



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.

All rights reserved. Any or all portions of this document may be reproduced and distributed without prior permission, provided the source is cited as: Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores Project (SCILLSS). (2017). Ensuring Rigor in Local Assessment Systems: A Self-Evaluation Protocol. Lincoln, NE: Nebraska Department of Education.

Table of Contents

Background	1
Why evaluate assessments?	1
What is this protocol designed to do?	2
Guidelines for Implementing the Self-Evaluation Protocol	2
Self-Evaluation Protocol, Step One: Articulate your current and planned needs for assessment scores and data	3
Self-Evaluation Protocol, Step Two: Identify all current and planned assessments.....	4
Self-Evaluation Protocol, Step Three: Gather and evaluate the evidence for each assessment.....	5
Evidence for Construct Coherence	5
Evidence for Comparability and Reliability	7
Evidence for Fairness and Accessibility.....	8
Evidence Related to Consequences and Use	10
Self-Evaluation Protocol, Step Four: Review the evidence across assessments.....	12
Construct Coherence	16
Comparability and Reliability	18
Fairness and Accessibility.....	20
Consequences and Use	22
Self-Evaluation Protocol, Step Four: Summary of Individual Assessment Reviews.....	24
Glossary.....	25

List of Exhibits

Exhibit 1. Assessment Uses and Associated Stakes	3
Exhibit 2. Evidence for Construct Coherence.....	6
Exhibit 3. Evidence for Comparability and Reliability	7
Exhibit 4. Evidence for Fairness and Accessibility.....	9
Exhibit 5. Evidence Related to Consequences and Use	10

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 modules, or 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 these modules are not a toolkit for developing assessments, they offer 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 first of the five modules is available on the SCILLSS website at <http://www.scillsspartners.org/scillss-resources/>. The subsequent modules will be developed and available for use by December 2018.

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:
<ul style="list-style-type: none"> • guide next steps in instruction • evaluate instruction • evaluate curriculum <p>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.</p>	<ul style="list-style-type: none"> • evaluate learning for calculating grades • determine eligibility for program entry or exit • diagnose learning difficulties <p>These uses have high stakes for individual students and scores must always be considered in combination with other information.</p>	<ul style="list-style-type: none"> • evaluate teachers • evaluate schools or districts • evaluate programs or services <p>These uses have high stakes for educators and scores must always be considered in combination with other information.</p>



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 provided on page 14 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. In addition, a direct link between each type of evidence and the *Standards* is supplied in Appendix A. 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. How clear are the definitions of the measurement target(s) ? How does/do this/these measurement target(s) align with your intended measurement target(s) for the content area and grade level?	The 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 target(s)? What evidence do the developers provide to support the quality of their development processes and their implementation?	The test publisher has documented the development process and it includes multiple reviews from experienced test development professionals who review every item before it is used.
3. How are 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? What evidence supports the quality of these reviews and the use of the feedback they provide to improve item quality?	The test publisher has documented a rigorous field test process; all test items are screened to ensure that they are appropriate for the test form.
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?	The test publisher provides a scoring report that documents the procedures used for scoring the assessment, and provides scorer training materials as appropriate, including rubrics, annotated student exemplars, etc.; the 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.
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?	An independent alignment study was completed and it demonstrates that the items address the intended measurement targets.
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 provides multiple score user guides, including guides for students, parents, and teachers.

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 yield accurate, reliable scores? What evidence supports score reliability?	The test has sufficient content coverage and items to ensure the reliability of the test forms. The test publisher provides documented evidence of the estimated reliability of test scores, including estimated error across the entire score range.
2. How is the assessment designed to support comparability of scores across forms? What evidence supports such comparability?	The test publisher produces technical reports that document the procedures for equating test forms and these procedures have been independently verified.
3. How is the assessment designed to support comparability of scores across time? What evidence supports such comparability?	
4. How is the assessment designed to support comparability of scores across administration sites? What evidence supports such comparability?	The test publisher provides sufficient training materials in order to ensure the consistency of administration practices across all test sites.
5. How is the assessment administered to protect against various types of cheating so that the scores reflect students’ knowledge and skills and not inappropriate access to testing materials? What evidence supports the implementation of these safeguards and security protocols?	The technical manual provides guidance for appropriate administration of the assessment and security of test materials to ensure a fair and standardized test administration.

6. How is the assessment scored such that scores reflect students' knowledge and skills and not inaccuracies or inconsistencies in the scoring process? What evidence supports implementation of these scoring procedures?	The test publisher provides a scoring report that documents the procedures used for scoring the assessment, and provides scorer training materials as appropriate, including rubrics, annotated student exemplars, etc.
7. How are scores reported in ways that appropriately support or disrupt comparability in score interpretation across time, administration sites, or variations in student characteristics?	The 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 accessible and fair 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

<p style="text-align: center;">Fairness and Accessibility Guiding Questions</p>	<p style="text-align: center;">Examples of Evidence for Fairness and Accessibility</p>
<p>1. How were the assessment questions developed to ensure that scores do not reflect student characteristics that are irrelevant to the measurement targets? What evidence supports this freedom from bias?</p>	<p>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).</p>
<p>2. How were the needs of students with disabilities addressed during assessment development? What evidence supports these efforts and their effectiveness?</p>	<p>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.</p>
<p>3. How were the needs of English learners addressed during assessment development? What evidence supports these efforts and their effectiveness?</p>	<p>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.</p>
<p>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 identification and use of these accommodations at the time of testing?</p>	<p>The test publisher’s technical manual provides evidence of pilot studies and/or cognitive labs to ensure that students with disabilities can demonstrate what they know and can do when responding to the assessment items.</p>
<p>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 identification and use of these accommodations at the time of testing?</p>	<p>The test publisher provides an accommodations manual that specifies the allowable accommodations for students who are English learners.</p>
<p>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?</p>	<p>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.</p>
<p>7. How are assessment scores interpreted in relation to knowledge and skills that test takers have had an opportunity to learn or are preparing to learn? What evidence supports the interpretation of students’ scores in relation to their learning opportunities?</p>	<p>Documentation from the district describes opportunities for teachers to evaluate assessment scores in relation to the curriculum, instruction, and learning taking place in the classroom.</p>

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. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your school or agency? How well do these uses align?	The technical manual documents the purposes and uses of the assessment scores, and the district documents the actual purposes and uses of the assessment scores at the district level.
2. If your school or agency is using test scores for purposes other than those for which the test developers intended, what evidence supports those uses?	The district provides documentation showing adequate evidence in support of the actual uses of assessment results.
3. If assessment scores are associated with recommendations for instruction or other interventions for individual students, 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 technical manual documents the purposes and uses of the assessment scores; the test publisher and/or district provides a means for educators to understand and implement interventions for students; individual student reports provide timely and appropriate information about student performance to support intended uses.
4. If assessment scores are associated with recommendations for whole-class or group instruction, 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 technical manual documents the purposes and uses of the assessment scores; the test publisher and/or district provides a means for educators to understand and implement interventions for whole-class or groups of students; aggregated performance reports provide timely and appropriate information about student performance to support intended uses.

5. 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 technical manual provides evidence of the reliability and validity of assessment scores for their intended uses.
6. 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 has released test score user guides, and has documented that the user guides have been reviewed by students, parents, and teachers.
7. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of cheating by students or others associated with the assessment or its stakes? What evidence supports the implementation and effectiveness of these efforts?	The technical manual provides guidance for appropriate administration of the assessment and security of test materials to ensure a fair and standardized test administration.

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, Step Three: Gather and Evaluate the Evidence for Each Assessment

Name of Assessment:
Who takes this test?

Key Validity Area	Score	Low (0-6)	Moderate (7-10)	Strong (11-14)
Construct Coherence: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparability & Reliability: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Fairness & Accessibility: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Consequences & Use: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

How are scores used?

Low stakes for educators and students	High stakes for students	High stakes for educators
To guide next steps in instruction <input type="checkbox"/>	To evaluate learning for calculating grades <input type="checkbox"/>	To evaluate teachers <input type="checkbox"/>
To evaluate instruction <input type="checkbox"/>	To determine eligibility for program entry or exit <input type="checkbox"/>	To evaluate schools or districts <input type="checkbox"/>
To evaluate curriculum <input type="checkbox"/>	To diagnose learning difficulties <input type="checkbox"/>	To evaluate programs or services <input type="checkbox"/>
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
<p>1. How clear are the definitions of the measurement target(s)? How does/do this/these measurement target(s) align with your intended measurement target(s) for the content area and grade level?</p>			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
<p>2. How was the assessment developed to measure the measurement target(s)? What evidence do the developers provide to support the quality of their development processes and their implementation?</p>			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
<p>3. How are 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? What evidence supports the quality of these reviews and the use of the feedback they provide to improve item quality?</p>			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
<p>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?</p>			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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)?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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 yield accurate, reliable scores? What evidence supports score reliability?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
2. How is the assessment designed to support comparability of scores across forms? What evidence supports such comparability?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
3. How is the assessment designed to support comparability of scores across time? What evidence supports such comparability?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
4. How is the assessment designed to support comparability of scores across administration sites? What evidence supports such comparability?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
5. How is the assessment administered to protect against various types of cheating so that the scores reflect students' knowledge and skills and not inappropriate access to testing materials? What evidence supports the implementation of these safeguards?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
6. How is the assessment scored such that scores reflect students' knowledge and skills and not inaccuracies or inconsistencies in the scoring process? What evidence supports implementation of these scoring procedures?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
7. How are scores reported in ways that appropriately support or disrupt comparability in score interpretation across time, administration sites, or variations in student characteristics?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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 assessment questions developed to ensure that scores do not reflect student characteristics that are irrelevant to the measurement targets? What evidence supports this freedom from bias?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
2. How were the needs of students with disabilities addressed during assessment development? What evidence supports these efforts and their effectiveness?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
3. How were the needs of English learners addressed during assessment development? What evidence supports these efforts and their effectiveness?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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 identification and use of these accommodations at the time of testing?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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 identification and use of these accommodations at the time of testing?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
7. How are assessment scores interpreted in relation to knowledge and skills that test takers have had an opportunity to learn or are preparing to learn? What evidence supports the interpretation of students' scores in relation to their learning opportunities?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your school or agency? How well do these uses align?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
2. If your school or agency is using test scores for purposes other than those for which the test developers intended, what evidence supports those uses?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
3. If assessment scores are associated with recommendations for instruction or other intervention for individual students, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
4. If assessment scores are associated with recommendations for whole-class or group instruction, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
5. 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
6. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
7. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of cheating by students or others associated with the assessment or its stakes? What evidence supports the implementation and effectiveness of these efforts?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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			
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



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.