Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS)

Ensuring Rigor in Local Assessment Systems:
A Self-Evaluation Protocol

Ensuring Rigor in Local Assessment Systems: A Self-Evaluation Protocol was developed with funding from the US Department of Education under Enhanced Assessment Grants Program CFDA 84.368A. The contents do not necessarily represent the policy of the US Department of Education, and no assumption of endorsement by the Federal government should be made.

Table of Contents

List of Exhibits

The *Standards for Educational and Psychological Testing* are referenced throughout this document. Its citation is:

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for Educational and Psychological testing*. Washington DC: AERA.

Background

Every US state and school district uses one or more assessments of students' academic knowledge and skills for a variety of purposes. This self-evaluation protocol is designed to support local educators (including but not limited to, district test coordinators, curriculum specialists, principals, and/or teachers) in evaluating each of these assessments as well as their local assessment system, as a whole. We suggest using an inclusive process with this protocol, with multiple individuals contributing as a team and with the understanding that this process may lead to some internal debate on the value and purpose of assessment within your school or district.

Why evaluate assessments?

An assessment system at a school or school district level should provide students, teachers, administrators, and school personnel with an accurate reflection of the key concepts, knowledge, and skills that students have achieved for a range of purposes. Each assessment within the local assessment system should yield information that is meaningful and useful for a particular purpose or purposes. The only way one can know if an assessment yields valid and useful information is to evaluate evidence in relation to how its scores are to be interpreted and used. This process, known as **validity evaluation**, is what this protocol is designed to support.

Addressing questions about the **validity** and **reliability** of assessments is an essential obligation of any person or agency using test scores to make judgments about any individual or group. This obligation applies whether a test is teacher-made for a class or produced commercially for large-scale use. What differs are expectations for the nature and degree of evidence necessary to support the interpretations and uses of the test scores. Note that validity and reliability are not characteristics of a test itself: they apply to the scores a test yields and the uses for those scores. A test is not inherently good or bad, but its scores can be used for appropriate or inappropriate purposes.

This notion of validity in relation to scores and score uses is so fundamental that it is the very first standard in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), the document that guides all educational and psychological assessment practices in the US.

> **"Standard 1.0**. Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided."
>
> (AERA, APA, & NCME, 2014, p. 23)

For the present purposes, we reflect this concept in foundational questions that underlie this self-evaluation protocol:

> For what purpose(s) was the assessment developed? Is the purpose for which you are using the assessment among those purposes for which it was developed?

The protocol begins and ends with a consideration of purpose.

What is this protocol designed to do?

This self-evaluation protocol provides a framework for educators at a school, school district, or local system level to use in considering how to best implement an assessment system. It is designed to focus on assessments that are state- or district-mandated, developed by an independent test vendor, and selected for use within a school throughout the school year. Scores from these assessments may be used as part of official accountability programs; others may inform instruction or yield information for use in assigning grades. All tests that yield scores used for any of these purposes are part of a school's or district's assessment system and should be evaluated on a regular basis.

This protocol is meant to support reviews for each assessment in a system and for a local assessment system, as a whole. Educators at a school or school district level can use and modify this protocol as needed to best suit their needs. It may be helpful to consider each assessment from multiple perspectives, such as those of test administrators, teachers, parents, and students. Different stakeholders may hold different views on what scores mean and how they should be used; it may be necessary to determine which interpretations and uses are supported by evidence and which are not.

The SCILLSS Digital Workbook on Educational Assessment Design and Evaluation is designed as a resource for those implementing the self-evaluation protocol. The workbook encompasses five chapters that together are intended to provide state and local educators with a grounding in the principles for high quality assessment. Such principles are critical to the appropriate selection, development, and use of assessments in educational settings. While this digital workbook is not a toolkit for developing assessments, it offers a framework for making decisions about whether to develop or adopt tests and for evaluating tests currently in use. The workbook is designed to be used on its own or as a resource for those completing the SCILLSS self-evaluation protocols at the local or state level. The five chapters comprising the digital workbook are available on the SCILLSS website here.

Guidelines for Implementing the Self-Evaluation Protocol

We recommend four steps in implementing this protocol:

1. Articulate your current and planned needs for assessment scores and data
2. Identify all current and planned assessments
3. Gather and evaluate the evidence for each assessment
4. Review the evidence across assessments

Next, we provide considerations and guidelines for preparation prior to implementing the self-evaluation protocol. We also recommend your team gather information on each of the assessments administered at the school or district level. This information includes, but is not limited to, assessment purposes and uses, assessment technical manuals, assessment research conducted by publishers/test vendors and/or by outside researchers, and administration manuals for the assessment.

We offer suggestions for implementing the self-evaluation protocol via the four steps that follow.

Self-Evaluation Protocol, Step One: Articulate your current and planned needs for assessment scores and data

Step one involves identification of your intended purposes and uses for test scores resulting from individual assessments. Assessments are tools for producing information to help answer questions. What questions do you have about student achievement and what information can assessments provide to help you answer those questions? In other words, what are your intended uses of assessment scores? For what purposes will they be used?

Further, it's important to identify the stakes associated with the intended uses of test scores. The higher the stakes, the greater the burden for adequate evidence to support score meaning and use.

Assessment uses and associated stakes often include those presented in Exhibit 1.

Exhibit 1. Assessment Uses and Associated Stakes

Information educators use to:	**Information educators use to:**	**Information administrators use to:**
• guide next steps in instruction	• evaluate learning for calculating grades	• evaluate teachers
• evaluate instruction	• determine eligibility for program entry or exit	• evaluate schools or districts
• evaluate curriculum	• diagnose learning difficulties	• evaluate programs or services
These uses are more formative. They have relatively **low stakes** for students and educators, as long as scores are considered in combination with other information and decisions allow for flexibility in implementation.	These uses have **high stakes** for individual students and scores must always be considered in combination with other information.	These uses have **high stakes** for educators and scores must always be considered in combination with other information.

Self-Evaluation Protocol, Step Two: Identify all current and planned assessments

The second step in using this self-evaluation protocol involves identification of the complete array of assessments you use or plan to use to address specified needs. You could organize your list of needs and associated assessments by content area, grade level, or another set of categories or dimensions to facilitate your review. It may be helpful, for example, to see assessments used for grading in math across grade levels or the set of assessments used at a particular grade level across content areas.

As you complete your identification of assessments, you may find areas where you have some overlap—two or more tests that yield scores used for a common purpose—as well as areas of gap where you do not have an assessment that could provide relevant information. Both situations can be appropriate. Areas of overlap can allow for multiple sources of data to enhance decisions based on those data. Likewise, a gap could mean that you gain adequate information from non-test sources. Alternatively, too much overlap can signal a need to reduce testing to conserve instructional time and other resources, and a gap could mean that you are missing a valuable piece of a puzzle. Only you and your decision-makers can determine what makes most sense in your system.

The Self-Evaluation Protocol, Steps One and Two worksheet is intended to guide you through the intended purposes, assessment uses, and associated stakes. Complete this form for each content area, grade level, or other set of dimensions. When complete, this form will help you take stock of the assessments across a specific content area/grade level and determine where gaps or overlaps exist, if any.

Self-Evaluation Protocol, Step Three: Gather and evaluate the evidence for each assessment

Once you have identified each need/purpose, intended uses, and stakes and associated assessments, it is time to compile and review evidence regarding the interpretations and uses of the assessment scores. First, it would be appropriate to consider what types of data and evidence are available to help support the use of test scores, and how and when this information was collected. The data and evidence should be available to address different aspects of the assessment, such as: 1) the development of the assessment, 2) how the assessment is administered, 3) how the assessment scores are created and distributed, and 4) how the assessment scores are used within your school or district. The data and evidence can come from various sources, including directly from the test publisher through a technical manual, special studies conducted by the test publisher, or independent research conducted by other entities.

As you review the evidence, you will reach a conclusion as to whether the evidence available can be considered *Adequate*, *Incomplete*, or *Lacking*. Evidence that is considered *Adequate* provides sufficient data and information to demonstrate that the components of the assessment—how it is designed, administered, scored, and reported—directly support the intended test score interpretations and uses across the full range of students taking the assessment. Evidence that is considered *Incomplete* provides some, but not all, of the necessary data and information; gaps may be evident and critical information for establishing a comprehensive argument for the validity of the intended test score interpretations and uses may be missing. Evidence that is considered *Lacking* provides little or no evidence, and does not provide sufficient data to support any of the intended test score interpretations and uses.

Below, we pose the four key validity questions necessary to guide the collection of evidence to support or refute the validity of interpretations and uses of the assessment scores. Each of these key validity questions is supported by the *Standards,* and the most critical related standard or standards are outlined within each section to highlight the relationship between the two. The guiding questions in each of the exhibits below are accompanied by evidence examples and are intended to support the evaluation and gathering of evidence for each assessment.

Evidence for Construct Coherence

Key Validity Questions: Does the assessment have evidence for construct coherence with your overall standards and curriculum? Has the assessment been designed in such a way to ensure that the content of the assessment is consistent with your state standards and the curriculum in the classroom? In other words, to what extent does the assessment as designed capture the knowledge and skills defined in the target domain?

Standard 1.1 demands "an analysis of the relationship between the content of a test and the construct it is intended to measure" (AERA, APA, & NCME, p. 23). A construct is the concept or characteristic that a test is designed to measure. Construct coherence ensures the assessment and its operational system have been designed to yield scores that reflect the construct represented in the academic content standards and that complement and support the knowledge and skills prioritized for instruction and assessment across the larger educational setting. Test developers should "document the extent to which the content domain of a test represents the *domain* defined in the test specifications" (AERA, APA, & NCME, p. 89).

Construct coherence strengthens the validity of interpretations and uses of assessment scores and their intended purposes. Guiding questions to support gathering evidence for construct coherence are included in Exhibit 2.

Exhibit 2. Evidence for Construct Coherence

Construct Coherence Guiding Questions	Examples of Evidence for Construct Coherence
1. What are you intending to measure with this test?	The test publisher clearly defines the purpose(s) of the test, what is to be measured by the test, the population(s) for which the test is intended, and how the test scores are to be interpreted and consequently used. A district provides documentation that summarizes the alignment between the measurement targets on the assessment and the academic content standards targeted through classroom instruction and assessment.
2. How was the assessment developed to measure the measurement targets?	The test publisher documents how the domain is defined, including how the test was designed and by whom to address an appropriate range of knowledge and skills at appropriate levels of difficulty and complexity. The test publisher provides a test blueprint or test specifications that define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of the items and sections, and has documented the extent to which the content of the test represents the domain defined in the test specifications. The test publisher provides item specifications and describes the qualifications of item writers and how they were trained to write items for the test.
3. How were items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement target(s) and not other content, skills, or irrelevant student characteristics?	The test publisher has documented a rigorous development and field-test process; all test items are reviewed multiple times by experienced test development professionals and are screened to ensure that they are appropriate for the test form. Documentation includes evidence of how and when items were pilot-tested and field-tested and includes information about how results of those processes were used to improve individual items and the bank of items as a whole.
4. How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes	The test publisher provides a scoring report that documents the procedures used for scoring the items, and provides scorer training materials as

Construct Coherence Guiding Questions	Examples of Evidence for Construct Coherence
evaluated to ensure they accurately capture and assign value to students' responses?	appropriate, including rubrics and examples of responses for each level in the scoring rubrics. A district provides documentation of efforts to ensure interrater reliability and the standardized application of scoring rules and procedures.
5. How are scores for individual items combined to yield a total test score? What evidence supports the meaning of this total score in relation to the measurement target(s)?	The test publisher provides technical documentation of all scaling procedures and clearly explains the characteristics, meaning, and intended interpretations of scale scores, as well as their limitations. The test publisher provides documentation of how, when, and by whom the performance level descriptors were established and of how, when, and by whom the cut scores that separate the score ranges for each performance level were determined.
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?	One or more entities who are independent of the test developers provide reports that describe their evaluations of alignment quality. These reports should describe the methodology used for the evaluations, the qualifications of the reviewers, the results of the evaluation, and specific recommendations to the test developer for how to improve item quality.
7. How are scores reported in relation to the measurement target(s)? Do the reports provide adequate guidance for interpreting and using the scores?	The test publisher should provide score reports and accompanying documentation meant to guide those who are expected to read and understand score reports, including teachers, parents, students, administrators, and the public. Documentation should include information that describes the purpose of the test, what the scores mean, what evidence supports score meaning, and any cautions for score use. A district that wishes to use test scores to make decisions about instruction or placement establishes clear evidence to support all relevant score-based recommendations.

Evidence for Comparability and Reliability

Key Validity Questions: Are the test scores comparable, or are the test scores reliable and consistent in meaning across all students, classes, and schools? For comparability, is there evidence to support the concept that the test scores mean the same thing for all students, regardless of which year the student takes the test or the exact test form that is taken? For reliability, is there evidence that demonstrates the test scores are free of random measurement errors, and are dependable and consistent for individual test takers?

Standard 2.0 demands “appropriate evidence of reliability/precision should be provided for the interpretation for each intended score use” (AERA, APA, & NCME, p. 42). Reliability/precision refers to the degree to which test scores for a group of test takers are consistent and dependable over repeated applications of a measurement procedure. Comparability ensures the assessment system operates as intended (e.g., administration, scoring, analyses, reporting) and yields scores that are comparable in meaning across sites and time.

As with construct coherence, comparability strengthens the validity of interpretations and uses of assessment scores by ensuring that assessment scores mean what they are intended to mean and are used appropriately. Guiding questions to support gathering evidence for comparability and reliability are included in Exhibit 3.

Exhibit 3. Evidence for Comparability and Reliability

Comparability and Reliability Guiding Questions	Examples of Evidence for Comparability and Reliability
1. How is the assessment designed to support comparability of scores across forms and formats?	To ensure comparability of scores across forms, the test publisher provides a blueprint or test map within their technical documentation that defines the set of items that make up the test in terms of how many items, what kinds of items, and what each item is supposed to measure. To ensure comparability of scores across formats, the test publisher documents results of studies of test performance for equivalent groups of students who take the test in different formats. Documentation considers differences in total test scores and how students perform on the items within the test.
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?	The test publisher provides test administration manuals that outline the standardized procedures and conditions for administration and provides a means for training school staff to administer the tests and to handle testing documents. A district designates a person responsible for coordinating the testing process by ensuring that the materials are kept secure, that those proctoring the test are appropriately trained, that the tests are given to the appropriate students, and that the conditions under which students are taking the test conform to the administration guidelines.
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?	The test publisher provides information in the technical manual about how items are designed and developed to be scored accurately and consistently and presents the rubrics, criteria, or other guidance for scoring constructed-response items. A district documents the training, protocols, and processes for teachers to score student writing responses from a district-wide writing assessment, and after each response is scored multiple times by different teachers, evaluates and documents the reliability of the scoring process.

Comparability and Reliability Guiding Questions	Examples of Evidence for Comparability and Reliability
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?	The test publisher provides a technical manual that summarizes how score scales were developed and evaluated to ensure that the scaled scores are accurate and meaningful, and how score scales are equated across test administrations to support the comparison of scores across forms, sites, and time.
5. To what extent are different groups of students who take a test in different sites or at different times comparable?	A district wanting to make score comparisons for groups (e.g., comparisons across sites; comparisons of the same cohort of students across years; comparisons of different student groups such as English learners and non-English learners) provides information about (a) policies about who is tested and included in the reporting of results, (b) students' opportunities to learn the material being tested, (c) the availability and use of testing accommodations, and (d) students' motivation to take the test.
6. How are scores reported in ways that support appropriate interpretations about comparability and disrupt inappropriate comparability interpretations?	A test publisher/user reports the reliability/precision information for each test score and for observed differences between scores. The test scores are accompanied by information about how the scores are to be interpreted and used and how they should not be interpreted and used, and the score reports are clear and accessible to those who are meant to interpret and use the scores, including students, parents, and educators.
7. What evidence supports the appropriate use of the scores involving comparisons across students, sites, forms, formats, and time?	A test publisher provides score reports that present performance in levels that includes information to help test users interpret the meaning of students' performance at each level and includes text associated with each level that describes the kinds of skills that students whose test score falls into that level may have. A district provides documentation to communicate changes or alterations to an assessment and its scores across years. The district provides documentation to show they are evaluating the comparability of test forms and scores across sites, time, and varying student characteristics.

Evidence for Fairness and Accessibility

Key Validity Questions: Are the tests fair and accessible for all students? Has the test publisher provided evidence that all students can complete the assessment and fully understand the concepts being assessed? To what extent are students able to demonstrate what they know and understand in your schools and within your current curriculum?

Standard 3.2 demands that tests be designed to measure the intended construct and minimize the potential for construct-irrelevant characteristics (AERA, APA, & NCME, p. 64). Further, Standard 3.6 demands that test developers examine the evidence for validity of score interpretations across subgroups in the intended examinee population (AERA, APA, & NCME, p. 65). Considering fairness and accessibility ensures all test takers can demonstrate what they know and can do on an assessment without being impeded by characteristics of the items that are irrelevant to the construct being measured.

Considerations of fairness and accessibility strengthen the validity of interpretations and uses of assessment scores by ensuring that assessment scores mean what they are intended to mean and are used appropriately. Guiding questions to support gathering evidence for fairness and accessibility are included in Exhibit 4.

Exhibit 4. Evidence for Fairness and Accessibility

Fairness and Accessibility Guiding Questions	Examples of Evidence for Fairness and Accessibility
1. How were the needs of all students addressed during assessment development? How were the assessment questions developed to ensure that scores reflect the intended measurement targets and not student characteristics or contexts that are irrelevant to the measurement targets?	The test publisher provides documentation that demonstrates how the principles of Universal Design guided the development process. This documentation includes item writing training materials and guidelines that aid item writers in creating items that are free of potentially biasing content or features. Once developed, all items are reviewed multiple times before being used, including a fairness review and the completion of statistical reviews, such as Differential Item Functioning (DIF).
2. How were the needs of students with disabilities addressed during assessment development?	The test publisher's technical manual provides evidence of considerations of Universal Design for the assessment, and the test publisher provides an accommodations manual that specifies the allowable accommodations during the administration of the assessment. The test publisher provides documentation to describe the qualifications and involvement of the experts that contributed to the development process. The test publisher has documented who the experts are in terms of their relevant professional qualifications and experience, what the experts do or did during the development process, and how the input from the experts was used.
3. How were the needs of English learners addressed during assessment development?	The test publisher reviews the performance of English learners on all test items and completes Differential Item Functioning (DIF) analyses to ensure that items do not unfairly disadvantage English learners. The test publisher provides documentation to describe the qualifications and involvement of the experts that contributed to the development process. The test publisher has documented who the experts are in terms of their relevant professional qualifications and experience, what the experts do or did during the development process, and how the input from the experts was used.
4. How are students with disabilities able to demonstrate their knowledge and skills	The test publisher's technical manual provides evidence of pilot studies and/or cognitive labs to

Fairness and Accessibility Guiding Questions	Examples of Evidence for Fairness and Accessibility
through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?	ensure that students with disabilities can demonstrate what they know and can do when responding to the assessment items. A district documents appropriate Individualized Education Program (IEP) and 504 processes and guidelines and educators understand which accommodations are available and most suitable for each student. Once an accommodation is assigned to a student, he or she has opportunities to use the accommodation in the instructional setting prior to use on the assessment.
5. How are English learners able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?	The test publisher provides an accommodations manual that specifies the allowable accommodations for students who are English learners. A district develops and implements guidance and forms for a team of local educators who know the student and have relevant academic and linguistic expertise to use when making and documenting decisions about how individual students should participate in assessments and provides effective training on test participation and the selection of accommodations for the decision-making teams for English learners.
6. How are students' responses scored in ways that reflect only the construct-relevant aspects of those responses? What evidence supports the minimization of construct-irrelevant influences on students' responses?	The test publisher provides technical documentation that describes how the scoring processes, including rubrics for scoring constructed-response items, have been designed to recognize and appropriately value construct-relevant aspects of students' responses and minimize the influence of construct-irrelevant aspects. For any items requiring human scoring, the test publisher has provided extensive training for all graders, including information to ensure that all scores are based upon key aspects of the measurement targets. The scoring process also has multiple quality control steps, such as auditing graders throughout the entire scoring window to ensure that all scoring is consistent with the item rubric.

Fairness and Accessibility Guiding Questions	Examples of Evidence for Fairness and Accessibility
	The test publisher provides the results of r analyses to identify and explain group differences in test or item performance.
7. What evidence supports the interpretation and use of students' scores in relation to their learning opportunities?	A district has documentation that describes opportunities for teachers to evaluate assessment scores in relation to the curriculum, instruction, and learning taking place in the classroom. A district takes into account not just scores from a single test but other relevant information when making a decision or characterization that will have a major impact on a student.

Evidence Related to Consequences and Use

Key Validity Questions: Does the use of the test scores lead to positive consequences for your students, schools, and teachers? To what extent does the test yield information that is used appropriately within a system to achieve specific goals? For example, has the test publisher provided sufficient information to allow school personnel to review the assessment results, determine appropriate follow-up steps, and identify the resources necessary to complete all follow-up activities?

Standard 7.0 demands that information relating to tests be clearly documented so that test users can "make informed decisions regarding which test to use for a specific purpose, how to administer the chosen test, and how to interpret the scores" (AERA, APA, & NCME, p. 125). Considering the implications of **consequences** when developing assessments ensures the assessment yields information that can be and is used appropriately within a system.

Considering the implications of consequences in conjunction with construct coherence, comparability and reliability, and fairness and accessibility strengthens the validity of interpretations and uses of assessment scores for their intended purpose(s). Guiding questions to support gathering evidence for consequences and use are included in Exhibit 5.

Exhibit 5. Evidence Related to Consequences and Use

Consequences and Use Guiding Questions	**Examples of Evidence Related to Consequences and Use**
1. Are the items and content of the test consistent with the standards being measured to ensure appropriate uses?	The test publisher provides technical documentation that includes a clear and specific definition of what the test is intended to measure. The test publisher provides a blueprint or other framework that defines what is on the test along with a description of how the framework was developed and how that process meets industry standards for quality and rigor; the test publisher also describes how items were developed to reflect the blueprint and how that process meets industry standards for quality and rigor. The test publisher provides reports from independent evaluations of the test framework and the test items that support the vendor's claims about what the test is designed to measure and how well it reflects that design.
2. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of inappropriate uses by students, teachers, or administrators? What evidence supports the implementation and effectiveness of these efforts?	A district establishes processes and procedures for ensuring that during test administration, teachers, administrators, and others serving as proctors, follow the test administration guidelines that the test publisher provides and remove any conditions that may be obstacles for students during testing. A test publisher provides score reports that include only scores for which there is adequate validity and reliability evidence. Accompanying documentation should provide interpretive guidance that characterizes the interpretations and uses that are intended and supported by adequate validity evidence and also cautions against interpretations and uses for which there is not adequate validity evidence.
3. What evidence is available to support the use of test scores across the entire score scale and all performance levels?	The test publisher provides reports and technical manuals that describe how the items and the performance or achievement level descriptors were developed to reflect the entirety of the score scale and to focus on points of the score scale associated with intended interpretations and uses.

Consequences and Use Guiding Questions	Examples of Evidence Related to Consequences and Use
	The test publisher provides overall reliability/precision indicators for the test and for the points on the score scale associated with specific intended interpretations and uses. This information would be updated after each administration cycle.
4. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your state or local district? How well do these uses align?	The test publisher provides technical manuals or reports on test development that clearly describe how the scores are intended to be interpreted and used and the evidence that supports these claims about score interpretation and use. The test publisher also cautions against unsupported score interpretations and uses and provides rationales for these cautions. A district evaluates how scores are, or are intended to be, interpreted and used and how these interpretations and uses compare with those the test publisher has described. A district evaluates the consequences associated with score interpretations and uses beyond those the test publisher describes.
5. If assessment scores are associated with recommendations for instruction or other interventions for individual students, groups of students, or the whole-class what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?	The test publisher documents evidence that the test was developed to yield scores that could identify students' instructional needs; this evidence includes information about the cognitive and learning models that provide the foundation for test development. The test publisher also documents empirical studies of the effectiveness of the instructional guidance for students in all levels of the score scale. A district evaluates the efficacy of the instructional guidance provided by the test publisher as it is implemented in classrooms to ensure the guidance is beneficial to teachers and student outcomes.
6. If assessment scores are associated with high stakes decisions for teachers, administrators, schools, or other entities or individuals, what evidence supports such interpretations and uses of these scores?	The test publisher provides technical documentation that describes how tests have been designed, developed, administered, scored, and reported in ways that support claims that the scores can be appropriately used in making high stakes decisions. A district develops a Theory of Action that describes how the scores are to be used along

Consequences and Use Guiding Questions	Examples of Evidence Related to Consequences and Use
	with the other information that will guide high stakes decisions and provides a rationale for why test scores contribute important information to the specific high stakes decisions to which they will be put. A district gathers evidence of the efficacy of test score use for high stakes decisions including evidence that the intended outcomes are being achieved and negative, unintended outcomes are being avoided.
7. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?	The test publisher includes information about score reliability/precision on all score reports and describes this information in the materials that accompany the reports in ways that people who are not testing specialists can understand. A district identifies the full range of student and parent communication needs and establishes strategies for addressing those; this would include the range of languages that students and their parents speak and read and the means for getting the necessary information to the right individuals. A district gathers or produces descriptive information about tests prior to testing and interpretive guidance to accompany score reports in as many languages and forms as possible and prepares teachers and administrators to help students and parents appropriately interpret test score information and use it in making sound decisions.

The Self-Evaluation Protocol, Step Three worksheets are intended to capture the necessary details for determining the adequacy of the evidence for each assessment in an assessment system. It may be helpful for each individual on your evaluation team to complete the Step Three worksheets independently to ensure multiple perspectives and viewpoints are represented as part of the assessment evaluation process. It will be necessary to compile all known and available assessment documentation prior to completing step three. For each question across each of the key validity categories, we recommend that you:

- consider and document the evidence for the interpretations and uses of the assessment scores for each question,
- summarize the evidence related to each question, and

- capture any important or useful comments that may support determination of the adequacy of the evidence.

The adequacy of the evidence is determined by your judgment in consideration of your state or local educational context and assessment system.

For each of the key validity areas (e.g., consequences and use, fairness and accessibility), the worksheets offer spaces for you to record ratings and capture total scores at the top of the first page of the worksheet. These scores provide a way to quantify the strength of the evidence: 1) Low (0-6 points), 2) Moderate (7-10 points), or 3) Strong (11-14 points).

Self-Evaluation Protocol, Step Four: Review the evidence across assessments

Once you have completed steps one through three of the self-evaluation protocol, it is time to review and evaluate how well your assessment system supports your primary purposes and uses. For this component of the work, it will be important to review each assessment purpose and use and identify areas with adequate evidence for the test score use and others where the degree of data and evidence is not as substantial.

As noted in step three, for each of the key validity areas, the total scores can be recorded at the top of the first page of the step three self-evaluation protocol worksheets. These scores provide a way to quantify the strength of the evidence: 1) Low (0-6 points), 2) Moderate (7-10 points), or 3) Strong (11-14 points). For each assessment, these scores can then be transferred to the Self-Evaluation Protocol, Step Four worksheet.

As you consider all of the characteristics of your assessment system and how it has been implemented in your school or district, it will be essential to view this evidence from a holistic perspective.

For uses of the test scores that appear to have *strong evidence*, consider whether the accumulated evidence gives you complete confidence in that particular use of the test scores.

In the event that data and evidence are missing, one important consideration is whether or not this purpose and use should be considered essential or even warranted. Another important question is whether there is a plan in place by the test publisher or others to evaluate the uses of the test scores.

Self-Evaluation Protocol, Steps One and Two: Identifying Purposes and Assessments Used to Serve those Purposes

Need/purpose	Assessment(s) Used to Serve this Purpose

Self-Evaluation Protocol, Step Three: Gather and Evaluate the Evidence for Each Assessment

Name of Assessment:

Who takes this test?

Key Validity Area	Score	Low	Moderate	Strong
		(0-6)	(7-10)	(11-14)
Construct Coherence:	________	□	□	□
Comparability & Reliability:	________	□	□	□
Fairness & Accessibility:	________	□	□	□
Consequences & Use:		□	□	□

How are scores used?

Low stakes for educators and students		**High stakes for students**		**High stakes for educators**	
To guide next steps in instruction	□	To evaluate learning for calculating grades	□	To evaluate teachers	□
To evaluate instruction	□	To determine eligibility for program entry or exit	□	To evaluate schools or districts	□
To evaluate curriculum	□	To diagnose learning difficulties	□	To evaluate programs or services	□
Other uses:		Other uses:		Other uses:	

Measurement targets (the concepts, knowledge, and skills this test is meant to measure):

When and how often is this test administered?

Construct Coherence

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
1. What are you intending to measure with this test?			□ Adequate □ Incomplete □ Lacking
2. How was the assessment developed to measure the measurement targets?			□ Adequate □ Incomplete □ Lacking
3. How were items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement target(s) and not other content, skills, or irrelevant student characteristics?			□ Adequate □ Incomplete □ Lacking
4. How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?			□ Adequate □ Incomplete □ Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
5. How are scores for individual items combined to yield a total test score? What evidence supports the meaning of this total score in relation to the measurement target(s)?			□ Adequate □ Incomplete □ Lacking
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?			□ Adequate □ Incomplete □ Lacking
7. How are scores reported in relation to the measurement target(s)? Do the reports provide adequate guidance for interpreting and using the scores?			□ Adequate □ Incomplete □ Lacking

Number of Adequate ratings: _____ X 2 =	
Number of Incomplete ratings: _____ X 1 =	
Number of Lacking ratings: _____ X 0 =	
Construct Coherence Total =	

Comparability and Reliability

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
1. How is the assessment designed to support comparability of scores across forms and formats?			□ Adequate □ Incomplete □ Lacking
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?			□ Adequate □ Incomplete □ Lacking
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?			□ Adequate □ Incomplete □ Lacking
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?			□ Adequate □ Incomplete □ Lacking
5. To what extent are different groups of students who take a test in different sites or at different times comparable?			□ Adequate □ Incomplete □ Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
6. How are scores reported in ways that support appropriate interpretations about comparability and disrupt inappropriate comparability interpretations?			□ Adequate □ Incomplete □ Lacking
7. What evidence supports the appropriate use of the scores involving comparisons across students, sites, forms, formats, and time?			□ Adequate □ Incomplete □ Lacking

Number of Adequate ratings: _____ X 2 =	
Number of Incomplete ratings: _____ X 1 =	
Number of Lacking ratings: _____ X 0 =	
Comparability & Reliability Total =	

Fairness and Accessibility

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
1. How were the needs of all students addressed during assessment development? How were the assessment questions developed to ensure that scores reflect the intended measurement targets and not student characteristics or contexts that are irrelevant to the measurement targets?			□ Adequate □ Incomplete □ Lacking
2. How were the needs of students with disabilities addressed during assessment development?			□ Adequate □ Incomplete □ Lacking
3. How were the needs of English learners addressed during assessment development?			□ Adequate □ Incomplete □ Lacking
4. How are students with disabilities able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?			□ Adequate □ Incomplete □ Lacking

Question	**Summary of Evidence**	**Comments on Evidence**	**Adequacy of Evidence**
5. How are English learners able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?			☐ Adequate ☐ Incomplete ☐ Lacking
6. How are students' responses scored in ways that reflect only the construct-relevant aspects of those responses? What evidence supports the minimization of construct-irrelevant influences on students' responses?			☐ Adequate ☐ Incomplete ☐ Lacking
7. What evidence supports the interpretation and use of students' scores in relation to their learning opportunities?			☐ Adequate ☐ Incomplete ☐ Lacking

Number of Adequate ratings: _____ X 2 =

Number of Incomplete ratings: _____ X 1 =

Number of Lacking ratings: _____ X 0 =

Fairness & Accessibility Total =

Consequences and Use

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
1. Are the items and content of the test consistent with the standards being measured to ensure appropriate uses?			□ Adequate □ Incomplete □ Lacking
2. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of inappropriate uses by students, teachers, or administrators? What evidence supports the implementation and effectiveness of these efforts?			□ Adequate □ Incomplete □ Lacking
3. What evidence is available to support the use of test scores across the entire score scale and all performance levels?			□ Adequate □ Incomplete □ Lacking
4. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your state or local district? How well do these uses align?			□ Adequate □ Incomplete □ Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
5. If assessment scores are associated with recommendations for instruction or other interventions for individual students, groups of students, or the whole-class what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?			□ Adequate □ Incomplete □ Lacking
6. If assessment scores are associated with high stakes decisions for teachers, administrators, schools, or other entities or individuals, what evidence supports such interpretations and uses of these scores?			□ Adequate □ Incomplete □ Lacking
7. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?			□ Adequate □ Incomplete □ Lacking

Number of Adequate ratings: _____ X 2 =

Number of Incomplete ratings: _____ X 1 =

Number of Lacking ratings: _____ X 0 =

Consequences & Use Total =

Self-Evaluation Protocol, Step Four: Summary of Individual Assessment Reviews

Name of Assessment	Summary of Evidence												Action		
	Construct Coherence			Comparability and Reliability			Fairness & Accessibility			Consequences & Use			Drop	Revisit	Keep as is
	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14			
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□
	□	□	□	□	□	□	□	□	□	□	□	□	□	□	□

Glossary

Comparability/ comparable scores
Scores from two or more tests that might reasonably be compared, or used interchangeably, because the tests have been shown to measure similar content and skills with about the same level of accuracy.

Construct
The psychological trait or characteristic that an assessment tool has been designed to measure. Examples include achievement, cognitive ability, and interests.

Construct-irrelevant
Situations in which the scores of test takers are influenced, positively or negatively, by factors that are different from those the test is intended to measure. For example, when the reading requirements for a science test interfere with the ability of some students to respond, reading comprehension is considered an irrelevant construct that diminishes the meaning of the science scores obtained.

Construct-relevant evidence
Information gathered to show that a score on a certain test is a measure of the construct intended by the developer or is not a measure of some competing construct.

Measurement targets
Measurement targets are a set of knowledge, skill, and competency expectations derived from a set of standards that inform test and item development procedures and determine what the assessment scores are meant to reflect.

Opportunity to learn
The extent to which test takers have had an opportunity to learn and develop the tested constructs through their educational program and have had experience with the language or the majority culture required to understand the test.

Reliability
The characteristic of a set of test scores regarding the amount of random error from the measurement process that might be embedded in the scores. Scores that are highly reliable are accurate, reproducible, and consistent from one testing occasion to another. Reliability coefficients have values ranging between 0.00 (low reliability) to approaching 1.00 (highly reliable), are usually used to indicate the amount of error in the scores.

Validity
The degree to which evidence and theory supports the interpretations of test scores for proposed uses of tests.

Validity Evaluation
The process of gathering and evaluating evidence related to the interpretation and use of scores from a particular test.