State determined that scores for students with disabilities that are based on accommodated administration conditions will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration conditions? **(c)** How has the State ensured that appropriate accommodations are available to limited English proficient students and that these accommodations are used as necessary to yield accurate and reliable information about what limited English proficient students know and can do? **(d)** How has the State determined that scores for limited English proficiency students that are based on accommodated administration circumstances will allow for valid inferences about these students' knowledge and skills and can be combined meaningfully with scores from non-accommodated administration circumstances?	The State provides for the use of appropriate accommodations and has conducted studies to ensure that scores based on accommodated administrations can be meaningfully combined with scores based on the standard administrations. *Possible Evidence* • The State has analyzed the use of specific accommodations for different groups of students with disabilities and has provided training to support sound decisions by IEP teams. • The State routinely monitors the extent to which test accommodations are consistent with those provided during instruction. • The State has analyzed the effect of specific accommodations for students with limited English proficiency and has shared results with LEAs and schools. • Documentation of the quality and consistency of the accommodations it offers for limited English proficient students (e.g., training of translators, simplified English, standardized translation of instructions for test administration that are comparable to the regular assessment).	No analyses have been carried out to determine whether specific accommodations produce the effect intended. The State does not require that decisions about how students with disabilities will participate in the assessment system be made on an individual basis or specify that these decisions must be consistent with the routine instructional approaches as identified by each student's IEP and/or 504 plan. The State uses the same accommodations for limited English proficient students as it uses for students with disabilities.

Table 8b. Critical Elements from Title III OELA Monitoring Reports for ELL Assessments (2006) Related to Accommodations

Critical Element
Per Title III OELA Monitoring Reports, if accommodations are provided on the ELP assessment to students with disabilities, then the state should provide documentation of which accommodations were provided, the method for determining accommodations, and the number and percentage of students receiving such accommodations

Table 9. Crosswalk Between Critical Elements Identified in Standards and Assessment Peer Review Guidance (USED, 2004) and Evaluation of the Technical Evidence of Assessments for Special Student Populations (AACC, 2007)

Notes: Table 9 provides another overview of technical criteria for evaluating the quality of assessments. It lists validated technical criteria by *type* (validity, reliability, bias and sensitivity) and *evidence/method elements* one would expect to see in support of each type vis-à-vis the various aspects of test development (e.g., test design and development, item level, test level). These criteria are cross-referenced with the critical elements for technical quality identified in *Standards and Assessment Peer Review Guidance* (USED, 2004). An "X" indicates evidence that state officials might consider in order to support the technical quality (per *Standards and Assessment Peer Review Guidance*) of their assessments for special student populations. For more information about the technical criteria presented here, see the document titled *Evaluation of the Technical Evidence of Assessments for Special Student Populations* at www.aacompcenter.org (see Special Populations page).

TECHNICAL CRITERIA			PEER REVIEW CRITICAL ELEMENTS: TECHNICAL QUALITY					
			4.1	4.2	4.3	4.4	4.5	4.6
	TYPE	ELEMENT: EVIDENCE/METHOD	Validity	Reliability	Fairness/Access	Comparability	Administration, Scoring, Analysis, Reporting	Accommodations
Test Design and Development								
Item/Test level	Construct validity	Test purpose	X					
		Population/classification	X	X	X	X	X	
		Theoretical foundation/framework	X					
		Universal design	X		X			
		Readability	X		X		X	
Test Design and Development								
Item level	Content validity	Alignment (items-to-standards)	X				X	
		Linkage (items-to-standards, standards-to-standards)	X				X	
		Expert judgment	X				X	
		p-values/point biserials	X	X			X	
		IRT/item fit	X				X	
		Structural equation modeling	X				X	
		t-tests	X				X	
		ANOVA	X				X	
		Factor analysis	X				X	
Test Design and Development								
Test level	Construct validity	Equivalence/comparability	X			X		X
		Multi-trait/multi-method/subtest inter-correlation	X			X	X	
	Content validity	Test blueprint	X					
		Alignment (test form-to-blueprint)	X				X	

TECHNICAL CRITERIA			PEER REVIEW CRITICAL ELEMENTS: TECHNICAL QUALITY					
			4.1	4.2	4.3	4.4	4.5	4.6
	TYPE	ELEMENT: EVIDENCE/METHOD	Validity	Reliability	Fairness/Access	Comparability	Administration, Scoring, Analysis, Reporting	Accommodations
Test level	Content validity	Descriptive statistics (e.g., central tendency, variation)	X	X			X	
		IRT/test fit	X				X	
		Linking/equating	X			X		
	Criterion validity (predictive/concurrent)	Cross tabulations	X				X	
		Pearson correlation	X				X	
	Consequential validity	Use of results	X	X	X	X	X	
Test Design and Development								
Administration	Construct validity	Accommodation	X	X	X	X	X	X
		Fidelity	X			X		X
		Standardization		X	X		X	
Test Design and Development								
Item/Test Level	Reliability—Stability & consistency	Standard error of measurement/ confidence intervals		X			X	
		Test-retest		X			X	
		Alternate form		X		X	X	
	Reliability—Internal consistency	Coefficient alpha		X			X	
		KR-21		X			X	
		Test length/power estimates		X			X	
		Split-half		X			X	
	Reliability—Generalizability	G-coefficient		X			X	
	Reliability—Classification consistency	Correlation coefficient		X			X	
		Percent correspondence		X			X	
		Classification error		X			X	
	Bias and sensitivity—Linguistic	Expert review	X		X	X		
		DIF analysis					X	
	Bias and sensitivity—Ethnicity/race	Expert review	X		X	X		
		DIF analysis					X	
	Bias and sensitivity—Cultural/religious	Expert review	X		X	X		
	Bias and sensitivity—Geographic	Expert review	X		X	X		
		DIF analysis					X	
	Bias and sensitivity—SES	Expert review	X		X	X		
		DIF analysis					X	
	Bias and sensitivity—Disability	Expert review	X		X	X		
		DIF analysis					X	

TECHNICAL CRITERIA			PEER REVIEW CRITICAL ELEMENTS: TECHNICAL QUALITY					
			4.1	4.2	4.3	4.4	4.5	4.6
	TYPE	ELEMENT: EVIDENCE/METHOD	Validity	Reliability	Fairness/Access	Comparability	Administration, Scoring, Analysis, Reporting	Accommodations
Item/Test level	Bias and sensitivity—Gender	Expert review	X		X	X		
		DIF analysis					X	
Field Testing								
	Content validity	Blueprint	X					
		Sampling	X	X				
		Norming	X			X	X	
Scoring								
	Content validity	Rubric	X	X			X	
		Scale	X	X			X	
		Standard setting (cut score and proficiency levels)	X	X	X		X	
		Training of scorers/scoring protocol		X			X	
	Reliability—Inter-rater	Correlation (kappa)		X			X	
		Percent correspondence		X			X	
Reporting								
	Consequential validity	Reporting category	X				X	X
		N	X	X			X	
		Central tendency/variation	X	X			X	
		Effect size	X			X	X	
Security								
	Consequential validity	Protocols	X	X			X	X

Test Accommodations

Many of the Peer Review comments to states emphasized a need for more evidence or additional work in the areas of test accommodations (discussed further here) and standard setting (discussed in the next section).

The list of allowable accommodations for SWDs and for ELLs differs across states (National Research Council 2002, 2004; Rivera & Collum, 2004). Providing students with appropriate test accommodations is critical because appropriate access to assessments is necessary to improve the validity of the results, and valid assessments are critical if results are used for accountability purposes.

Test accommodations tend to fall into one of four categories: presentation, response, timing/scheduling, and setting. Presentation accommodations include alterations to the way in which the test is presented to students, such as an oral presentation or a Brailled version of the test. Response accommodations involve changes to the way students

Accommodations are intended to provide students with the maximally appropriate conditions to access the tested content and demonstrate their knowledge and skills.

are expected to provide their responses; such accommodations could include oral rather than written responses, or the use of an assistive device to demonstrate a response. Accommodations related to timing/scheduling may include extended time or frequent breaks during testing. And accommodations to setting include changes to the test location or conditions, such as administering the test individually or in a small group setting rather than in a regular classroom (Thurlow, House, Boys, Scott, & Ysseldyke, 1999). Accommodations are intended to provide students with the maximally appropriate conditions to access the tested content and demonstrate their knowledge and skills.

For SWDs served under IDEA, appropriate assessment accommodations should be consistent with IEP practices. Generally, the IEP must consider the student's present level of educational performance; that is, "…how the child's disability affects the child's involvement and progress in the general education curriculum…" (IDEA, 2004, Sec. 614 [d][1][A][i][I]). More specifically related to assessment, the IEP must include descriptions of "…any individual appropriate accommodations that are necessary to measure the academic achievement and functional performance of the child on state and district-wide assessments…" (IDEA, 2004, Sec. 614[d][1][A][i][VI][aa]). Thus, IDEA requires that the individual student's needs—rather than the student's disability category—should determine the appropriate accommodations for both instruction and assessment.

For ELLs the selection of accommodations should involve the consideration of the student's English language proficiency level as well as the extent to which the student has been instructed in the content of the test and the language of that instruction. In addition to the consideration of such student variables, the amount of appropriate direct linguistic support should be considered. All linguistic accommodations are intended to reduce the construct-irrelevant language demands on students in a test. That is, they are designed to reduce instances where the language of the test not associated with what is being assessed becomes a barrier to students' understanding of what is asked and how to respond. Direct linguistic support includes accommodations that address the construct-irrelevant language of the test (in either English or the student's native language). Examples of direct linguistic accommodation include oral presentation, linguistic simplification, in which the text is modified to reduce complex vocabulary and sentence structure, and bilingual glossaries or bilingual dictionaries, which allow students to translate unfamiliar terms. Indirect linguistic support accommodations also are used to reduce construct-irrelevant language barriers, but these supports usually address the testing conditions or environment (i.e., setting, schedule) (Center for Equity and Excellence in Education, 2005; Rivera & Collum, 2004).

The individual student's needs—rather than the student's disability category—should determine the appropriate accommodations for both instruction and assessment.

Currently, allowable accommodation practices vary greatly across states, and research on the effectiveness of accommodations for SWDs and ELLs is inconclusive. Nonetheless, much has been learned about test accommodations. Table 10 and Table 11 below list a selection of resources that state officials can use to inform their thinking about the appropriateness and effectiveness of various accommodations for SWDs and ELLs, respectively. These tables are not exhaustive because the body of rigorous research systematically examining the use of accommodations with SWDs and ELLs continues to grow[5]. In particular, more research is needed regarding other accommodations typically used with ELLs, such as accommodations related to presentation and response, which tend to lend themselves to accommodation of language. Common practice for selecting accommodations for ELLs suggests that decisions on accommodations often are based on research that focused on SWDs, rather than ELLs (Rivera & Collum, 2004).

[5] Lists of accommodations and relevant research/references will be updated as additional information becomes available and is reviewed using the AACC vetting criteria.

For English language learners, the selection of accommodations should involve the consideration of the amount of appropriate direct linguistic support needed vis-à-vis the student's English language proficiency level, the extent to which the student has been instructed in the content of the test, and the language of that instruction.

Table 10. Resources on Accommodations for Students with Disabilities

Resource / Accommodation	Presentation				Response	Timing/ Scheduling
	Calculators	Orally read directions/ oral presentation	Computer-assisted testing	Universal design	Dictation	Extended time & multi-day sessions
Calhoon, Fuchs, & Hamlett, 2000			X			
Fuchs, Fuchs, Eaton, Hamlett, Binkley, & Crouch, 2000					X	X
Fuchs, Fuchs, Eaton, & Karns, 2000	X	X				X
Johnstone, 2003				X		
Johnstone, Thompson, Moen, Bolt, & Kato, 2005				X		
Kosciolek & Ysseldyke, 2000		X				
Russell & Plati, 2000			X			
Thompson, Johnstone, & Thurlow, 2002				X		
Tindal, Heath, Hollenbeck, Almond, & Harniss, 1998		X				
Walz, Albus, Thompson, & Thurlow, 2000						X
Weston, 2002		X				

Note: Additional resources will be provided as they become available and are reviewed using the AACC vetting criteria.

Table 11. Resources on Accommodations for English Language Learners

Resource / Accommodation	Presentation				Timing/ Scheduling
	Orally read directions/ oral presentation	English dictionaries, customized dictionaries, & glossaries	Bilingual dictionaries & glossaries	Linguistic simplification & modification	Extended time & multi-day sessions
Type	Direct	Direct	Direct	Direct	Indirect
Abedi, 2001		X	X		
Abedi, Courtney, & Leon, 2003		X	X	X	
Abedi, Courtney, Mirocha, Leon, & Goldberg, 2005		X	X	X	
Abedi, Hofstetter, Baker, & Lord, 2001		X			X
Abedi, Hofstetter, & Lord, 2004		X		X	
Abedi & Lord, 2001				X	
Abedi, Lord, Hofstetter, & Baker, 2000		X		X	X
Abedi, Lord, Kim, & Miyoshi, 2000		X	X		
Albus, Bielinski, Thurlow, & Liu, 2001		X			
Castellon-Wellington, 2000	X				X
Kopriva, 2000, Ch. 6	X	X	X		X
Mazzeo, Carlson, Voelkl, & Lutkus, 2000	X	X*	X*		X
Rivera & Stansfield, 2004				X	

Note: Additional resources will be provided as they become available and are reviewed using the AACC vetting criteria.

*Resources do not specify whether glossaries discussed are monolingual or bilingual.

Key considerations regarding test accommodations

Research has shown that there are issues related to the validity of inferences drawn from the scores of students who have taken accommodated tests. Therefore, as state officials consider the appropriateness of their assessment accommodations, they ought to consider the following questions:

- Is the accommodation appropriate for the student/group of students?
- Does the accommodation give an unfair advantage to SWDs, ELLs, or to subgroups of either?
- Does the accommodation change the assessed construct?
- Does the accommodation (e.g., computer administration, assistive devices) change item/test comparability?

See resources listed in Tables 10 and 11 for relevant research.

Additional resources relevant to accommodations are as follows:

A Decision Framework for IEP Teams Related to Methods for Individual Student Participation in State Accountability Assessments, 2005
http://www.ed.gov/admins/lead/speced/toolkit/iep-teams.doc

National Council on Disability: *Improving Educational Outcomes for Students with Disabilities*, 2004
http://www.ncd.gov/newsroom/publications/2004/educationoutcomes.htm

Office of Special Education Programs (OSEP)
http://www.ed.gov/about/offices/list/osers/osep/index.html?src=mr

Office of English Language Acquisition (OELA) with link to National Dissemination Center for Children with Disabilities
http://www.ed.gov/about/offices/list/oela/index.html?src=oc

OELA National Clearinghouse (NCELA)
http://www.ncela.gwu.edu/

National Center for Research on Evaluation, Standards, and Student Testing (CRESST)
http://www.cresst.org/

Council of Chief State School Officers (CCSSO)
http://www.ccsso.org/

National Center on Educational Outcomes (NCEO)
http://education.umn.edu/nceo/

NCEO Online Accommodations Bibliography
http://education.umn.edu/NCEO/AccomStudies.htm

Council for Exceptional Children (CEC)
http://www.cec.sped.org/

Center for Equity and Excellence in Education
http://ceee.gwu.edu/

National Alternate Assessment Center
http://www.naacpartners.org/

Note: Additional resources will be provided as they become available and are reviewed using the AACC vetting criteria.

Standard Setting

Setting defensible cut scores and establishing meaningful performance levels are key concerns for state departments of education. While there are a number of standard setting methods used across states, there is no agreed-upon best method for setting standards (Berk, 1986; Linn, 2003).

Here are general descriptions of several standard setting methods:

- *Reasoned judgment*: The full range of possible scores (score scale) is divided into categories determined by experts. Exemplars and decision rules are used to connect descriptors with student work (Kingston, Kahl, Sweeney, & Bay, 2001).
- *Contrasting groups:* Comparisons are made between the expected performance and actual performance of different ability groups. Prior to testing, teachers familiar with the students separate students into pre-defined ability groups. The distribution of test scores across the groups is then examined (Livingston & Zeikey, 1982).
- *Modified Angoff:* Experts examine the test items and estimate the percentage of students at the bottom of the score range who will be able to pass each item. The estimates are summed and result in an overall percentage of items correct that correspond to the minimum passing score for a given level. This is typically used with multiple-choice items (Berk, 1986).
- *Bookmarking:* Experts review an ordered item booklet that contains test items arranged in order of difficulty. The experts are asked to mark the places in the booklet (i.e., between sequential items) where the skill range for one level ends and the next begins (Lewis, Mitzel, & Green, 1996).
- *Body of work:* Experts examine all student work and use this information to place the student in a performance level. Standard-setters are given a set of papers that exemplify the complete range of possible scores from low to high. Thus, for a given student, standard-setters determine which performance level placement most reasonably reflects the work of that student (Kingston, Kahl, Sweeney, & Bay, 2001).

As mentioned previously, there is no agreed-upon best method for setting standards, and research suggests that there is considerable variability in the standards set across methods due to, for example, variability across groups of standard-setters as well as variability due to the methods themselves (Jaeger, 1989).

Therefore, the use of multiple standard setting methods, with the results of the different methods considered together to determine cut scores (Jaeger, 1989) seems apt. Although the use of multiple methods may be cost prohibitive, such practice warrants consideration, given the consequences associated with the results of standard setting efforts.

Resources for standard setting

Guidelines and criteria are available for the selection and implementation of a standard setting method or methods. The following resources contain such guidelines and considerations for general education assessments.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Raymond, M. R., & Reid, J. B. (2001). Who made thee judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum.

The following resources offer guidelines and considerations for defensible adaptation of traditional general education standard setting methods for tests for students with disabilities.

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (NCEO Synthesis Report 47). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Roeber, E. (2002). *Setting standards on alternate assessments* (NCEO Synthesis Report 42). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Thurlow, M. L., & Ysseldyke, J. E. (2001). Standard-setting challenges for special populations. In G. Cizek, (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 387–410). Mahwah, NJ: Lawrence Erlbaum.

Additional resources relevant to standard setting are as follows:

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

Cizek, G. (2001). *Setting performance standards: Concepts, methods, and perspectives*. Mahwah, NJ: Lawrence Erlbaum.

Mitzel, H. C. (2005). *Consistency for state achievement standards under NCLB*. Paper presented to CAS SCASS Study Group. Washington, DC: Council of Chief State School Officers.

Note: Additional resources will be provided as they become available and are reviewed using the AACC vetting criteria.

Annual Measurable Achievement Objectives (AMAOs) for ELLs

An area of need across states that is requiring more attention is the setting of Annual Measurable Achievement Objectives for ELLs. NCLB's Title III requires that each state establish three AMAOs.

- AMAO 1: The number or percentage of ELLs making progress toward English language proficiency (one level per year) until reaching proficiency.
- AMAO 2: The annual increase in the number or percentage of students attaining English language proficiency.
- AMAO 3: As a subgroup (per Title I), ELLs' adequate yearly progress (AYP) toward meeting grade-level academic achievement standards in English language arts and math.

States are accountable for meeting their AMAOs, and receipt of Title III funding is contingent on this.

As of 2007, all 50 states report having an English language proficiency assessment for their ELL students. All states have set their AMAOs (U.S. Office of Management and Budget and Federal Agencies, 2006). However, many have not set all three AMAOs, and AMAOs vary widely across states, making cross-state comparisons difficult (Center on Education Policy, 2006). In order to provide states with information related to setting and monitoring progress toward meeting AMAOs, the AACC has identified the following resources for their consideration.

Resources regarding AMAOs

The following resources offer considerations for states in relation to their AMAOs. This list will be updated as additional resources are reviewed using the AACC vetting criteria.

U.S. Congress. (2002). *No Child Left Behind Act of 2001.* Public Law 107–110, 107th Congress. Washington, DC: Government Printing Office.

Center on Education Policy. (2006, March). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Author.

Note: Additional resources will be provided as they become available and are reviewed using the AACC vetting criteria. ❖

References

Abedi, J. (1999, April). *Examining the effectiveness of accommodation on math performance of English language learners*. Paper presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada.

Abedi, J. (2001). *Validity of accommodations for English language learners*. Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Abedi, J. (2004). The No Child Left Behind Act and English language learners: Assessment and accountability issues. *Educational Researcher*, 33(1), 4–14.

Abedi, J., Courtney, M., & Leon, S. (2003). *Effectiveness and validity of accommodations for English language learners in large-scale assessments* (CSE Report No. 608). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., Courtney, M., Mirocha, J., Leon, S., & Goldberg, J. (2005). *Language accommodations for English language learners in large-scale assessments: Bilingual dictionaries and linguistic modification* (CSE Report No. 666). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., & Dietel, R. (2004). Challenges in the No Child Left Behind Act for English language learners. *Phi Delta Kappan*, 85, 782–785.

Abedi, J., & Hejri, F. (2004). Accommodations for students with limited English proficiency in the National Assessment of Educational Progress. *Applied Measurement in Education*, 17, 371–392.

Abedi, J., Hofstetter, C., Baker, E., & Lord, C. (1998). *NAEP math performance and test accommodations: Interactions with student language background* (CSE Technical Report No. 536). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., Hofstetter, C., & Lord, C. (2004). Assessment accommodations for English language learners: Implications for policy-based empirical research. *Review of Educational Research*, 74, 1–28.

Abedi, J., Leon, D., & Mirocha, J. (2001). *Impact of students' language background on standardized achievement test results: Analyses of extant data.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., & Lord, C. (2001). The language factor in mathematics tests. *Applied Measurement in Education*, 14, 219–234.

As mentioned at the beginning of this document, the Assessment and Accountability Comprehensive Center (AACC) will update these guidelines as new, relevant research, guidance, and strategies become available. Additionally, these guidelines are designed to evolve with the changing needs of RCCs and states. Future guidelines will include relevant information and research related to: ELL alignment and linkage; linguistic modification and access strategies; and the framework for English language proficiency standards and assessments.

For questions related to these guidelines, please contact:
Edynn Sato, Ph.D.
Director, Special Populations
Assessment and Accountability Comprehensive Center, WestEd
email: esato@wested.org.

Please cite as: Sato, E., Rabinowitz, S., Worth, P., Gallagher, C., Lagunoff, R., & McKeag, H. (2007). *Guidelines for Ensuring the Technical Quality of Assessments Affecting English Language Learners and Students with Disabilities: Development and Implementation of Regulations*. (Assessment and Accountability Comprehensive Center report). San Francisco: WestEd.

For more information about the AACC, visit www.aacompcenter.org.

The contents of this guide were developed under a grant from the Department of Education. However, those contents do not necessarily represent the policy of the Department of Education, and you should not assume endorsement by the Federal Government.

Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice, 19*(3), 16–26.

Abedi, J., Lord, C., Kim, C., & Miyoshi, J. (2001). *The effects of accommodations on the assessment of LEP students in NAEP* (CSE Technical Report No. 537). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., Lord, C., & Plummer, J. R. (1995). *Language background as a variable in NAEP mathematics performance: NAEP TRP task 3d: Language background study.* Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Abedi, J., Lord, C., & Plummer, J. R. (1997). *Final report of language background as a variable in NAEP mathematics performance* (CSE Technical Report No. 429). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Albus, D., Bielinski, J., Thurlow, M., & Liu, K. (2001). *The effect of a simplified English language dictionary on a reading test* (LEP Projects Report No. 1). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved September 26, 2007, from http://education.umn.edu/NCEO/OnlinePubs/LEP1.html

American Diploma Project. (2004, February). *Ready or not: Creating a high school diploma that counts.* Washington, DC: Achieve.

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

Anastasi, A. (1988). *Psychological testing.* New York: Macmillan.

Assessment and Accountability Comprehensive Center. (2007). *Evaluation of the technical evidence of assessments for special student populations.* San Francisco: WestEd.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56,* 137–172.

Berkowitz, D., Wolkowitz, B., Fitch, R., & Kopriva, R. (2000). *The use of tests as part of high-stakes decision-making for students: A resource guide for educators and policy-makers.* Washington, DC: U.S. Department of Education Office for Civil Rights.

Calhoon, M. B., Fuchs, L. S., & Hamlett, C. L. (2000). Effects of computer-based test accommodations on mathematics performance assessments for secondary students with learning disabilities. *Learning Disability Quarterly, 23,* 271–282.

Castellon-Wellington, M. (2000). *The impact of preference for accommodations: The performance of ELLs on large-scale academic achievement tests* (CSE Technical Report No. 524). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Center for Equity and Excellence in Education. (2005). *Recommendations for state ELLs accommodation policies.* Retrieved May 31, 2006, from http://ceee.gwu.edu/AA/Accommodations_Recos.html

Center on Education Policy. (2005, March). *From the capital to the classroom: Year 3 of the No Child Left Behind Act.* Washington, DC: Author.

Center on Education Policy. (2006, March). *From the capital to the classroom: Year 4 of the No Child Left Behind Act.* Washington, DC: Author.

Cizek, G. (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Education Week. (2004). Quality counts. *Education Week*, 23(17).

Education Week. (2005). Quality counts. *Education Week*, 24(17).

Elliot, S. N., Kratochwill, T. R., McKevitt, B., Schulte, A. G., Marquart, A., & Mroch, A. (1999, June). *Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities: Mid-project results.* Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference, Snowbird, UT.

Fast, E. F., & Hebbler, S. (2004). *A framework for evaluating validity in state accountability systems.* Washington, DC: Council of Chief State School Officers.

Fuchs, L. S., Fuchs, D., Eaton, S. B., Hamlett, C. L., Binkley, E., & Crouch, R. (2000). Using objective data sources to enhance teacher judgments about test accommodations. *Exceptional Children*, 67, 67–81.

Fuchs, L. S., Fuchs, D., Eaton, S. B., & Karns, K. M. (2000). Supplementing teacher judgments of mathematics test accommodations with objective data sources. *School Psychology Review*, 29(1), 65–85.

Gong, B. (2005). *Documenting validity of accountability systems: A progress report.* Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing, Center for the Study of Evaluation.

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89–116). Mahwah, NJ: Lawrence Erlbaum.

Helwig, R., Stieber, S., Tindal, G., Hollenbeck, K., Heath, B., & Almond, P. (2000). *A comparison of factor analyses of handwritten and word-processed writing of middle school students.* Eugene, OR: RCTP.

Herman, J. L., & Dietel, R. (2005). *What school board leaders should know about standards-based accountability.* Los Angeles: University of California, National Center for Research on Evaluation, Standards and Student Testing, Center for the Study of Evaluation.

Hollenbeck, K., Tindal, G., & Almond, P. (1998). Teachers' knowledge of accommodations as a validity issue in high-stakes testing. *Journal of Special Education,* 32(3), 175–183.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Johnson, E., Kimball, K., & Brown, S. (2001). American sign language as an accommodation during standards-based assessments. *Assessment for Effective Intervention,* 26(2), 39–47.

Johnstone, C. J. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report No. 37). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 1, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Technical37.htm

Johnstone, C. J., Thompson, S. J., Moen, R. E., Bolt, S., & Kato, K. (2005). *Analyzing results of large-scale assessments to ensure universal design* (Technical Report No. 41). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved, May 19, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Technical41.htm

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 53–88). Mahwah, NJ: Lawrence Erlbaum.

Kane, M. T. (2002). Validating high-stakes testing programs. *Educational Measurement: Issues and Practice, 21*(1), 31–41.

Kingston, N., Kahl, S. R., Sweeney, K., & Bay, L. (2001). Setting performance standards using the body of work method. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 219–248). Mahwah, NJ: Lawrence Erlbaum.

Kopriva, R. (2000). *Ensuring accuracy in testing for English language learners.* Washington, DC: Council of Chief State School Officers.

Kosciolek, S., & Ysseldyke, J. E. (2000). *Effects of a reading accommodation on the validity of a reading test* (Technical Report No. 28). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996, June). *Standard setting: A bookmark approach.* Paper presented at the Council of Chief State School Officers Large-Scale Assessment Conference, Colorado Springs, CO.

Linn, R. L. (2003). Performance standards: Utility for different uses of assessments. *Education Policy Analysis Archives, 11*(31). Retrieved March 8, 2006, from http://epaa.asu.edu/epaa/v11n31

Linn, R. L., Baker, E. L., & Betebenner, D. W. (2002). Accountability systems: Implications of requirements of the No Child Left Behind Act of 2001. *Educational Researcher, 31*(6), 3–16.

Livingston, S. A., & Zeikey, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests.* Princeton, NJ: Educational Testing Service.

Mazzeo, J., Carlson, J. E., Voelkl, K. E., & Lutkus, A. D. (2000). *Increasing the participation of special needs students in NAEP: A report on 1996 NAEP research activities* (NCES Publication No. 2000-473). Washington, DC: National Center for Education Statistics.

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

Mitzel, H. C. (2005). *Consistency for state achievement standards under NCLB.* Paper presented to CAS SCASS Study Group. Washington, DC: Council of Chief State School Officers.

National Research Council. (2002). In J. Koenig (Ed.), *Reporting test results for students with disabilities and English-language learners.* Washington, DC: National Academies.

National Research Council. (2004). In J. Koenig & L. Bachman (Eds.), *Keeping score for all: The effects of inclusion and accommodation policies on large-scale educational assessments.* Washington, DC: National Academies.

No Child Left Behind Act of 2001, 20 U.S.C. 6301 *et seq.* (2002).

Olson, B., Mead, R., & Payne, D. (2002). *A report of a standard setting method for alternate assessments for students with significant disabilities* (NCEO Synthesis Report No. 47). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Rabinowitz, S. (2004). *Design considerations for building out NCLB state assessment systems.* San Francisco: WestEd.

Rabinowitz, S., & Ananda, S. (2002). *Assessment and accountability provisions for the No Child Left Behind Act: Key decision points for states.* San Francisco: WestEd.

Rabinowitz, S., & Sato, E. (2005). *Evidence-based plan: Technical adequacy of assessments for alternate student populations. A technical review of high-stakes assessments for English language learners.* San Francisco: WestEd.

Rabinowitz, S., & Sato, E. (2006). *The technical adequacy of assessments for alternate student populations: Guidelines for consumers and developers.* San Francisco: WestEd.

Raymond, M. R., & Reid, J. B. (2001). Who made thee judge? Selecting and training participants for standard setting. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 119–157). Mahwah, NJ: Lawrence Erlbaum.

Rivera, C., & Collum, E. (2004, January). *An analysis of state assessment policies addressing the accommodation of English language learners.* Issue paper commissioned for the National Assessment Governing Board Conference on Increasing the Participation of SD and LEP Students in NAEP. Arlington, VA: George Washington University, Center for Equity and Excellence in Education.

Rivera, C., & Stansfield, C. W. (1998). Leveling the playing field for English language learners: Increasing participation in state and local assessments through accommodations. In R. Brandt (Ed.), *Assessing student learning: New rules, new realities* (pp. 65–92). Arlington, VA: Educational Research Service.

Rivera, C., & Stansfield, C. W. (2004). The effect of linguistic simplification of science test items on score comparability. *Educational Assessment*, 9(3&4), 79–105.

Roeber, E. (2002). *Setting standards on alternate assessments* (NCEO Synthesis Report No. 42). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Russell, M., & Plati, T. (2000). M*ode of administration effects on MCAS composition performance for grades four, eight, and ten. A report of findings submitted to the Massachusetts Department of Education.* Chestnut Hill, MA: National Board on Educational Testing and Public Policy.

Schulte, A. G., Elliott, S. N., & Kratochwill, T. R. (2000). Educators' perceptions and documentation of testing accommodations for students with disabilities. *Special Services in the Schools, 16,* 35–56.

Shepard, L., Taylor, G., & Beterbenner, D. (1998). *Inclusion of limited-English-proficient students in Rhode Island's grade 4 mathematics performance assessment* (CSE Technical Report No. 486). Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing, Center for the Study of Evaluation.

Sireci, S. G. (2004). Computerized adaptive testing: An introduction. In J. E. Wall & G. R. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 685–694). Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.

Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large-scale assessments* (Synthesis Report No. 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved September 11, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html

Thurlow, M. L., & Ysseldyke, J. E. (2001). Standard-setting challenges for special populations. In G. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Thurlow, M. L., House, A., Boys, C., Scott, D., & Ysseldyke, J. (2000). *State participation and accommodation policies for students with disabilities: 1999 update* (Synthesis Report No. 33). Minneapolis: University of Minnesota, National Center on Educational Outcomes. Retrieved June 30, 2006, from http://education.umn.edu/NCEO/OnlinePubs/Synthesis33.html

Tindal, G., Heath, B., Hollenbeck, K., Almond, P., & Harniss, M. (1998). Accommodating students with disabilities on large-scale tests: An experimental study. *Exceptional Children, 64,* 439–450.

Tindal, G., & Ketterlin-Geller, L. R. (2004). *Research on mathematics test accommodations relevant to NAEP Testing.* Washington, DC: National Assessment Governing Board. Retrieved March 21, 2006, http://www.nagb.org/pubs/conferences/tindal.pdf

U.S. Department of Education, Office of Elementary and Secondary Education. (2004). *Standards and assessments peer review guidance: Information and examples for meeting requirements of the No Child Left Behind Act of 2001.* Washington, DC: Author.

U.S. Department of Education, Office of English Language Acquisition, Language Enhancement, and Academic Achievement for Limited English Proficient Students. (2006). *Title III OELA Monitoring Reports.* Retrieved May 5, 2006 from http://www.ncela.gwu.edu/oela/OELAprograms/MonitoringReports

U.S. Office of Management and Budget and Federal Agencies. (2006). *Detailed information on the English language acquisition state grants assessment.* Retrieved July 30, 2007, from http://www.whitehouse.gov/omb/expectmore/detail/10003329.2006.html

Uzzell, L. A. (2005). *No Child Left Behind—The dangers of centralized education policy* (Policy Analysis Report No. 544). Washington, DC: CATO Institute.

Walz, L., Albus, D., Thompson, S., & Thurlow, M. (2000). *Effect of a multiple day test accommodation on the performance of special education students* (Report No. 34). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Weston, T. J. (2002, July). *The validity of oral accommodation in testing.* Paper presented to the NAEP Validity Studies Panel, Washington, DC.