THE CENTER ON STANDARDS & ASSESSMENT IMPLEMENTATION
WestEd CRESST

Key Questions for Test Development Practices

The Key Questions for Test Development Practices document highlights guiding questions and examples of evidence for various categories of reliability and validity concerns. State and Local Education Agencies may benefit from this framework, as it provides guidance around the high-priority elements of evidence for emerging assessments based on the Next-Generation Science Standards. These high-priority elements are grouped by evidence categories, and include guiding questions and examples of evidence.

CATEGORY OF EVIDENCE	GUIDING QUESTIONS	EXAMPLES OF EVIDENCE
Construct Validity	• Are the test purpose, target audience, and intended uses clearly specified? • What evidence suggests that the test captures all elements of the science constructs as intended? • What theory of action or theoretical foundation describes the connection between test results and intended claims? • What evidence supports the intended interpretations of scores from these measures? • What information was considered during identification of the most appropriate item types for use on this assessment? What measurement theory of action or theoretical foundation supports those decisions? • Did evidence of construct underrepresentation or introduction of construct irrelevant variance emerge? • What efforts were taken to ensure standardized administration conditions? • What steps were taken to maintain test security?	• Communiqués to stakeholders that communicate test purpose, target audience, and intended uses • Theory of action diagrams • Relevant research citations in documentation • Documentation of steps taken during item development • Evidence of teacher participation during all steps of work • Findings from alignment studies that verify depth of knowledge measured • Findings from dimensionality studies • Administration guides • Test security procedures and protocols • Technical reports

CATEGORY OF EVIDENCE	GUIDING QUESTIONS	EXAMPLES OF EVIDENCE
Content Validity	• What evidence suggests that test items are measuring the full depth and breadth of the NGSS? In what ways was matrix sampling used to promote full coverage? • In what ways were state educators at key grades involved in decision-making about the appropriateness of content assessed? • How are states ensuring that students have the opportunity to learn tested content? • Were items appropriately field-tested prior to operational testing and feedback collected from educators and students? • If multiple test forms are used, how are forms linked or equated?	• Findings from studies that verify alignment to intended content • Documentation of content sampling plan • Recommendations from content reviews • Findings from expert reviews • Item-level statistics collected during field-testing such as p-values, point biserial, and item characteristic curves • Test-level descriptive statistics such as mean, standard deviation, N-counts, and test characteristic curves • Report on equating methodology and findings • Findings from tests of fit, structural equation modeling, ANOVAs, or factor analyses • Test blueprints • Field-test sampling plan and findings • (e.g., placement, classification, measuring growth/achievement)
Consequential Validity	• In what ways were plausible unintended outcomes considered? • How were performance levels and cut-scores determined? • How are results being used? • Was the test susceptible to misuse? • Have any issues related to ethics or equity emerged in relation to this assessment? • To what extent are results meeting the needs of or benefitting different stakeholder groups? • Has new evidence emerged that calls into question the current interpretation of test scores?	• Findings from surveys of students, teachers, or parents • Documentation from standard setting (participants, methodology, outcomes) • Documentation of test security violations

CATEGORY OF EVIDENCE	GUIDING QUESTIONS	EXAMPLES OF EVIDENCE
Reliability	• What evidence suggests the tests meet industry standards for reliability as a measure of students' annual achievement in relation to the Next Generation Science Standards? • Were items appropriately field-tested prior to operational testing?	• Documentation of scoring methods • Estimates of internal consistency • Documentation of test What evidence was collected to support use of this assessment with the target population? • What efforts were taken to examine the appropriateness of this measure for students from diverse geographic, cultural, linguistic, religious, or socioeconomic backgrounds? • What evidence suggests this measure is appropriate for students with disabilities or English language learners?length • Information about scale used and range of student performance • Scoring guides and rubrics • Manuals for training hand-scorers • Findings from inter-rater reliability, split-half, test-retest, alternate forms, or generalizability studies • Reporting of standard error of measurement or use of confidence intervals
Fairness	• What evidence was collected to support use of this assessment with the target population? • What efforts were taken to examine the appropriateness of this measure for students from diverse geographic, cultural, linguistic, religious, or socioeconomic backgrounds? • What evidence suggests this measure is appropriate for students with disabilities or English language learners?	• Documentation of application of principles of universal design for assessment during development • Recommendations from bias/sensitivity reviews • Results from DIF analyses • Findings from expert reviews • Documentation of population included for field-testing • Description of allowable test accommodations
Feasibility	• Are the assessments administered, responses scored, and results reported in cost-efficient and responsible ways? • How much time is required to administer this test at each grade? • What special skills are required for test administrators and scorers? • In what ways was technology used to promote administration, scoring, or reporting efficiencies?	• Findings from surveys of students, teachers, or parents • Administration windows and testing time estimates • Findings from impact analyses, utilization studies, or benefit-cost analyses

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.