WestEd excellence in research, development, & service

The Technical Adequacy of Assessments for Alternate Student Populations

GUIDELINES FOR CONSUMERS AND DEVELOPERS

By Stanley N. Rabinowitz and Edynn Sato

technical GUIDELINES

Abstract

These guidelines are designed to assist developers and consumers of assessments for English language learners in particular and special student populations in general. They are intended to help evaluate the technical adequacy (i.e., validity, reliability, freedom from bias) of assessments used to meet relevant Title I and Title III requirements under No Child Left Behind. This document includes a comprehensive set of validated criteria against which to review the technical evidence associated with assessments for special student populations. These criteria are sensitive to the unique characteristics of the student population, the particular purpose of the assessment, and the stage of development and maturity of the assessment. Also included in this document are examples of promising practices for ensuring validity, reliability, and freedom from bias, as well as a framework for integrating this technical evidence in response to federal requirements such as the NCLB-driven peer review of assessment programs. Using these guidelines, assessment developers and consumers will be able to gauge the technical adequacy of their assessments for special student populations and identify appropriate and necessary next steps for ensuring their validity and, ultimately, the defensibility of their assessment systems and results.

Introduction

Students who are English language learners (ELLs) are held accountable in two ways under the No Child Left Behind Act (NCLB): they must meet state-developed Annual Measurable Achievement Objectives (AMAOs) under Title III, and as a legislatively mandated subgroup, they must meet Adequate Yearly Progress (AYP) expectations for reading and math under Title I. Under the Title III requirements, states must show increases in the number and percent of ELL students (a) making progress in learning English and (b) attaining English proficiency. The Title I AYP requirement, on the other hand, relates to the need for this subgroup of students to meet rigorous state academic achievement standards.

The technical adequacy (i.e., validity, reliability, freedom from bias) of the assessments used to gather evidence of meeting AMAOs and AYP is critical, since test results impact related rewards and sanctions. Yet, until recently the technical adequacy of assessments for ELL students had not been examined in any methodical fashion. During 2005, Rabinowitz and Sato (n.d.) assembled available documentation related to such assessments and subjected the documents to a rigorous review for scientific quality. Results revealed great variability in the type and quality of technical evidence provided, potentially leaving states hard pressed to defend the technical adequacy of their ELL assessments and, consequently, the validity and reliability of their assessment results.

The guidelines presented in this document, an outcome of Rabinowitz and Sato's technical review, are intended to assist assessment developers and consumers in evaluating the technical adequacy of their ELL assessments in particular and assessments for special student population in general. Collectively, the guidelines present a systematic method for documenting and gauging the quality of the evidence states need in order to establish the technical adequacy of the assessments they are using to measure the progress of their special student populations, and, therefore, the effectiveness of their programs for preparing these students to meet challenging academic achievement standards.

Background

Recent developments in education have converged to create a critical need for valid, reliable, unbiased methods for conducting high-stakes assessments for students who are English language learners (Kubiszyn & Borich, 2003; Heubert & Hauser, 1999). Foremost is the movement toward ensuring education accountability for all students in a school system. In this effort, testing plays the central role of supplying evidence to parents, policymakers, politicians, and taxpayers about the degree to which students meet high standards. The high stakes attached to such assessments include more than decisions regarding the individual student

(e.g., achievement, promotion, graduation); they also include decisions regarding the quality of education in individual schools, districts, and the state as a whole. Accordingly, the interest in having a variety of accessible assessments that measure high content standards while meeting strict technical requirements has increased. Another development has been the demographic shift toward larger numbers and a larger proportion of public school students who speak a language other than English at home and are not yet proficient in academic English. As English language learners, these students are entitled to full and fair access to educational opportunities, including full inclusion in their state's assessment program. These trends merged with the signing of NCLB in 2002; all states are now required to conduct assessments to ascertain whether all children — including traditionally exempted special populations — are meeting high standards.

Assessing the linguistic progress of English learners and their proficiency in English per Title III has created some challenges for states, which must set goals for student language development and achievement of proficiency (AMAOs), define levels of proficiency, and measure growth. In measuring students' attainment of these goals, they must be certain that the assessments they use are valid for their needs, as well as appropriate for the ELL student population.

Assessing the achievement of English learners in content areas also has created a range of difficulties for states implementing the high expectations of NCLB per Title I[1].

To assess this population, states implement one or more of the following practices: (1) provide accommodations that suit the particular needs and characteristics of this student group (Center for Equity and Excellence in Education, 2005), (2) incorporate principles of Universal Design (Bowe, 2000; Rochester, 2004), and (3) develop alternative assessments based on the same content and performance standards used for the general student population. Each approach presents considerable technical challenges, especially in regard to the critical requirement that assessments be valid, reliable, and free of bias.

Policymakers and test developers have attempted in various ways to implement these seemingly straightforward practices. However, demonstrating technical adequacy has proven difficult, undermining states' confidence in how they judge the academic achievement of their special student populations (Rabinowitz, Ananda, & Bell, 2005). Key challenges that must be overcome so that assessment results can better support and inform policy include:

- ***Demonstrating technical adequacy.*** Once targeted tests are developed for a particular subgroup, the state must demonstrate that they meet the same technical standards as are applied to the equivalent tests for the general student population including:

  > **Lack of Bias** - Most methods to examine bias have focused on gender and ethnicity concerns. Because the English learner population tends to be more heterogeneous than typically has been assumed of males or African Americans as a group, for example, these methods will need some degree of reconceptualization to work with assessments for English learners.

  > **Reliability and Validity** - Reliability and validity studies are expensive and are difficult to implement with relatively small numbers of English learner students. At a minimum, evidence must show that the targeted assessments are at least as reliable and valid for the subpopulation of students as are other tests in the testing program. More ideally, evidence should show that the targeted assessments (and any approved accommodations) add sufficient validity for the targeted student population to justify the additional costs. Absent that evidence, states would be hard-pressed to justify costs of targeted assessments for special populations. What constitutes sufficient incremental validity to justify the use of these assessments is subject to debate and must be studied more fully in the future.

- ***Ensuring consistency of meaning across assessments for general population students and those who are English learners***. As Universal Design principles get more fully implemented and the list of allowable accommodations grows, ensuring comparability of content and construct across a range of assessments becomes more difficult. When alternative assessments are added to the equation, the difficulty of ensuring consistency of meaning for assessment scores increases yet again. This issue is especially relevant for those states attempting to use assessments developed for Title III purposes to assess the academic achievement of their ELL population per the requirements of Title I.

- ***Developing various support documents***. As the use of accommodations and alternative formats grows, so does

[1] NCLB provisions for assessing the academic achievement of English learners under Title I require that states and districts:

- assess in a valid and reliable manner;
- make reasonable accommodations;
- use the language and form "most likely to yield accurate and reliable information (to the extent practicable)"; and
- make every effort to develop linguistically accessible academic assessment measures.

the need for a wider range of support materials (e.g., manuals, answer documents, score reports, interpretive guides), potentially produced in a range of languages.

A good deal of research has been conducted on how to provide students who are English learners with access to valid, reliable, and bias-free assessments to measure their academic achievement. For most large-scale assessments, a wide range of technical data and evidence is available, most commonly in the form of technical manuals and reports. Research studies have been conducted to evaluate the performance of some of these assessments. Other evidence, such as materials presented in conferences and workshops, or internal working papers and reports, also exists. Yet, until recently these various sources of evidence had not been aggregated in any methodical fashion; nor had a protocol existed to evaluate the technical quality (i.e., validity, reliability, freedom from bias) of the evidence used to determine technical adequacy.

In their technical analysis of state assessments, Rabinowitz and Sato assembled available documentation related to these assessments for ELL students and used rigorous criteria to review the scientific quality of the evidence. Their technical analysis aimed to address the following questions:

1. What is the technical adequacy of high-stakes statewide assessments developed to assess English learner student populations, that is, is there adequate supporting evidence that assessments are valid, reliable, and unbiased?

2. What protocol and criteria can be used to evaluate the quality of evidence (i.e., validity, reliability, freedom from bias) presented?

3. What are "state-of-the-art" practices for developing, planning, and implementing bias, reliability, and validity studies for assessments developed to measure the performance of special student populations, in particular, English language learners?

The information presented in these guidelines is derived from this technical analysis. Included are

- A comprehensive set of validated criteria against which assessment consumers and developers can review the technical evidence of their assessments for special student populations;

- Examples of promising practices for providing evidence of validity, reliability, and freedom from bias appropriate for special student population assessments; and

- A framework for integrating technical evidence in response to federal requirements (i.e., Peer Review).

I. Technical Criteria

A worksheet for documenting and evaluating the technical evidence associated with an assessment geared toward ELL students (and that is adaptable for other special populations) is available in Appendix A on page 11. Technical evidence can be found in sources such as journal articles, conference presentations, technical and administration manuals, technical reports, and Web postings. The criteria included on the worksheet are described below starting on page 5.

PURPOSE, POPULATION, AND CONTENT

A technical evaluation of assessments for special student populations necessarily requires more than an examination of the methods and outcomes used to substantiate claims of validity, reliability, and freedom from bias; it must also consider the content and context of the assessment. More specifically, technical evidence must be evaluated with respect to the particular special population (e.g., ELL) and the specific purpose(s) of the assessment. Information that should be considered when judging the technical adequacy of evidence associated with an assessment includes:

- Statement of the test's purpose;

- Description of the target population and definitions of key characteristics; and

- Description of the content and constructs assessed.

Such information is important because it reflects the assumptions upon which the test was designed. If this foundational information was incomplete or inaccurate, the assessment might not be appropriate for the intended population or purposes and, consequently, the validity, reliability, and defensibility of the measure and its results would be questionable. The following discussion of purpose, target population, and content and constructs applies specifically to assessments for students who are English learners.

Purpose

An assessment's purpose for ELL students usually falls into one of the following categories (Zehler, Hopstock, Fleischman, & Greniuk, 1994):

1. Identification — to initially determine whether a student is Limited English Proficient

2. Placement — to assign a student to appropriate services (based on the proficiency level designations)

3. Placement Review — to review a student's service placement (thereby assessing progress)

4. Exit — to review a student's proficiency status

Under NCLB, states need to be cognizant of the purpose of the assessment(s) they use in order to ensure that the student performance data yielded are appropriate for the specific required accountability reports.

Target Population

To ensure that an assessment is valid and reliable for all populations of students, including the full range of English learner subgroups, consumers need to verify that the assessment was field-tested with a population of students that reflects the same demographic background, including education and language background (e.g., home language, native language, level of English proficiency, length of time in a U.S. school), as those students with whom the assessment will be used. Additionally, consumers should be clear about the definitions or descriptions test developers are using for key characteristics of the target population, as well as the methodology used to classify the population (e.g., English Only, Fluent English Proficient). If the test consumer and developer define the population of students differently or if students are classified according to methods that are traditionally difficult to standardize (e.g., teacher judgment), the interpretation of student scores may be unreliable. Such understanding will inform the appropriate identification of students for the test's administration as well as support valid interpretation of results.

Content and Constructs

According to the documents reviewed by Rabinowitz and Sato, in many instances a test developer first selected an assessment or test-item pool and then conducted an alignment between the items and the state English language development standards. Only in cases where states developed a new assessment or augmented an existing assessment was a test blueprint used, with test items developed specifically to state standards. To ensure that states are adequately assessing the full breadth and depth of a domain, as required by NCLB, states or test developers should articulate up front the range of skills and concepts they believe should be assessed in order to appropriately determine the language proficiency of a student.

SUFFICIENCY, QUALITY, AND ADEQUACY

Once the purpose, target population, and content of an assessment are determined to be an appropriate match for the consumer's needs, the next step is to evaluate the nature of the technical evidence provided — its sufficiency, quality, and adequacy. More specifically, consumers should consider the following:

- Information Provided: Is the information an assertion, a summary, or a detailed description?
- Type of Data: Are the data provided quantitative, qualitative, or both?
- Sufficiency: Is the information comprehensive (e.g., quantitative information is presented with supporting textual context and interpretation)?
- Quality: Does the method satisfy statistical assumptions? Is the method replicable? Is the outcome accurate (i.e., minimal or acceptable measurement error)? Is the outcome generalizable and/or broadly applicable?
- Adequacy: Is the information credible? Does the information directly support the evidence being evaluated?

These considerations should be applied to the technical criteria presented below.

CRITERIA

The following criteria are based on widely known and respected standards for technical adequacy of assessments and research (e.g., What Works Clearinghouse 2004a & b; American Educational Research Council, American Psychological Association, & National Council on Measurement in Education, 1999; Becker & Camilli, 2004), as well as on principles underlying rigorous, scientifically based research. The appropriateness and comprehensiveness of these criteria have been corroborated by a panel of experts in assessment, measurement, and applied linguistics.[2] The criteria are organized according to the following three categories, each reflecting a different level of specificity: (1) Criteria Clusters (i.e., validity, reliability, testing system criteria), the broadest category; (2) Criterion (e.g., field-testing, content validity, criterion validity, freedom from bias, reporting); and (3) Specific Evidence (e.g., randomization, accommodations, p-values, bias/DIF analyses, Universal Design, internal consistency, equating, scaling), the most narrowly defined category. These three categories allow for the variation in specificity of evidence that assessments may have, depending on their stage of development and maturity.

[2] The expert panel convened for this study included Dr. Jamal Abedi (UCLA/UC Davis/CRESST), Dr. Frances Butler (CRESST), and Dr. Steven Sireci (University of Massachusetts).

Technical Criteria

Criteria Cluster	Criterion	Specific Evidence[3]
Validity	Field Testing	Field Test Sampling Design: Representativeness and Norming
		Field Test Sampling Design: Currency (at least, dates documented)
		Field Test Sampling Design: Randomization
		Fidelity (link of test to stated purpose of the test)
	Design	Attrition of Persons (for Pre/Post Designs)
		Test Blueprint
		Scoring Rubric for OE Items: Construction and Validation
		Accommodations
	Content	Content Alignment Studies
		Expert judgments
		p-values
		Discrimination (Item-test Correlations)
		Bias/DIF analysis
		IRT/Item fit (ICC)
		Distractor Analysis
	Construct	Factorial Validity (structural equation modeling)
		Multi-Trait/Multi-Method
		Equivalence/Comparability (construct the same regardless of examinee's ability)
	Criterion	Predictive validity - Validation to the Referent
		Predictive validity - Individual and group scores
		Concurrent validity - Validation to External Criteria
		Concurrent validity - Validity of External Criteria
		Concurrent validity - Individual and group scores
	Consequential	Evaluation of Testing Consequences
		Individual and group scores
	Freedom from Bias	Content
		Ethnicity
		Cultural
		Linguistic
		Socio-economic
		Geographic
		Students with disabilities
		Universal Design

[3] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

Technical Criteria (continued)

Criteria Cluster	Criterion	Specific Evidence[3]
Reliability	Reliability: Single Administration	Scale
		Internal Consistency
		Split-half
		Scorer / Hand-scoring
	Reliability: Multiple Administrations	Test-retest
	Reliability: Either Single or Multiple Administrations	Alternate form
		Individual and group scores
		Classification consistency
		Generalizability
Testing System (Super-ordinate) Criteria	Form-Level Analyses	N
		Central Tendency (Mean, Median, Mode)
		Variation (Range, Variance, Standard Deviation)
		Standard Error of Measurement
		Bias
		IRT fit (TCC)
		Equating
		Scaling
	Reporting	Student level
		NCLB Subgroups
		Class
		District
		State
		Population
		Description of Standards Setting: Methods, Participants, Group Size
	Report Format	Basic
		Custom

In applying these technical criteria for their differentiated purposes, consumers and test developers need to keep in mind the following considerations, some of them identified through Rabinowitz and Sato's technical study:

- There is substantial but not total overlap between the procedures and criteria found appropriate and essential for reviews of ELL assessments and those for assessments for the general student population. Some criteria were found not to transfer directly or to be less critical in this type of technical review. For example, defining the referent group for bias/DIF studies was more difficult since the typical default group (white, English-speaking students) does not typically participate in ELL assessment technical studies. Thus, procedures widely used for technical reviews of assessments for non-ELL populations may need to be modified before they may be applied to more specialized populations.

- Technical reviews for ELL assessment should include both psychometrically trained reviewers and specialists in language acquisition. The study suggests there is overlap but incomplete agreement in the final reviews of these different types of experts. Ideally, content specialists would review ELL assessments to ensure sufficient content/construct validity before a more comprehensive technical review is carried out.

- Differential weighting or prioritization of review criteria may result in more efficient, valid technical reviews. Absent a priori knowledge of which technical criteria might be more integral to a technical review of ELL assessments, Rabinowitz and Sato's procedures made no attempt to provide different weights for different types of evidence. Future studies should focus on both potential differential weighting schemes and how such rules may evolve through the various stages of assessment development. For example, assessments in early years of development may focus primarily on construct and content validity. As these assessments become more widely used, reviewers may expect and, therefore, give added weight to consequential validity evidence.

- The large number of review criteria was cumbersome at times for reviewers. Future technical reviews should examine how to reorganize the review criteria to reflect or maintain the right level of granularity (i.e., Specific Evidence level) but also be more manageable. Some application rules may be more applicable for large comprehensive documents, such as technical manuals, than for a study looking more narrowly at a specific technical aspect of an assessment (e.g., bias study).

- Reviewing technical evidence is a necessary but insufficient screen for determining technical adequacy. Potential test consumers should always supplement these reviews by actually examining the test itself. This will ensure a full understanding of the quality of the items and of how the construct underlying the assessment is being operationalized.

- States and other consumers of ELL assessments need reviews of this type in order to proceed with accountability and education reform. Use of appropriate and technically defensible assessments is a key to reform in the NCLB era. States are undergoing peer review of their assessment systems that includes review of their assessments for ELL students. These peer review studies provide users with the following tools to decide how to select and use ELL assessments: examples of promising practices for establishing/presenting evidence for validity, reliability, bias; recommendations for priorities, timeline, strategies and guidelines for reporting results; and modes for integrating research and responding to federal regulations.

- Developers of ELL assessments need to step up their efforts to ensure technical adequacy and make all relevant technical evidence available for public and consumer review. When reviewing state ELL assessments, Rabinowitz and Sato found none of the instruments to be technically sufficient. Given the relatively short amount of time most ELL assessments have been under development and given the higher student and system stakes NCLB has required them to support, this finding is not surprising. But over time these assessments must become more technically sufficient. Assessment consumers must have adequate tools to support improvement of services to ELL students and other subpopulations at risk of poor academic achievement. For English learners in particular, the need will only grow as the percentage of ELL students continues to increase, not just in states such as California, Texas, New York, and Florida, but throughout the nation.

II. Examples of Promising Practices

In Rabinowitz and Sato's review, none of the tests developed to assess the language acquisition of ELL students met the full range of technical expectations for a high-stakes assessment used to measure both student achievement and school accountability. However, Rabinowitz and Sato did observe examples of technical evidence that satisfied the content and context, as well as the sufficiency, quality, and adequacy, criteria listed above. This section describes some of these promising practices so developers and consumers, respectively, can build

from these modest samples of success in developing and evaluating the next generation of these assessments.

VALIDITY

Content Validity

- While most assessment summaries described the need for demonstrating content validity, some provided evidence of expert review of items against state standards and test specifications. At least one assessment consumer has commissioned an independent alignment study based on rigorous criteria of alignment to both academic and English-language development standards.

- P-values and standard deviations for each item within each test form were presented. These values were accompanied by a discussion of the appropriateness of the assessment items for discriminating among Limited English Proficient, Non-English Proficient, Fluent English Proficient, and English Only students. Limitations of the field-test sample (e.g., participation of a significant number of students rated average to above average by their teachers) and their impact on the interpretation of the *p*-values also were presented.

Concurrent Validity

- Crosstabulation tables and Pearson correlation coefficients were presented to show the relationship between student performance on the assessment and teacher ratings of student academic ability and language ability (i.e., reading and writing).

- Correspondence between student scores on the assessment and other established, comparable language measures was presented. A narrative describing the degree of correspondence accompanied the data.

- Detailed description of the methodology (i.e., subjects, materials, procedures, design) and outcomes of concurrent validity studies were presented. Implications of the findings were described.

More typically, evidence of validity was limited to content validity and consisted of an assertion that content experts had reviewed the assessment items, or results of analyses (e.g., correlation coefficients) were presented with no or limited discussion of context or meaning. For the most part, little evidence was provided for the theoretical basis or theory of language acquisition supporting the assessment or other estimates of construct validity.

RELIABILITY

- Most tests in the review sample included both a description of reliability and some evidence of a reliability study, however rudimentary. Typically, both measures of internal consistency reliability (e.g., coefficient Alpha) and the standard error of measurement (SEM) were provided for each test form along with discussion of the interpretation of these values with respect to the reliability of the assessment scores.

Similar to evidence of validity typically found in the assessment documentation, reliability evidence often consisted of statements that the assessment was reliable or data were presented with insufficient description or discussion to fully judge the adequacy of the supporting reliability study.

BIAS

Bias is not addressed in most English learner assessment documentation beyond an assertion that the assessment is free of bias and sensitivity issues. The documentation of one assessment addressed bias in terms of gender bias, but it did not address other potential sources of bias that would seem to be more relevant to the target population (e.g., linguistic, cultural).

In seeking evidence to establish and defend the technical adequacy of the English learner assessments, test consumers must distinguish between mere claims and substantiated research findings. "Evidence" presented in technical documentation can range from assertions of findings (without support) to research summaries (without any evidence) up to detailed descriptions of formal research-based technical evidence. It is important for consumers to consider the nature and source of the evidence when evaluating its technical adequacy.

III. A Framework for Integrating Technical Evidence in Response to Federal Requirements

Use of appropriate and technically defensible assessments is a key to reform in the NCLB era. Assessments for special student populations (e.g., alternate assessments for students with disabilities, Title III assessments) have, to date, been receiving the most negative feedback in initial NCLB-required peer reviews. This is not surprising since the assessment of special student populations is a relatively new area of work within the parameters of the legislation. To satisfy NCLB's goals, assessments for special student populations must have evidence of validity, reliability, and freedom from bias so that states using these assessments can fully and equitably include these students in their testing program.

These guidelines, including the review criteria in section II, can help states prepare for the peer review of their assessment systems, which includes review of their assessments for students who are English learners. Both sets of criteria (i.e., the technical criteria presented in these guidelines and the peer review criteria) are based on well-established

practice in the areas of assessment and accountability; therefore, it is not surprising that there would be substantial overlap between the two. As shown in Appendix B on page 16, the criteria in these guidelines overlap with the peer review criteria, with the exception of the following indicators dealing primarily with test administration protocols and procedures:

- Test security/protection of student confidentiality/ethical considerations
- Test administration protocol/training of test administrators
- Information about cognitive complexity of items (DOK, process dimensions)
- Assessment feedback loop/continuous improvement/self-evaluation stipulations
- Degree to which test is dynamic/can accommodate ongoing changes over time in core values, instructional practices
- Test inclusiveness/levels of participation

These guidelines are intended to inform the next generation of assessments for special student populations. Thus, the next iteration will consider inclusion of the criteria listed above in order to increase alignment with the expectations of the peer review and to expand the basis for supporting an assessment's validity, reliability, and freedom from bias. ■

References

American Educational Research Council, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing (3rd ed.). Washington, DC: Author.*

Becker, D., & Camilli, G. (2004). Standardizing the standards: In search of uniform guidelines for state technical reports. National Council on Measurement in Education, 12, 6–8.

Bowe, F. G. (2000). *Universal design in education: Teaching non-traditional students. Westport, CT: Bergen & Garvey.*

Center for Equity and Excellence in Education. (2005). Recommendations for State ELLs Accommodation Policies. Retrieved July 10, 2005, from http://ceee.gwu.edu/AA/Accommodations_Recos.html.

Heubert, J., & Hauser, R. (Eds.). (1999). *High stakes: Testing for tracking, promotion, and graduation. Washington, DC: National Research Council, National Academy Press.*

Kubiszyn, T., & Borich, G. (2003). *Educational testing and measurement: Classroom application and practice. (7th ed.). New York: John Wiley and Sons.*

Rabinowitz, S., Ananda, S., & Bell, A. (2004, December 7). Strategies to assess the core academic knowledge of English-language learners. Journal of Applied Testing Technology. Retrieved July 12, 2005, from http://www.testpublishers.org/Documents/JATTdec2004_ELL_strategies_pape.pdf.

Rabinowitz, S., & Sato, E. (n.d.). *A technical review of high-stakes assessments for English language learners. San Francisco, CA: WestEd.*

Rochester Institute of Technology. (2004). *Class act: Access for deaf and hard-of-hearing students. Retrieved July 12, 2005, from http://www.rit.edu/~classact/side/universaldesign.html.*

What Works Clearinghouse. (2004a). *WWC study review standards. Retrieved July 18, 2005, from http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf.*

What Works Clearinghouse. (2004b). WWC evidence standards. Retrieved July 18, 2005, from http://www.whatworks.ed.gov/reviewprocess/standards.html.

Zehler, A., Hopstock, P.J., Fleischman, H.L., & Greniuk, C. (1994). *An Examination of Assessment of Limited English Proficient Students (Task Order Report. D070). Special Issues Analysis Center.*

WestEd, a nonprofit research, development, and service agency, works with education and other communities to promote excellence, achieve equity, and improve learning for children, youth, and adults. While WestEd serves the states of Arizona, California, Nevada, and Utah as one of the nation's Regional Educational Laboratories, our agency's work extends throughout the United States and abroad. It has 15 offices nationwide, from Washington and Boston to Arizona, Southern California, and its headquarters in San Francisco. For more information about WestEd, visit our website: WestEd.org; call 415.565.3000 or, toll-free, (877) 4-WestEd; or write: WestEd / 730 Harrison Street / San Francisco, CA 94107-1242.

This paper was produced in whole or in part with funds from the Institute of Education Sciences, U.S. Department of Education, under contract #ED-01-CO-0012. Its contents do not necessarily reflect the views or policies of the Department of Education.

Appendix A: Technical Evidence Worksheet

This worksheet guides assessment consumers through an evaluation of the context and content of an assessment intended for use with special populations (Section I) and of the technical evidence presented to establish the assessment's validity, reliability, and freedom from bias (Sections II and III). Assessment producers can use this worksheet to verify that such information is presented in their assessment's documentation. Although the worksheet is focused on assessments for use with students who are English language learners (ELL), it can be easily adapted for use with assessments targeting other special populations.

Test Name: _____
Publisher: _____
Year of Publication: _____

Section I

Instructions: First articulate your intended purpose, student population, and assessed content needs. Then, review the assessment's documentation (e.g., technical report, manuals) and determine whether this assessment is appropriate for your intents and needs.

Purpose: We need an assessment for the following purpose:

- ☐ Identification
- ☐ Placement
- ☐ Placement Review
- ☐ Exit
- ☐ Other: _____

This assessment is appropriate for our intended use.

☐ YES ☐ NO ☐ UNSURE (Reason): _____
_____

Target Population: The students we will test can be characterized as follows:
ELL Subgroup Characteristics

- ☐ _____
- ☐ _____
- ☐ _____
- ☐ _____
- ☐ _____

This assessment is appropriate for the population of ELL students we will test.

☐ YES ☐ NO ☐ UNSURE (Reason): _____
_____

Content/Constructs Assessed: Given the purpose of the test (as stated above), we expect students to be able to demonstrate what they know and can do with regard to the following:

Target Content/Constructs

☐_____

☐_____

☐_____

☐_____

☐_____

This assessment tests the critical skills and concepts we believe ought to be assessed in order to satisfy the purpose stated above.

☐ YES ☐ NO ☐ UNSURE (Reason): _____

_____

Section II

Instructions: Review the assessment's documentation and evaluate (i.e., 0-5, as defined in the column "Evaluation") the presence and presentation of the technical evidence listed below. To the degree possible, evaluate the assessment's "Specific Evidence."

Criteria Cluster	Criterion	Specific Evidence[1]	Evaluation 0 = Unsure/unclear 1 = No information presented 2 = Evidence addressed in an assertion 3 = Evidence presented in a summary without data 4 = Evidence presented in a summary with data 5 = Evidence presented in a detailed description with data	Notes (e.g., if Specific Evidence is not available, you may choose to note the level of detail at which evidence is available – Criteria Cluster, Criterion; questions or concerns; reference to other documentation that may further address a piece of technical evidence)
Validity	Field Testing	Field Test Sampling Design: Representativeness and Norming		
		Field Test Sampling Design: Currency (at least, dates documented)		
		Field Test Sampling Design: Randomization		
		Fidelity (link of test to stated purpose of the test)		
	Design	Attrition of Persons (for Pre/Post Designs)		
		Test Blueprint		
		Scoring Rubric for OE Items: Construction and Validation		
		Accommodations		
	Content	Content Alignment Studies		
		Expert judgments		
		p-values		
		Discrimination (Item-test Correlations)		
		Bias/DIF analysis		
		IRT/Item fit (ICC)		
		Distractor Analysis		
	Construct	Factorial Validity (structural equation modeling)		
		Multi-Trait/Multi-Method		
		Equivalence/Comparability (construct the same regardless of examinee's ability)		
	Criterion	Predictive validity - Validation to the Referent		
		Predictive validity - Individual and group scores		
		Concurrent validity - Validation to External Criteria		
		Concurrent validity - Validity of External Criteria		
		Concurrent validity - Individual and group scores		
	Consequential	Evaluation of Testing Consequences		
		Individual and group scores		

[1] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

Criteria Cluster	Criterion	Specific Evidence[1]	Evaluation 0 = Unsure/unclear 1 = No information presented 2 = Evidence addressed in an assertion 3 = Evidence presented in a summary without data 4 = Evidence presented in a summary with data 5 = Evidence presented in a detailed description with data	Notes (e.g., if Specific Evidence is not available, you may choose to note the level of detail at which evidence is available – Criteria Cluster, Criterion; questions or concerns; reference to other documentation that may further address a piece of technical evidence)
Validity, continued	Freedom from Bias	Content		
		Ethnicity		
		Cultural		
		Linguistic		
		Socio-economic		
		Geographic		
		Students with disabilities		
		Universal Design		
Reliability	Reliability: Single Administration	Scale		
		Internal Consistency		
		Split-half		
		Scorer/Hand-scoring		
	Reliability: Multiple Administrations	Test-retest		
	Reliability: Either Single or Multiple Administrations	Alternate form		
		Individual and group scores		
		Classification consistency		
		Generalizability		
Testing System (Superordinate) Criteria	Form-Level Analyses	N		
		Central Tendency (Mean, Median, Mode)		
		Variation (Range, Variance, Standard Deviation)		
		Standard Error of Measurement		
		Bias		
		IRT fit (TCC)		
		Equating		
		Scaling		
	Reporting	Student level		
		NCLB Subgroups		
		Class		
		District		
		State		
		Population		
		Description of Standards Setting: Methods, Participants, Group Size		
	Report Format	Basic		
		Custom		

[1] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

Section III

Instructions: Once you've completed your evaluation of the assessment's technical evidence, consider the following:

1. The information related to the assessment's technical evidence consists mostly of:
   - ☐ Assertions
   - ☐ Summaries
   - ☐ Detailed descriptions

2. The evidence and related information provided in the assessment's documentation is (check all that apply):
   - ☐ Comprehensive (e.g., quantitative information is accompanied by supporting text and interpretations)
   - ☐ Accurate and directly supports the evidence being evaluated
   - ☐ Generalizable or broadly applicable
   - ☐ Credible

3. The data presented to support the technical evidence discussed in the assessment's documentation are mostly:
   - ☐ Quantitative
   - ☐ Qualitative
   - ☐ Both quantitative and qualitative
   - ☐ There are no data presented

4. There is enough evidence to start using the test.
   - ☐ Yes
   - ☐ Yes with reservations (reason): _____
   - ☐ No

5. If there is currently insufficient evidence, there is a plan to gather the evidence needed.
   - ☐ Yes (explanation): _____
   - ☐ No

Appendix B: Overlap Between Technical Criteria and Peer Review Criteria

Technical Criteria			Peer Review Criteria																
			3.4	3.7	4.1	4.2	4.3, 6.1	3.3, 4.4	4.5	4.5	4.5, 5.6	4.6	5.1 - 5.4	5.5 - 5.7	6.2 - 6.4	7.1	7.2	7.3	7.4	7.5
			Comparability Across Grades	Alternate Assess. for SD	Evidence of Validity	Evidence of Reliability	Evidence of Fairness/ Accessibility	Forms Equivalence	Test Administration Protocol	Scoring Protocol	Reporting Protocol	Accommodations	Alignment with Content Standards	Alignment with Achiev./Perf. Levels	Participation of Special Populations	Valid Interpretation of Results from Reports	Required Subgroups in Reporting System	Student Level Reports	Protection of Confidentiality	Itemized Score Analyses Reported
Validity	Field Test	Design	X		X		X	X				X	X	X						
		Forms Equating				X		X					X	X						
		Norming			X		X								X	X				
	Design	Construct Comparability	X		X	X		X	X	X	X	X	X	X	X	X	X	X		X
		Blueprint	X		X		X	X					X	X		X				
		OE Scoring Rubric	X		X	X	X			X			X	X		X				X
		Accommodations	X		X	X	X		X		X	X		X	X	X		X	X	X
	Content	Alignment Studies	X		X			X		X	X		X	X						
	Construct	SEM/MTMM	X		X		X								X					
	Criterion	Predictive			X											X				
		Concurrent			X											X				
	Consequential	Eval Testing Consequences	X	X	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X
	Freedom from Bias	Content		X	X		X	X				X	X							
		Ethnicity			X		X			X					X	X			X	
		Gender			X		X			X						X			X	
		Linguistic			X		X		X	X	X	X			X	X			X	
		Socio-economic			X		X			X					X	X			X	
		Geographic			X		X			X					X	X			X	
		Disabilities		X	X		X		X	X	X	X			X	X			X	
		Universal Design	X	X	X	X	X		X	X	X				X	X	X	X		
		DIF/Other Review			X	X	X								X	X				X

Technical Criteria			Peer Review Criteria																	
			3.4	3.7	4.1	4.2	4.3, 6.1	3.3, 4.4	4.5	4.5	4.5, 5.6	4.6	5.1 - 5.4	5.5 - 5.7	6.2 - 6.4	7.1	7.2	7.3	7.4	7.5
			Comparability Across Grades	Alternate Assess. for SD	Evidence of Validity	Evidence of Reliability	Evidence of Fairness/ Accessibility	Forms Equivalence	Test Administration Protocol	Scoring Protocol	Reporting Protocol	Accommodations	Alignment with Content Standards	Alignment with Achiev./Perf. Levels	Participation of Special Populations	Valid Interpretation of Results from Reports	Required Subgroups in Reporting System	Student Level Reports	Protection of Confidentiality	Itemized Score Analyses Reported
Reliability	Single Admin.	Scale	X			X	X			X		X		X	X	X		X		X
		Internal Consistency				X		X					X			X				
		Split-Half				X		X		X						X				X
		Interrater				X				X						X				
	Multiple Admin.	Test-Retest				X				X						X				
		Alternate Form				X		X		X						X				
		Classification Consistency	X		X	X	X	X		X		X		X	X	X		X		
		Generalizability	X		X	X	X	X	X			X			X	X				
Testing Syste Adequacy	Item Level	P-values			X		X	X					X							X
		Discrimination			X		X	X					X							X
		Item Fit/ICCs						X					X							X
		Distracters			X		X	X		X			X							X
	Form Level	Ns	X			X									X	X	X	X		
		Descriptive Statistics	X			X		X							X	X	X	X		X
		SEMs				X				X						X		X		
		Test Fit/TCCs	X	X	X	X	X	X					X	X	X	X				
		Equating			X		X	X					X	X		X				X
		Scaling	X			X	X			X		X		X		X				
		Standards Setting	X		X	X	X				X			X		X				
	Reporting Level	Student		X	X	X	X	X			X	X				X		X	X	X
		NCLB Groups		X	X	X	X				X	X		X	X	X	X		X	X
		Class/Grade									X					X	X		X	X
		District	X								X	X				X	X		X	X
		State	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X
		Population	X								X	X			X	X			X	X
	Report Format	Basic									X					X	X	X	X	X
		Custom	X	X							X	X			X	X	X	X	X	X