THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

Framework for Collecting Evidence for Test Validation

The Framework for Collecting Evidence for Test Validation[1] defines categories of evidence and their supporting documentation at each phase of the test development process. State and Local Education Agencies may benefit from this framework, as it provides guidance about the specific types of evidence that should be collected during each of the following phases: (1) Test Design and Development; (2) Field Testing; (3) Test Administration; (4) Scoring; and (5) Reporting. Within the Test Design and Development phase, categories of evidence have been grouped by item level validity, test level validity, item level reliability, and test level reliability.

PHASE I: TEST DESIGN AND DEVELOPMENT[2]

Validity: Item Level

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	Ensure test captures all elements of construct as intended
Test purpose	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 testing (e.g., placement, classification, measuring growth/achievement)
Population/ Classification	The set of examinees for whom the test is intended for the purpose(s) stated.	Clearly defined population; geographical location

[1] Unless otherwise noted, guidelines provided in this framework are drawn from Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014).
[2] Documentation is required at both the item and test levels, so evidence collection strategies are provided separately.

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
Theoretical foundation/ framework	Underlying framework, model, or perspective that defines the domain being measured and how best to measure it.	Clearly stated, coherent, current/accepted theories
Universal Design (UD)	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.	Specific and/or explicit evidence of application of UD principles during design and development phase
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.)
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).	Ensure test captures all elements of content as intended
Alignment (items-to-standards)	In-process alignment is a procedure for ensuring that test items under development are aligned to existing content standards. Ex post facto alignment is a process—usually a formal study—for evaluating whether existing items are aligned to existing content standards.	Alignment studies completed using appropriate unit(s) of analysis and appropriate model; explanation of process or results (including limitations). In-process alignment may be done by writers, editors, or other developers and expert reviewers during the item development process. Ex post facto alignment should be done by independent experts in assessment, standards, 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) of the set of standards.
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
p-values/ point biserials	P-values are the probability of correctly answering an item. Point biserials are correlations between the total test score and item score.	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
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)

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	One, two, or three parameter IRT are okay
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).	Report on the relative contribution of each factor examined; support/ verification of predictions of the relationship of the construct to the measurable factors
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 t statistic and its significance level; explanation of results
Factor analysis	Factor analysis is a statistical technique to determine if multiple variables can be described by a few factors (unidimensionality) (Fraenkel & Wallen, 2000).	Value for f statistic and its significance level; explanation of results
Bias and Sensitivity (Linguistic, Ethnicity/Race, Cultural/Religious, Geographic, SES, Disability, Gender)	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.	Correlations (factor loadings); explanation of results
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.	Expert review of item content and wording as well as associated stimuli

Validity: Test Level

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	Ensure test captures all elements of construct as intended
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
Multi-trait/Multi-method/Subtest inter-correlation	Subtest inter-correlation is evidence that the pieces of the test are measuring the same construct (e.g., subtests within the reading section). 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. 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 subtest inter-correlation be presented as evidence of construct validity.	Correlation table or MTMM matrix
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).	Ensure test captures all elements of content as intended
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 and item type, etc.
Alignment (test form-to-blueprint)	Alignment (test form-to-blueprint) is the degree to which a test form reflects the intended breadth, depth, and emphasis of content specified in the test blueprint.	Alignment studies done (independent); appropriate unit(s) of analysis and model/appropriate dimensions evaluated; explanation of results (including limitations)
Descriptive statistics	Descriptive statistics are summary measures of a distribution of scores, providing information about central tendency, location and variability.	Mean; standard deviation; N; explanation of results (e.g., evidence that field-test results were used to select appropriate items)
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)
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 instruments administered at about the same time (Fraenkel & Wallen, 2000).	

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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
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.	Ensure stated test purpose matches test use and anticipate plausible unintended outcomes
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, and entry/exit).	Proficiency level descriptors; description of range of levels of performance; 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

Reliability: Item Level

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
Reliability: Internal Consistency	Internal consistency is the extent to which items on a test measure a construct consistently.	
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).	
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.	
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)
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: Test Level

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	
SEM/Confidence Intervals	Standard Error of Measurement (SEM) 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.	
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
Alternate Form	In alternate forms reliability, two or more tests are designed to measure the same construct (McDonald, 1999).	Correlation; explanation of results
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	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
Correlation coefficient	Correlation coefficient is a statistical measure that compares the strength and degree of agreement between two binding classification decisions.	Correlation
Percent correspondence	Percent correspondence is a degree of agreement between two binding classification determinants.	Percent of agreement, classification error
Classification error	Classification error is the likelihood that an examinee is classified incorrectly.	Probability of (mis)classification

PHASE II: FIELD TESTING

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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 of 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
Test 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. Make transparent any changes in content assessed
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; characteristics; 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")
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

PHASE III: TEST ADMINISTRATION

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	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
Accommodations	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.	Theoretically, allowed accommodations do not alter the construct assessed and do not affect reliability of measure
Fidelity	Fidelity is the degree to which the protocol for standardized test administration is followed.	Test administration conditions/procedures do not alter the construct; make transparent any changes in administration guidelines
Standardization	Standardization means having rules and specifications for testing procedures that are intended to ensure testing conditions are the same for all examinees.	Level of detail and degree to which they ensure standardized testing conditions
Validity: Consequential (Test Security)	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.	Security protocol for development, administration, scoring, and reporting (nondisclosure, confidentiality, erasure analysis)
Protocols	Test security protocols are systems established to prevent viewing, publication, or unauthorized copying of test materials.	Systematic; clear; adequate/appropriate for ensuring security (including limiting access/distribution)

PHASE IV: SCORING

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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
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 discernable and real. Make transparent any changes in scoring procedures
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
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
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 and degree of fidelity to rating process
Correlation (kappa)	Correlation (kappa) is a statistical measure that compares the strength and degree of agreement between two (or more) different raters.	Coefficient
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.
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.	DIF analyses at subgroup level (e.g., Linguistic, Ethnicity/Race, Cultural/ Religious, Geographic, SES, Disability, Gender)
DIF analysis	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

PHASE V: REPORTING AND INTERPRETATION OF SCORES

CATEGORY OF EVIDENCE	OPERATIONAL DEFINITION	COMMENTS ABOUT DOCUMENTATION
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.	Monitor intended and unintended outcomes and/or test misuse
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).	Score reports have an appropriate level of granularity/detail (unit of analysis); consistent with purpose of assessment and intended use of results; clarity and coherence of presentation
N	N is the number of examinees tested.	Subgroup numbers; minimum N (which examinees/groups are excluded)
Central tendency/ Variation	Central tendency/variation is the average or typical score attained by a group of subjects.	Means (average)/medians (middle score); standard deviation (variability from the mean); range; shape of distribution; frequencies
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
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, and entry/exit).	Fidelity between stated purpose of assessment; guidelines for how results should be interpreted, reported, and used -- look at, for example, sample reports, scoring outcomes/results; includes item release strategy

THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION WestEd CRESST

CSAI Update is produced by the The Center on Standards and Assessment Implementation (CSAI). CSAI, a collaboration between WestEd and CRESST, provides state education agencies (SEAs) and Regional Comprehensive Centers (RCCs) with research support, technical assistance, tools, and other resources to help inform decisions about standards, assessment, and accountability. Visit www.csai-online.org for more information.

This document was produced under prime award #S283B050022A between the U.S. Department of Education and WestEd. The findings and opinions expressed herein are those of the author(s) and do not reflect the positions or policies of the U.S. Department of Education.

WestEd

WestEd is a nonpartisan, nonprofit research, development, and service agency that partners with education and other communities throughout the United States and abroad to promote excellence, achieve equity, and improve learning for children, youth, and adults. WestEd has more than a dozen offices nationwide, from Massachusetts, Vermont and Georgia, to Illinois, Arizona and California, with headquarters in San Francisco.

For more information, visit WestEd.org; call 415.565.3000 or, toll-free, (877) 4-WestEd; or write: WestEd / 730 Harrison Street San Francisco, CA 94107-1242.