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

Evaluation of the Technical Evidence of Assessments for Special Student Populations

Assessment and Accountability Comprehensive Center
Special Populations Strand
August 2007

Overview and Background

Since the enactment of the No Child Left Behind Act of 2001 (NCLB), and particularly in relation to Title I and Title III, assessments for special student populations are undergoing increased scrutiny. Ensuring the technical adequacy of assessments for English language learners (ELLs) and students with disabilities (SWDs) is critical, given the high stakes associated with the outcomes of these measures. State departments of education, policymakers, and test developers have attempted several strategies to satisfy the NCLB requirements for valid, reliable, and accessible assessments for special student populations. In practice, however, several challenges affecting technical adequacy have impacted states' abilities to meet the needs of their special student populations (Rabinowitz, Ananda, & Bell, 2005). These challenges include demonstrating technical adequacy and comparability of assessments for special student populations and their general education counterparts and ensuring consistency of meaning across assessments for general population students and students with special needs (Abedi, 2004; Bowe, 2000; Center for Equity and Excellence in Education, 2005; Rochester, 2004).

Several challenges affecting technical adequacy have impacted states' abilities to meet the needs of their special student populations.

Recent research has shown that the technical adequacy of assessments for special student populations is relatively undeveloped compared to their general education counterparts; that is, the technical evidence provided and the methods by which this evidence is established do not necessarily account for the unique characteristics of special needs populations (i.e., ELLs, SWDs) or the assessed domains (e.g., English language proficiency) (Rabinowitz & Sato, 2005). While there is substantial overlap between the procedures and criteria found appropriate and essential for determining the technical adequacy of special population assessments versus their general education counterparts, there is not complete overlap. Some technical criteria do not transfer directly or are less critical when applied to a technical review of assessments for special student populations (Rabinowitz & Sato, 2005).

Technical criteria and the methods by which these criteria are applied must account for the unique characteristics of special student populations.

This document presents an ongoing evaluation conducted by the Special Populations Strand of the Assessment

WestEd
excellence in research, development, & service

and Accountability Comprehensive Center (AACC) and is intended to inform developers and consumers of assessments for special student populations (ELLs and SWDs). The evaluation focuses on the technical adequacy (i.e., validity, reliability, freedom from bias) of evidence related to assessments used to meet relevant Title I and Title III requirements under NCLB. The technical criteria used in this evaluation are validated and are sensitive to the unique characteristics of special student populations, the particular purposes of the assessments, and the stage of development and maturity of the assessments.[1]

The technical criteria used in this evaluation are validated and are sensitive to the unique characteristics of special student populations, the particular purposes of the assessments, and the stage of development and maturity of the assessments.

According to the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999) there are multiple elements that contribute to the technical adequacy of high-quality assessments. Until recently, the range of sources of technical evidence available had not been aggregated and rigorously evaluated in any methodical fashion. Therefore, the AACC is applying a comprehensive set of validated criteria based on those developed by Rabinowitz and Sato (2005) to evaluate the technical evidence associated with assessments for special student populations (see Appendix A for a list of the technical criteria and Appendix B for the operational definitions of the criteria). As mentioned previously, these criteria are sensitive to the unique characteristics of special student populations, the particular purposes of the assessments, and the stage of development and maturity of the assessments. Using the results of this evaluation, assessment developers and consumers will be able to gauge the technical adequacy of the assessments they are using or are considering for use with their special student populations and identify appropriate and necessary next steps for ensuring the assessments' validity and, ultimately, the defensibility of the assessment and its results.

As mentioned previously, this evaluation builds off of research conducted by Rabinowitz and Sato (2005) that examined the technical adequacy of evidence associated with ELL assessment types (i.e., consortium-developed, custom-developed [publisher], and not custom-developed [publisher]). The AACC evaluations presented here are reviews of technical evidence related to specific ELL assessments (rather than assessment types) in order to better inform states about the technical adequacy (i.e., validity, reliability, freedom from bias) of the assessments available for use with their ELLs (technical reviews of assessments for SWDs will be made available as this evaluation work continues). As suggested by the research conducted by Rabinowitz and Sato (2005), many assessments for special student populations currently used for accountability

Many assessments for special student populations currently used for accountability purposes should be considered works in progress.

[1] The first set of assessments evaluated vis-à-vis the technical criteria is English language proficiency assessments for ELLs. For evaluation summaries of the assessments reviewed, please go to http://www.aacompcenter.org and click on the Special Populations link. Assessments for SWDs will be reviewed subsequently.

purposes should be considered works in progress because they have been developed since the advent of NCLB, making current technical evidence preliminary, at best.

The Evaluation of Technical Evidence: Materials and Procedures

Materials

As mentioned previously, the first set of assessments evaluated was for ELLs. AACC staff compiled a state-by-state listing of English language proficiency (ELP) assessments used for ELL[2] students and collected available technical documentation related to specific assessments. Generally, the ELP assessments fell into three main categories: consortium-developed; custom-developed (publisher); and not custom-developed (publisher). Attention was focused on those more formal assessment efforts with sufficient resources behind them to likely meet technical adequacy requirements.

Available technical documentation published by the test developers was collected for each assessment (e.g., technical manuals and reports). Additional documents were identified for review by using the following resources:

- Experts in the areas of assessment, accountability, and ELLs;
- State-level contacts;
- Journal articles, technical bulletins, and reports;
- Conference presentations;
- Manuals and other documentation from state and national assessment programs; and
- Reliability and validity studies from relevant assessment programs.

Criteria

Technical evidence was evaluated against a comprehensive set of criteria based on widely known and respected standards (e.g., What Works Clearinghouse, 2004a, 2004b; AERA, APA, & NCME Joint Standards, 1999; Becker & Camilli, 2004) as well as principles underlying rigorous, scientifically-based research (see Appendix A for a list of the technical criteria and Appendix B for the operational definitions of the criteria). Generally, the criteria are related to assessment validity, reliability, and freedom from bias. A group of experts that collectively had experience in large-scale test development, psychometrics, English language development, the English language learner population, and technical assistance/consultation to state departments of education were convened to review and validate the criteria and their operational definitions.[3]

The evaluation necessarily involved both *technical* and *content* perspectives.

[2] For the purpose of considering materials for inclusion in this study, the category of English language learner (ELL) was defined in its broadest sense because part of this analysis involved the evaluation of the adequacy of the target population definition and subsequently implications for the technical adequacy of evidence presented, as it relates to this population (e.g., sampling and bias).

[3] The technical criteria and operational definitions developed by Rabinowitz and Sato (2005) were originally developed for research purposes. Therefore, with the consent of the authors, the AACC has made refinements to both the technical criteria and operational definitions in order to make this information more practicable to states and test publishers.

Use of appropriate and technically defensible assessments is a key to reform in the NCLB era.

Analysts

The evaluation necessarily involved both *technical* and *content* perspectives. Analysts possessed experience in large-scale test development, and collectively, they possessed training in measurement, statistics, and applied linguistics and had English language development expertise as well as experience teaching ELLs. The expertise of the analysts was critical because a technical understanding of evidence related to validity, reliability, and bias was not sufficient to determine whether the evidence was fully appropriate for the ELL population and purposes of the assessments (i.e., the target population is much more narrowly defined than that of a typical statewide achievement test, and the ELP domain is necessarily assessed in a manner that differs from academic content domains). For example, if a section within a document described Differential Item Functioning (DIF), then from a *technical* perspective (measurement, statistics) the analyst would determine whether the DIF analyses were appropriately implemented (e.g., sufficient sample sizes, appropriate significance tests). From a *content* perspective (applied linguistics, ELD), the analyst would determine whether the application of DIF from traditional testing programs was fully appropriate for the ELL population and purposes of the assessments that are the focus of this technical evaluation.

Analysts were provided training on the technical criteria and evaluation protocol. As necessary, analysts created *decision rules*, which are guidelines for the application of the criteria that help to ensure accurate and consistent application throughout the evaluation process. Training ended when analysts could ensure consistent and accurate understanding and application of the criteria.

Procedure

Following their training on the technical criteria and evaluation protocol, analysts evaluated the available documentation for selected ELP assessments. A preliminary summary describing the technical evidence associated with an assessment vis-à-vis the technical criteria was then written. This summary was sent to the test developer in order to solicit any comments and additional technical evidence the developer might have that would further inform the evaluation of the assessment's technical quality.

If a test developer submitted additional documentation for review, analysts would evaluate the documentation for consideration in the assessment's body of evidence for validity, reliability, and freedom from bias. If a test developer chose not to submit additional documentation or did not provide comment on the preliminary summary, then the test developer's response (or lack thereof) was noted in the body of evidence summary for the assessment.

Test developers were provided the opportunity to review and comment on the body of evidence lists and summaries resulting from AACC evaluations. Once a test developer had the opportunity to provide additional documentation for consideration and to comment on the body of evidence summary created for its assessment, the final summary was prepared for posting on the AACC website.

As noted previously, this evaluation is ongoing; therefore, additional technical evaluations of assessments will be conducted and summaries will be posted as they are completed. Currently, the AACC evaluations have focused on ELP assessments for ELLs. As the work continues, technical evaluations will include assessments for SWDs.

States and other consumers of assessments for special student populations need reviews of this type to proceed with accountability and educational reform. Use of appropriate and technically defensible assessments is a key to reform in the NCLB era. Developers of assessments for special student populations are encouraged to step up their efforts to ensure the technical adequacy of their assessments and to make all relevant technical evidence available for consumer review, particularly given the higher student and system stakes NCLB requires. States must have adequate tools to support improvement of services to their special student populations. This need will only grow as the percentage of students who are English language learners and students with disabilities continues to increase throughout the nation. ❖

This need will only grow as the percentage of students who are English language learners and students with disabilities continues to increase throughout the nation.

References

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

AERA, APA, & NCME. (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*. NCME Newsletter.

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 the World Wide Web: http://ceee.gwu.edu/AA/Accommodations_Recos.html.

Rabinowitz, S., Ananda, S., & Bell, A. (2005). Strategies to assess the core academic knowledge of English-language learners. *Journal of Applied Testing Technology*.

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, CA: WestEd.

Rochester Institute of Technology. (2004). *Class act: Access for deaf and hard-of-hearing students*. Online source: http://www.rit.edu/~classact/side/universaldesign.html.

What Works Clearinghouse. (2004a). *WWC study review standards*. Online publication: http://www.whatworks.ed.gov/reviewprocess/study_standards_final.pdf.

What Works Clearinghouse. (2004b). *WWC evidence standards*. Online publication: http://www.whatworks.ed.gov/reviewprocess/standards.html.

Evaluation summaries of assessments reviewed can be found at http://www.aacompcenter.org (see Special Populations page).

For questions about these technical criteria, contact:
Edynn Sato, Ph.D.
Director, Special Populations
Assessment and Accountability Comprehensive Center, WestEd
email: esato@wested.org.

For information about the AACC, visit http://www.aacompcenter.org.

Appendix A

Assessments for English Language Learners: Technical Adequacy Criteria—Tiers

Notes: Types of validity, reliability, and bias and sensitivity evidence associated with various phases of test development are presented in the table below. Tier 1 elements ought to be part of a test's body of evidence. Tier 2 elements are important, but may or may not be available, depending on the nature and maturity of a particular test.

Type	Phase	Tier 1 Elements	Tier 2 Elements
Validity			
Construct validity	Test design/development (10)	Test purpose	Universal design
		Population/classification	Readability
		Theoretical foundation/framework	Multi-trait/multi-method/subtest inter-correlation
		Standardization	Equivalence/comparability
			Fidelity
			Accommodation
Content validity	Test design/development (14)	Alignment (items-to-standards)	Linkage
		Expert judgment	Structural equation modeling
		Item fit IRT/item fit **OR** *p*-values/point biserials	*t*-tests
		Test blueprint	ANOVA
		Alignment (test form-to-blueprint)	Factor analysis
		Test fit IRT/test fit **OR** Descriptive statistics	Linking/equating*
	Field testing (3)	Sampling **OR** Norming	Blueprint
	Scoring (4)	Scale	Rubric
		Standard setting*	Training of scorers/scoring protocol
Criterion validity	Test design/ development (2)	Cross tabulations **OR** Pearson correlation	
Consequential validity	Test design/ development (1)	Use of results	
	Reporting (4)	N	Effect size
		Central tendency/ variation	
		Reporting category	
	Security (1)	Protocols	

*Placement in this column is related to the nature and maturity of the instrument.

Type	Phase	Tier 1 Elements	Tier 2 Elements
Reliability			
Reliability	Test design and development (11+1)	*Internal consistency* KR-21 **OR** Coefficient alpha **OR** Split-half	Test length/power estimates
		Standard error of measurement/ confidence intervals	*Generalizability* *G*-coefficient
		Test-retest **OR** Alternate form*	*Classification consistency* Percent correspondence **OR** Correlation coefficient **OR** Classification error
	Scoring (2)	*Inter-rater reliability* Percent correspondence **OR** Correlation (kappa)	
Bias and Sensitivity			
Bias and sensitivity	Test design/development (13)	*Expert review:* Linguistic Ethnicity/race Cultural/religious Geographic SES Disability Gender	*DIF analyses:* Linguistic Ethnicity/race Geographic SES Disability Gender

*Placement in this column is related to the nature and maturity of the instrument.

Appendix B

Assessments for English Language Learners: Technical Adequacy Criteria—Operational Definitions

		Type Element: Evidence/Method	Operational Definition	Notes
Test design and development				
	Validity		Validity (content, construct, and consequential) is the degree to which an assessment measures what it intends to measure, given the specific purpose(s)/use(s) of the assessment, the specified content, and the circumstances related to the assessment (context, population, etc.) (Messick, 1981).	
		Item/Test Level		
	Validity	Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test (based on Joint Standards, 1999).	
1	Validity	Test purpose	The test purpose is the reason or object for which an assessment is designed, developed, and intended to be used.	clearly stated purpose related to range of appropriate purposes for current ELD testing (e.g., placement, classification, redesignation)
2	Validity	Population/classification	The set of examinees for whom the test is intended for the purpose(s) stated.	clearly defined population: non-native English speakers; geographical location
3	Validity	Theoretical foundation/framework	The theoretical foundation is the underlying framework, model, or perspective that defines the domain being measured and how best to measure it.	clearly stated, coherent, current/accepted theories
4	Validity	Universal design	Universal design is a process of incorporating considerations and features into an instrument to promote its accessibility and validity for the widest range of examinees, including examinees with disabilities and examinees with limited English proficiency.	evidence of application of UD principles
5	Validity	Readability	Readability is the measure of the complexity of the language in the text and directions.	expert judgment; documentation; number; statement that text is grade-appropriate and appropriate for the population and purpose; protocol; readability formulae (e.g., Lexile, Dale-Chall, etc.)
		Item Level		
	Validity	Content	Content is the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test. Content validity indicates the degree to which the items measure the content (i.e., knowledge/skills/abilities).	

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
6	Validity	Alignment (items-to-standards)	Alignment refers to the degree to which content (e.g. skills, concepts) in an assessment and set of standards match in breadth, depth, and range of complexity. Alignment relationships tend to be direct relationships.	alignment at key points throughout the development process (internal, external) evaluated; explanation of process or results (including limitations); Internal: alignment may be done by writers, editors, or other developers and expert reviewers during the item development process. External: alignment should be done by independent experts in assessment, standards, the special student population, and relevant content areas. Alignment procedures and studies should look for appropriateness of item content and cognitive level as described in individual standards, and coverage (breadth and depth) as reflected by the set of standards.
7	Validity	Linkage (items-to-standards; standards-to-standards)	Linkage refers to relationships that tend to be developmental, foundational, or proximal (as opposed to direct), and are typically observed between standards and assessments or two sets of standards developed for different populations (e.g., general education standards and alternate standards).	linkage at key points throughout the development process (internal, external); appropriate criteria and model used; explanation of process or results (including limitations); Internal: linkage may be done by writers, editors, or other developers and expert reviewers during the item development process. External: linkage should be done by independent experts in assessment, standards, the special student population, and relevant content areas. Linkage procedures and studies should look for appropriate developmental, foundational, or proximal relationships of item content and cognitive level as described in individual standards, and coverage (breadth and depth) as reflected by the set of standards.
8	Validity	Expert judgment	Expert judgment of content validity is the use of individuals with relevant knowledge and background for verifying the degree to which the test's questions are representative of the content that the test questions are intended to assess.	credible experts; methodology/protocol described; explanation of findings; distracter analysis
9	Validity	*p*-values/point biserials	*P*-values are the probabilities of correctly answering items. Point biserials are correlations between the total test score and individual item scores.	(high *p*-value reflects an "easy" item; looking for a range of difficulty appropriate to test purpose) discussion of how *p*-values relate to the items' ability to discriminate among the target (sub)groups of examinees

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
10	Validity	IRT/item fit	IRT/item fit relates the probability of a correct response to an examinee's ability level on the construct (latent trait).	description of model; explanation of results; Item Characteristic Curve (ICC) one, two, or three parameter IRT are okay
11	Validity	Structural equation modeling	Structural equation modeling shows the relationship between the construct and the measurable factors that affect it and traces the relationships within a network of variables.	report on the relative contribution of each factor examined; support/verification of predictions of the relationship of the construct to the measurable factors
12	Validity	*t*-tests	*T*-tests are statistical hypothesis tests that examine the equality of the means of two variables or two groups on the same variable (Fraenkel & Wallen, 2000).	value for *t* statistic and its significance level; explanation of results
13	Validity	ANOVA	ANOVA is a statistical procedure that examines the equality of differences between the means of more than two groups and the interaction among effects (Fraenkel & Wallen, 2000).	value for *f* statistic and its significance level; explanation of results
14	Validity	Factor analysis	Factor analysis is a statistical technique to determine if multiple variables can be described by a few factors (unidimensionality) (based on Fraenkel & Wallen, 2000).	correlations (factor loadings); explanation of results
		Test Level		
	Validity	Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test (based on Joint Standards, 1999).	
15	Validity	Equivalence/comparability	Equivalence/comparability means that two or more tests/test forms measure the same construct and/or are interchangeable.	description of method of analysis (typically involves expert judgment; unit of analysis reflects the entire construct); subtest intercorrelations
16	Validity	Multi-trait/multi-method/ subtest inter-correlation	Multi-trait/multi-method matrices display evidence of the relationships/factors (convergence or divergence) related to examinee performance that can be compared so that the validity of the assessment can be determined/evaluated. Subtest inter-correlation is evidence that the pieces of the test are measuring the same construct (e.g., subtests within the reading section). Note: Subtest inter-correlation may appear as evidence of internal consistency. However, we believe that there is other stronger evidence for internal consistency. Therefore, our recommendation is that it be presented as evidence of construct validity.	correlation table or MTMM matrix

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
	Validity	Content	Content is the set of behaviors, knowledge, skills, abilities, attitudes, or other characteristics to be measured by a test. Content validity indicates the degree to which the items measure the content (i.e., knowledge/skills/abilities).	
17	Validity	Test blueprint	The test blueprint communicates the structure and contents of a test, including the relative weighting or distribution of strands of content.	table or chart showing the content distribution, item type, etc.
18	Validity	Alignment (test form-to-blueprint)	Alignment (test form-to-blueprint) is the degree to which the test form reflects the intended breadth, depth, and emphasis of content specified in the test blueprint.	in-process alignment and/or ex post facto alignment studies done (independent); appropriate unit(s) of analysis and model/ appropriate dimensions evaluated; explanation of process or results (including limitations)
19	Validity	Descriptive statistics (e.g., central tendency, variation)	Descriptive statistics are summary measures of a distribution of scores, providing information about location and variability.	mean; standard deviation; *N*; explanation of results (e.g., evidence that field test results were used to select appropriate items)
20	Validity	IRT/test fit	IRT/test fit relates the proportion of correct responses to an examinee's ability level on the construct (latent trait). One, two, or three parameter IRT are okay.	description of model; explanation of results; Test Characteristic Curve (TCC)
21	Validity	Linking/equating	Linking is putting two or more tests on a common scale to show that the scores can be compared. If the two tests are essentially parallel, the process is termed equating, a special case of linking.	report of linking/equating error; description of linking/equating methods (including assumptions, feasibility); reference to dimensionality; factor analysis; correlations; DIF; structural equation modeling
	Validity	Criterion (predictive/concurrent)	Criterion validity is the extent of the relationship of a test score to an external criterion. The extent to which a score can predict the value of a criterion measure is predictive validity (McDonald, 1999). Concurrent validity compares scores of two measures collected at about the same time (Fraenkel & Wallen, 2000).	must include explanation of how data are used
22	Validity	Cross tabulations	Cross tabulations are tabular representations of the relationships (categorical or continuous) among two or more different measures.	description of relationships; explanation of results; description of measures; includes expectancy tables
23	Validity	Pearson correlation	Pearson correlation is the number between -1 and 1 that indicates the degree to which two quantitative variables are related (shows strength and direction of relationship).	correlation coefficient; description of measures; explanation of results
	Validity	Consequential	Consequential validity is the degree to which results are used in a manner consistent with the *intended* purpose and uses of the assessment.	

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
24	Validity	Use of results	Use of results refers to the intended and unintended ways in which test scores are analyzed, reported, and/or brought into service to inform and facilitate decision-making (i.e., diagnosis, evaluation, classification, selection, promotion, placement, entry/exit).	fidelity between stated purpose of assessment and how results are reported/guidelines for use of results—look at stated purpose of the assessment along with, for example, sample reports, scoring outcomes/results; includes item release strategy
		Administration		
	Validity	Construct	A construct is the concept or the characteristic that a test is designed to measure. Construct validity indicates that the test scores reflect the examinee's standing on the psychological construct measured by the test (based on Joint Standards, 1999).	test administration (e.g., accommodations provided, fidelity to standard protocol) does not alter the construct being tested—for example, reading aloud the reading comprehension section of the assessment alters the construct
25	Validity	Accommodation	Accommodations are changes made to the test itself or its administration procedures in order to accommodate examinees who require such changes in order to be able to show what they know and can do. In theory, changes do not alter the construct, and are intended to minimize the influence of construct-irrelevant factors.	allowed accommodations do not alter the construct accommodations do not affect reliability or validity
26	Validity	Fidelity	Fidelity is the degree to which the protocol for standardized test administration is followed.	Test administration conditions/procedures do not alter the construct.
27	Validity	Standardization	Standardization means having rules and specifications for testing procedures that are intended to ensure testing conditions are the same for all examinees (Joint Standards, 1999).	level of detail and degree to which they ensure standardized testing conditions
	Reliability		Reliability is the degree to which test scores for a group of examinees are stable and consistent over repeated applications of a test and are therefore inferred to be dependable and repeatable for an individual examinee. It is also the degree to which scores are free of errors of measurement for a given group (Joint Standards, 1999).	
		Item/Test level		
28	Reliability	Stability & consistency	Stability is the extent to which scores on a test are essentially invariant over time. Consistency is the extent to which multiple forms of a test measure a construct consistently.	

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
29	Reliability	SEM/confidence intervals	Standard error of measurement indicates the dispersion of measurement errors when estimating examinees' true scores from their observed test scores. Confidence intervals are bands defining score zones in which the true scores are believed to lie, with a given level of confidence.	
30	Reliability	Test-retest	Test-retest reliability is a correlational measure based on the administration of the same test twice to the same group of examinees after a (brief) time interval has elapsed.	time between administrations; correlation coefficient
31	Reliability	Alternate form	In alternate forms reliability, two or more tests are designed to measure the same construct (McDonald, 1999).	correlation; explanation of results
	Reliability	Internal consistency	Internal consistency is the extent to which items on a test measure a construct consistently.	
32	Reliability	Coefficient alpha	Coefficient alpha is an internal consistency reliability coefficient based on the number of parts into which the test is partitioned (e.g., items, subtests, or raters), the interrelationships of the parts, and the total test score variance (Joint Standards, 1999).	
33	Reliability	KR-21	KR-21 is a reliability formula based on the number of items on a test, the mean, and the standard deviation (between 0 and 1). It should be interpreted like a correlation coefficient.	
34	Reliability	Test length/power estimates	Power estimates are statistical measures that indicate the probability that the null hypothesis will be rejected when there is a true difference (no Type II error).	probability that the test will correctly lead to the conclusion that there is a difference in performance when an alternative hypothesis is specified *t*-test, ANOVA, chi square number of items for entire test as well as reporting category (not format or number of pages)
35	Reliability	Split-half	Split-half reliability is an internal consistency reliability coefficient obtained by using half the items on the test to yield one score and the other half of the items to yield a second, independent score.	correlation coefficient with Spearman-Brown
	Reliability	Generalizability	Generalizability is the dependability of an observed score (of an individual or group of individuals) and the accuracy with which this observed score generalizes (to an individual's overall performance or to a larger group).	*G*-coefficient; percentage of variance that is explained (as opposed to due to chance)

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
36	Reliability	*G*-coefficient	*G*-coefficient is a reliability index encompassing one or more independent sources of error. It is formed as the ratio of (a) the sum of variances that are considered components of test score variance in the setting under study to (b) the foregoing sum plus the weighted sum of variances attributable to various error sources in this setting.	includes standard error of measurement, confidence intervals
	Reliability	Classification consistency	Classification consistency is the property of an instrument whereby classification decisions based on the instrument's scores are accurate and consistent. At the system level, classification consistency implies that decisions about performance drawn across measures/processes are consistent.	percentage of agreement; rationale must include explanation of how data are used discriminant analysis; mean scores and standard deviations for each performance level; kappa
37	Reliability	Correlation coefficient	Correlation coefficient is a statistical measure that compares the strength and degree of agreement between two binding classification decisions	correlation
38	Reliability	Percent correspondence	Percent correspondence is a degree of agreement between two binding classification determinants.	percent of agreement, classification error
39	Reliability	Classification error	Classification error is the likelihood that an examinee is classified (in)correctly.	probability of (mis)classification
40	Bias and sensitivity		Bias is the presence of construct-irrelevant elements that potentially advantage or disadvantage any examinee subgroup. Sensitivity is the presence of content that evokes an emotional response that inhibits examinees' ability to demonstrate what they know and can do.	
	Bias and sensitivity	Linguistic	Linguistic issues pertain to examinees' native and/or home language or dialect (e.g., Spanish, Tagalog, Mandarin), and to examinees' status as English learners in general.	expert review of item content and wording (e.g., avoidance of topics or idioms that may not be familiar to English learners) DIF analysis on subgroups with different home/native languages
41	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	review by representative experts and/or members of the community/target population(s)
42	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	significance level and discussion of interpretation

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
	Bias and sensitivity	Ethnicity/race	Ethnicity and race issues pertain to examinees' ethnicity and/or race (e.g., Asian, African-American).	expert review of item content and wording (e.g., avoidance of topics or vocabulary that may be offensive or sensitive for people belonging to a particular ethnicity or race) DIF analysis on ethnic/racial subgroups
43	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
44	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	
	Sensitivity	Cultural/religious	Culture and religious issues pertain to examinees' cultural and religious practices (e.g., celebration of holidays such as birthdays, Halloween, or Christmas).	expert review of item content and wording (e.g., avoidance of topics or vocabulary that may be offensive or sensitive for people belonging to a particular cultural or religious group)
46	Sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
	Bias and sensitivity	Geographic	Geographic issues pertain to examinees' home geographic location and urbanicity (e.g., Northern, Southern, rural, urban).	expert review of item content and wording (e.g., avoidance of topics or vocabulary that may be unfamiliar, offensive, or sensitive for people living in a particular geographic location) DIF analysis on geographic subgroups
47	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
48	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	
	Bias and sensitivity	SES	SES issues pertain to examinees' socio-economic status (typically determined by indicators such as eligible to receive free or reduced-price school lunch).	expert review of item content and wording (e.g., avoidance of topics or vocabulary that may be unfamiliar, offensive, or sensitive for people of a particular SES) DIF analysis on SES subgroups

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
49	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
50	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	
	Bias and sensitivity	Disability	Disability issues pertain to examinees' physical and/or cognitive ability (e.g., blind, deaf, wheelchair user, learning disabled).	expert review of item content, wording, and format (e.g., avoidance of topics or vocabulary that may be unfamiliar, offensive, or sensitive for people having a particular disability; accessibility of the format) DIF analysis on disability subgroups
51	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
52	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	
	Bias and sensitivity	Gender	Gender issues pertain to examinees' gender.	expert review of item content and wording (e.g., avoidance of topics or vocabulary that may be offensive or sensitive for people of a particular gender or gender orientation) DIF analysis on male/female subgroups
53	Bias and sensitivity	Expert review	Expert review for bias and sensitivity is a method in which individuals with knowledge of (and often, membership in) a subgroup evaluate the items in a test or item pool to ensure that the items do not give unfair advantage or disadvantage to any examinee subgroup.	
54	Bias and sensitivity	DIF analysis	DIF is a statistical property of a test item in which different but otherwise comparable groups of examinees who have the same total test score have different average item scores or, in some cases, different response patterns.	

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
Field testing				
	Validity	Content	Content validity is the degree to which the items on an instrument are representative of the questions that could be asked about the content.	not embedded: degree to which the items are representative of the questions that could be asked about the content; the degree to which the pool of items contains the breadth and depth of the content/standards that are assessed embedded: degree to which the forms reflect the requirements of the test blueprint—this may occur over time
55	Validity	Blueprint	The field test blueprint communicates the structure and contents of a field test, including the relative weighting or distribution of strands of content.	could occur over multiple administrations if specified table or chart showing the content distribution and item type, etc.
56	Validity	Sampling	Sampling is the process of selecting a number of examinees from a population in such a way that they are representative of the population intended to be tested.	method (random sampling, as opposed to convenience sampling, is preferred); description of sample (e.g., home language, time in US schools, native language); characteristics (language-related/ language-development-related, in particular); the quality of sampling is that it shows fidelity to the assessment's intended purpose (fidelity is the degree to which the norming population is representative of an instrument's identified target population); sample size (N) is large enough to cover the range of examinees/population characteristics targeted (e.g., 30 examinees per "cell")
57	Validity	Norming	Norming is the use of field test results to make decisions about test performance with respect to a reference group that permits meaningful comparisons to other individuals or generalizations to the population.	descriptive statistics or IRT statistics; how the items performed for the range of examinees (degree to which items performed with respect to the purpose of the test and the population tested); should have a purposive sample which shows oversampling of target subgroups tending to have low numbers and include calibration for these subgroups
Scoring				
	Validity	Content	Content validity is the degree to which the items on an instrument are representative of the questions that could be asked about the content. For scoring, content validity is the degree to which the test content is meaningfully measured quantitatively or qualitatively.	
58	Validity	Rubric	A rubric is the established criteria, including rules, principles, and illustrations, used in scoring responses.	rubric standardizes the scoring process; levels/elements within a rubric are discernible and real
59	Validity	Scale	Scores are arrayed on a numerical scale with the intention of quantifying examinee performances and providing a means for comparing scores across performances/examinees.	meaningful differentiation of examinee performance; appropriate range; lends itself to evaluation of examinee performance

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
60	Validity	Standard setting (cut score and proficiency levels)	Standard setting is a method/process for establishing points on a scale such that scores at or above a point are interpreted differently from scores below that point (NCES).	defensible; cut scores are neither arbitrary nor capricious; method(s)/experts used; standard error of measurement; number of participants
61	Validity	Training of scorers/scoring protocol	Training of scorers/scoring protocol is an established system with materials for training scorers.	clear protocol; evidence of calibration; anchor papers, etc. (as appropriate); monitoring/auditing procedure
	Reliability	Inter-rater reliability	Inter-rater reliability is an approach to reliability where the researcher compares the scores generated by two (or more) raters.	level of agreement; stated rating process; degree of fidelity to rating process
62	Reliability	Correlation (kappa)	Correlation (kappa) is a statistical measure that compares the strength and degree of agreement between two (or more) different raters.	coefficient
63	Reliability	Percent correspondence	Percent correspondence is a measure of inter-rater agreement, usually reported at the item level, defined as the share of examinee responses on which multiple raters agree.	percent of agreement; classification error; rationale Agreement can also be defined as within one rating category, within two, etc.
Reporting				
	Validity	Consequential	Consequential validity is the degree to which results are used in a manner consistent with the intended purpose and uses of the assessment.	
64	Validity	Reporting category	Reporting categories are the categories/labels associated with scores (e.g., standard, objective, examinee-level expectation, examinee-level, school-level, state-level, performance-level).	appropriate level of granularity/detail (unit of analysis); consistent with purpose of assessment and intended use of results
65	Validity	N	N is the number of examinees tested.	subgroup numbers; minimum N (which examinees/groups are excluded)
66	Validity	Central tendency/variation	Central tendency/ variation is the average or typical score attained by a group of subjects.	mean (average); median (middle score); standard deviation (variability from the mean); range; shape of distribution; frequencies
67	Validity	Effect size	Effect size is a statistic representing the magnitude of an effect and its practical significance so that outcomes of the assessment(s) can be compared to other measures for validation (N size taking ELP tests tends to be small; therefore, effect size is a means for examining practical significance for the population of examinees even with an absence of statistical significance).	method/formula

Appendix B: ELL Assessment Technical Criteria—Operational Definitions

		Type / Element: Evidence/Method	Operational Definition	Notes
Security				
	Validity	Consequential	Consequential validity is the degree to which results are used in a manner consistent with the *intended* purpose and uses of the assessment. In terms of security, scores can be used/interpreted in a manner consistent with the test's purpose.	
68	Validity	Protocols	Test security protocols are systems established to ensure test security—that is, the extent to which access to the specific content of a test has been limited to those who need to know it for test development, test scoring, and test evaluation (Based on Joint Standards, 1999).	systematic; clear; adequate/appropriate for ensuring security (including limiting access/distribution)