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

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

Ensuring Rigor in State 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

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 evidence for each assessment 5
Evidence for Construct Coherence 6
Evidence for Comparability and Reliability 9
Evidence for Fairness and Accessibility 12
Evidence for Consequences and Use 16
Self-Evaluation Protocol, Step Four: Review the evidence across assessments 21
Self-Evaluation Protocol, Steps One and Two: Identifying Purposes and Assessments Used to Serve those Purposes 22
Self-Evaluation Protocol, Step Three: Gather and Evaluate Evidence for Each Assessment 23
Construct Coherence 24
Comparability and Reliability 26
Fairness and Accessibility 28
Consequences and Use 30
Self-Evaluation Protocol, Step Four: Summary of Individual Assessment Reviews 32
Glossary 33
References 34
Appendix A: Review of State Provided Supplementary Resources 35
Scenario #1 – State-supplied Interim Assessments 35
Evidence for Construct Coherence 35
Scenario #2 – State-supplied Item Bank 36
Evidence for Comparability and Reliability 36
Scenario #3 – State-supplied Pre-K Readiness Tool 38
Evidence for Fairness and Accessibility 38
Scenario #4 – State-supplied Professional Development Materials 39
Evidence for Consequences and Use 39

List of Exhibits

Exhibit 1. Assessment Uses and Associated Stakes 3
Exhibit 2. Evidence for Construct Coherence 7
Exhibit 3. Evidence for Comparability and Reliability 10
Exhibit 4. Evidence for Fairness and Accessibility 13
Exhibit 5. Evidence for Consequences and Use 17
Exhibit A1. Evidence for Construct Coherence 36
Exhibit A2. Evidence for Comparability and Reliability 37
Exhibit A3. Evidence for Fairness and Accessibility 38
Exhibit A4. Evidence for Consequences and Use 40

Background

Every US state uses one or more assessments of students' academic knowledge and skills for a variety of purposes. This self-evaluation protocol is designed to support state departments of education in evaluating each of these assessments as well as their overall assessment system. 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 state.

Why evaluate assessments?

An assessment system mandated by a state department of education should provide state administrators, as well as students, teachers, and school personnel with an accurate reflection of the key concepts, knowledge, and skills that students have achieved. Each assessment 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, interpretation of scores, 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* (herein the *Standards*; 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?

In many scenarios, an assessment may purport to serve multiple purposes, but in each scenario, it is essential to consider the data and evidence to support each purpose. In many scenarios, while an assessment may claim to serve multiple purposes, its initial design and development were focused on one, and only one purpose. Because the question about the purpose of assessments is so critical, this 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 state level to use in any evaluation of aspects of their state assessment system. It is designed to focus on assessments that are state-mandated assessments or on support programs that are supplied by the state. Scores from these assessments may be used as part of official accountability programs or as critical pieces in assigning grades or determining student promotion. All tests that yield scores used for any of these purposes are part of a state's assessment system and should be evaluated on a regular basis.

Educators at a state 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 state assessment directors, state accountability officials, test administrators, as well as 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. As noted above, 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 state.

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 mandated by your state. This information includes, but is not limited to, assessment purposes and uses, assessment technical manuals, assessment research conducted by the publisher/test vendor and/or by outside researchers, score reports and any interpretive guides, 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. In many states, the state is focused on one or a small number of assessments used within the state accountability system. However, in other states, there may be multiple assessments used, and they may be used for a variety of purposes. Regardless of the specific scenario, assessments are tools for producing information to help answer questions. As this protocol is completed, it is essential to consider what questions 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 need 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 instruction • evaluate curriculum	• evaluate learning for calculating grades • determine eligibility for program entry or exit • diagnose learning difficulties	• evaluate teachers • evaluate schools or districts • 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, test purpose, or another set of categories or dimensions to facilitate your review. It may be helpful, for example, to see assessments used in science across grade levels or the set of assessments used at a particular grade level across content areas. In addition to state-mandated assessment programs, your state may also provide other optional resources or supplemental materials to schools and districts (e.g., interim assessments, item banks). While these auxiliary materials will not necessarily require validity evidence as detailed as discussed here, it would still be appropriate to review and consider the evidence available to support the use of these materials. Appendix A of this document provides further guidance on the manner and type of evidence that would be appropriate for these types of resources.

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 with gaps where you do not have an assessment that could provide relevant information. As you populate the self-evaluation protocol for steps one and two, the results will highlight these overlaps and gaps. It should also be noted that the presence of either overlap or a gap does not necessarily mean a serious issue exists; 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 worksheet within this protocol titled, “Self-Evaluation Protocol, Steps One and Two” is intended to first guide you through the intended purposes, assessment uses, and associated stakes for your assessment program. In step two, you will want to consider the current or planned assessments that support the purposes, uses, and stakes that have been outlined. 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 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 design and implementation process, such as: 1) Test Design and Development, 2) Test Administration, 3) Test Scoring, 4) Test Analysis, 5) Test Score Reporting, and 6) Test Score Use. As you review each assessment, the data and evidence that is available to support the interpretation and use of the assessment scores for their intended purpose(s) can and should come from all six of these aspects of the assessment program. The data and evidence can also 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 consider each goal or objective of your assessment system, it is recommended that you consider 1) what specific data and/or evidence the test publisher has provided and how this data and evidence directly supports the interpretation of the test scores; 2) whether any entities other than the test publisher produced similar data or evidence that provide further support for the test score interpretation and use (typically found in published research articles or reviews of the tests by other entities); and 3) in the event that data and evidence are not available, whether there is a structured plan in place to evaluate this test score interpretation and use that will yield the required information.

As you review the evidence, you will reach a conclusion regarding whether the evidence available can be considered *Adequate*, *Incomplete*, or *Lacking*. Evidence that is considered *Adequate* provides sufficient data and information and supports a comprehensive framework that directly addresses the given test score interpretation and use. Adequate evidence also supports the interpretation across the full range of students that take the assessment. Evidence that would be considered *Incomplete* may provide some of the necessary data, but may be missing some critical information, such as the appropriate use of the test scores across the full range of students, and across all test score interpretations. Evidence that may be considered *Lacking* provides little or no evidence to support the intended test score interpretations.

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 test scores. Each of these key validity questions is supported by the *Standards* (AERA, APA, & NCME, 2014), 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 (in Exhibit 2, Exhibit 3, Exhibit 4, and Exhibit 5) are accompanied by evidence examples and are intended to support evaluation and the gathering of evidence for each test score interpretation and use.

Evidence for Construct Coherence

Key Validity Question(s): Does the assessment have evidence for construct coherence with your overall standards? 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. In addition to providing evidence that the appropriate construct is being assessed, test publishers should also provide evidence that test scores are not confounded by other irrelevant factors, such as knowledge about a particular sport that influences a candidate's ability to answer items that appear on a science test. 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). In addition, Standard 12.4 suggests "when a test is used as an indicator of achievement in an instructional domain or with respect to specified content standards, evidence of the extent to which the test samples the range of knowledge and elicits the processes reflected in the target domain should be provided. Both the tested and the target domains should be described in sufficient detail for their relationship to be evaluated. The analyses should make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent" (AERA, APA, & NCME, p. 196).

Construct coherence strengthens the validity of interpretations and uses of assessment scores and their intended purposes. Exhibit 2 provides guiding questions and sample evidence for consideration when evaluating whether an assessment captures the knowledge and skills defined in the target domain. This exhibit is not intended to provide an exhaustive list of evidence; persons implementing the protocol will want to consider the extent to which additional evidence is available and appropriate for evaluating construct coherence at each phase of assessment development and implementation.

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. The test publisher 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 evaluated to ensure they accurately	The test publisher provides a scoring report that documents the procedures used for scoring the items, and provides scorer training materials as appropriate, including rubrics and examples of responses for each level in the scoring rubrics.

capture and assign value to students' responses?	The state 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. The test publisher and state have engaged in steps to ensure that all students have had the opportunity to practice completing the test on the platform which it will be administered upon. The test publisher and state engage in steps to document how, when, and by whom the performance level descriptors were established and 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)?	An independent alignment study was completed and demonstrates that the items address the intended measurement targets. The alignment has also provided evidence that the test forms have appropriate coverage of the state standards. The study should describe the methodology used for the evaluation, 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 provides 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. The test publisher provides sufficient information to understand students' progress toward meeting the measurement targets.

Evidence for Comparability and Reliability

Key Validity Question: 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 includes reliability estimates, including documentation for how the estimates were determined and if the estimates are applicable across students that take the assessment? Assuming your test places students into performance categories, what evidence is available to document that the decision rules for placing students into performance categories were determined through a rigorous process that allowed for multiple parties to be involved and to help determine the rules?

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, and that assessment scores are comparable to other external indicators (classroom and district measures) of student achievement.

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. Exhibit 3 provides guiding questions and sample evidence for consideration when evaluating whether an assessment sufficiently supports the comparability of test scores. This exhibit is not intended to provide an exhaustive list of evidence; persons implementing the protocol will want to consider the extent to which additional evidence is available and appropriate for evaluating test comparability and reliability.

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 a test administration manual that outlines the standardized procedures and conditions for administration across all test sites. The administration guidelines should clearly specify whether students are allowed to use resources such as dictionaries, formula sheets, or calculators while they are testing. These constraints are necessary to ensure that all students take the tests under the same conditions. The manual should also clearly specify testing accommodations available to some students with disabilities and some English learners. The decisions about which accommodations students are and are not allowed to use while testing should depend in part on individual students' needs and also on what the test is meant to be measuring.
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. The test publisher provides evaluation information in the technical manual after every administration about the scoring processes and procedures as designed and implemented, including the qualifications of those scoring constructed-response items and the accuracy of algorithms when items are machine-scored, and any errors that occurred during scoring and how these were resolved.

4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?	The test publisher produces technical reports that document 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 times. The test publisher ensures these procedures have been independently verified by a third party.
5. To what extent are different groups of students who take a test in different sites or at different times comparable?	A state 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 state provides documentation to communicate changes or alterations to an assessment and its scores across years. The state 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 Question(s): 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 state and within your current curriculum?

Standard 3.2 indicates that tests should 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 or testing context that are irrelevant to the construct being measured. Construct-irrelevant characteristics are extraneous factors that distort the meaning of test scores, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. Universal design is an approach to assessment development that attempts to maximize the accessibility of a test for all its intended test takers.

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 for all students. Exhibit 4 provides guiding questions and sample evidence for consideration when evaluating fairness and accessibility. This exhibit is not intended to provide an exhaustive list of evidence; persons implementing the protocol will want to consider the extent to which additional evidence is available and appropriate for evaluating fairness and accessibility at each phase of assessment development and implementation.

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 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'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 with any necessary accommodations. The test publisher provides an accommodations manual that specifies the allowable accommodations for students with disabilities.
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's technical manual provides evidence of pilot studies and/or cognitive labs to ensure that English learners can demonstrate what they know and can do when responding to the assessment items with any necessary accommodations. The test publisher provides an accommodations manual that specifies the allowable accommodations for students who are 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. 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?	Documentation from the test publisher describes opportunities for teachers to evaluate assessment scores in relation to the curriculum, instruction, and learning taking place in the classroom.

	A state 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 for Consequences and Use

Key Validity Question(s): Does the use of the test scores lead to positive consequences and not negative unintended 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 state 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, both positive and unintended negative consequences, when developing assessments ensures the assessment yields information that can be and is used appropriately within a system to achieve specific goals and that the assessment outcomes contribute to improvements in teachers' capacity to provide academic instruction and support learning for all students. This is fundamental as "consequences are the first and most important consideration in establishing the validity of the assessment" (International Reading Association and the National Council of Teachers of English, 1994, p. 17).

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). Exhibit 5 provides guiding questions and sample evidence for consideration when evaluating the consequences associated with an assessment. This exhibit is not intended to provide an exhaustive list of evidence; persons implementing the protocol will want to consider the extent to which additional evidence is available and appropriate for evaluating consequences associated with the assessment.

Exhibit 5. Evidence for 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. The type of items included in the test and the content and skills coverage of the test are consistent with the expected knowledge, skills, and abilities of students.
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?	The technical manual provides guidance for appropriate administration of the assessment and security of test materials to ensure a fair and standardized test administration. A state 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. 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 state 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 state 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 produces a technical manual that documents the purposes and uses of the assessment scores; the test publisher and/or state provides a means for educators to understand and implement interventions for students; individual student reports provide timely, appropriate, and reliable information about student performance to support intended uses. The test publisher and/or state provides a means for educators to understand and implement interventions for whole-class or groups of students; aggregated performance reports provide timely, reliable, and appropriate information about student performance to support intended uses.

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 state develops a Theory of Action that describes how the scores are to be used along 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 state 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 state 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 state 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 worksheet is intended to capture the necessary details for determining the adequacy of the evidence for each assessment in an assessment system. The worksheets that follow can be completed with a team of people from your state and can support critical conversations within your state personnel. 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 (e.g., consequences and use, fairness and accessibility), you captured the total score at the top of the first page of the step 3 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). For each assessment, these scores should 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 state, it will be essential to view this evidence from a holistic perspective. Across the key validity components of the test, your review team can consider if the assessments adequately meet system goals and objectives and then determine subsequent actions to take regarding each assessment in your state.

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 and does not result in unintended negative consequences. Additional critical issues that can be considered are whether or not the data and evidence were collected in a manner consistent with the intended use of the test scores in your state. For example, was research conducted with samples of students that are consistent with your expected population and was the test administered using a similar model as your planned administration model?

If data and evidence are missing, one important consideration is whether or not the purpose and use should be considered essential or if it could be considered not as critical. 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. In some scenarios, it is not feasible to have all the evidence required as soon as a test is being introduced or being used in a new environment.

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 Evidence for Each Assessment

Name of Assessment:
Who takes this test (e.g., grade, all or particular groups of students)?

Key Validity Area	**Score**	**Low** (0-6)	**Moderate** (7-10)	**Strong** (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 (what 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
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

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
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 & Reliability			Accessibility & Fairness			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 target	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 reproducible and consistent from one testing occasion to another. Reliability coefficients have values ranging between 0.00 (low reliability) and 1.00 (highly reliable), are usually used to indicate the amount of error in the scores.
Validity	The degree to which evidence and theory support 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.

References

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.

International Reading Association and the National Council of Teachers of English. (1994). *Standards for the assessment of reading and writing*. Urbana, IL: Author.

Appendix A: Review of State Provided Supplementary Resources

In addition to mandating statewide assessments, many states offer a variety of resources designed to support schools and districts as they work to prepare all their students to show what they know and can do. These tools can be of a variety of forms, such as interim assessments designed to provide snapshots of student progress towards yearly goals, item banks designed to measure state standards, or other supplementary materials such as teacher professional development or teacher curriculum development support materials.

Most often, the stakes for these supplementary materials would not be considered as high as other statewide mandated assessments. Nonetheless, as the statewide entity providing access to these resources, it is imperative that some data and evidence be available to support the recommended uses of these materials.

This Appendix is designed as a high-level framework for the review and consideration of these materials. For each of the four validity questions described earlier in this protocol, one scenario is described, and some examples of the types of data and evidence that could be gathered are provided. ***It is important to keep in mind that these scenarios are provided as exemplars, and that all four validity questions should be considered for any auxiliary materials supplied by your state. This framework should be adjusted and tailored to the specific state offerings under review.***

Scenario #1 – State-supplied Interim Assessments

The state has supplied access to interim assessments designed to monitor student progress within a given instructional year. These assessments are not designed to be used within any accountability systems nor are they recommended for use in making decisions about student advancement to the next grade.

Evidence for Construct Coherence

As mentioned above, construct coherence is not the only validity question that should be considered when evaluating the validity of score interpretations and uses from interim assessments. Examples of the type of questions that could be considered when reviewing construct coherence are presented below, but all aspects of validity should be considered.

Key Validity Question: Do the interim assessments and their auxiliary materials provide evidence for construct coherence with your overall standards, curriculum, and statewide assessments?

In this scenario, construct coherence ensures that the interim assessments have been designed to yield data and information that is consistent with your standards, curriculum, and statewide assessment program. Any review of the assessments and their auxiliary materials "make explicit those aspects of the target domain that the test represents, as well as those aspects that the test fails to represent" (AERA, APA, & NCME, p. 196). Exhibit A1 provides examples of the type of evidence that would be expected to be observed with a system of interim assessments that are designed to provide snapshots of student progress throughout the school year.

Exhibit A1. Evidence for Construct Coherence

Construct Coherence Guiding Questions	Examples of Evidence for Construct Coherence
1. What are the measurement targets for this test?	The test publisher provides documentation that summarizes the alignment between the measurement targets on the assessment and the academic content standards targeted through classroom instruction and assessment as well as your statewide assessment.
2. How was the assessment developed to measure these measurement targets?	The test publisher has documented the development process and it includes reviews from experienced test development professionals. The test publisher has also provided evidence that supports the link between student performance on any interim assessments and other high stakes statewide assessments.
3. 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. The user guides provide support to help all parties fully understand what each interim assessment measures as well as enough information to allow for the identification of academic areas where students appear to have a strong grasp of the materials as well as areas where students require further instruction. The user guides explain how test scores can be used in relation to other curriculum materials and provides instructionally relevant feedback to aid educators in developing lesson plans for students.

Scenario #2 – State-supplied Item Bank

Within our hypothetical state, the education community determined that there was a need to assist teachers in the development of assessments that could be used to assess students' progression toward meeting the goals for each given subject area. It was also determined that teachers wished to have a system that allowed them an appropriate amount of autonomy to determine when and how to best assess their own students. As a result, the state has contracted with a test vendor who has supplied access to an item bank that can be used to create teacher-formulated assessments. The item bank is designed to align with the state standards and should contain appropriate coverage of all state standards at each given grade level. Teachers are able to select items and create test forms that can be administered to their students.

Evidence for Comparability and Reliability

When resources are provided that allow the assessments to be created within each school, one of the critical questions that must be addressed is whether different tests can be considered comparable and whether the results are sufficiently reliable for use by teachers.

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, has the test publisher provided guidance for how test comparability can be supported? For reliability, has the test publisher provided guidance for the appropriate procedures to ensure minimum levels of reliability?

In this scenario, it will be essential that the test publisher has provided sufficient guidance to support teachers in the development of assessments that result in reliable and comparable scores across time, forms, and administration sites. The publisher should provide information on topics such as the required length and content coverage of the test to ensure test scores meet minimal reliability requirements. In addition, teachers should be provided information to help them understand why comparability is critical, and what steps can be followed to help ensure comparability across test forms. As an example, Exhibit A2 provides examples of the type of data and evidence that should be provided if the state provides access to an item bank designed to help in the measurement of state and local standards.

Exhibit A2. 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.
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 has provided recommendations for appropriate administration models, the amount of time that should be provided for students to complete the assessments, and appropriate methods for collection and scoring of student performance.
3. To what extent are different groups of students who take a test in different sites or at different times comparable?	The test publisher provides guidance for how a district can establish policies and procedures for the development of assessments to be administered to students. For a district wanting to make score comparisons for groups (e.g., comparisons across schools or classrooms; comparisons of the same cohort of students across years; comparisons of different student groups such as English learners and non-English learners), the district should establish guidelines 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.

Scenario #3 – State-supplied Pre-K Readiness Tool

In scenario #3, the state has identified a need to better understand the status of children as they enter the school system. A better understanding of the key concepts and knowledge that a child has as they enter kindergarten will assist the school in developing activities for children as they enter kindergarten. The state has provided access to a pre-K readiness tool designed to evaluate how well-prepared children are to enter kindergarten. The tool is designed to provide a snapshot of the strengths and weaknesses of each child to help schools better prepare a curriculum that is appropriate for each child as they enter the school system.

Evidence for Fairness and Accessibility

In this scenario, the assessment of children at this young age presents some unique challenges. It is critical that the test publisher demonstrate and provide evidence to support the idea that these assessments are reliable, valid, and fair for children at this age.

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 have been presented with the opportunity to address the concepts being assessed? Has the test publisher provided information and guidance on how other users of the test scores, such as families, can use the information and score reports provided? To what extent are students able to demonstrate what they know and understand in your state and within your current curriculum?

Exhibit A3 provides examples of the types of data and evidence that would be expected in the event that a state provides access to early childhood or pre-K readiness tools to help support schools and districts as they initially enroll students.

Exhibit A3. 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 has produced evidence that children at the target age are able to review and appropriately respond to all the items that are presented. The test publisher has demonstrated that the administration models (i.e., the test administrator/child ratio, the duration of the assessment, directions provided) are appropriate across all ranges, and adjusted based upon the specific age of the student.
2. How are students with disabilities and 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 has produced evidence that students with disabilities and English learners can demonstrate what they know and can do when responding to the assessment items with any necessary accommodations. The test publisher provides an accommodations manual that specifies the allowable accommodations for students.

3. What evidence supports the interpretation and use of students' scores in relation to their learning opportunities?	Score reports have been developed that are specifically targeted to families of young children, with simple and clear explanations of what the scores do and do not mean, along with sources of other additional information. Scores are reported in a manner that respects the expected differences in performance from children, even within the narrow age ranges where differences in performance would be expected.

Scenario #4 – State-supplied Professional Development Materials

In this scenario, a state has determined that it is essential to provide teacher professional development activities in their state. A set of auxiliary materials are provided to assist teachers in the development of classroom materials, at-home exercises for students, and classroom-based assessments. For all of these activities, the stated goal is aiding students as they progress toward meeting the goals for each given subject area. It would be expected that all materials should be directly tied to the state standards and curriculum and should aid teachers as they prepare their classroom activities and use assessments to monitor student progress throughout the school year.

Evidence for Consequences and Use

When supplying such auxiliary materials, it will be essential for the state to consider how supplying these materials will impact teachers and other educators in their state.

Key Validity Questions: Does the use of the test scores and other materials lead to positive consequences for your students, schools, and teachers? Has the use of the test scores and materials led to any unintended consequences that had a negative impact on your educational programs? To what extent does the test and other materials yield information that is used appropriately within a system to achieve specific goals?

Exhibit A4 provides examples that would be expected if a state were to provide these additional resources. When considering the consequences of providing these materials, it is essential that the outcomes of the use of these materials be considered, along with any unintended consequences that may also arise.

Exhibit A4. Evidence for Consequences and Use

Consequences and Use Guiding Questions	Examples of Evidence for Consequences and Use
1. Is the content of the materials consistent with the standards being measured to ensure appropriate uses?	The publisher of the material has produced evidence that demonstrates that the content of the materials is consistent with the state and district expectations for teachers and students. Teachers have reviewed the materials and confirmed that the materials are appropriate and useful for them in their work. The publisher has produced materials that support the focus of the materials and that the content coverage of the materials is consistent with the expectations of teachers and other educators. The materials do not omit critical areas of the standards or curriculum that could lead to the exclusion of content from teacher's coverage in the classroom.
2. How are the materials intended to be used as described by the publishers and how are they used by your state or local district? How well do these uses align?	The publisher has produced evidence that users of the materials can effectively translate the materials into meaningful changes in their classroom curriculum. The publisher has also produced evidence that the changes in curriculum and activities lead to meaningful increases in student performance or the effectiveness of teachers.

As with the evaluation of the statewide assessment program, it is recommended that you review all data and evidence available for any materials that are of interest. As you review the evidence, you will reach a conclusion regarding whether the evidence available can be considered *Adequate*, *Incomplete*, or *Lacking*. Evidence that is considered *Adequate* provides sufficient data and information that provides a comprehensive framework that directly addresses the test use and interpretation and supports the interpretation across the full range of students that take the assessment. Evidence that would be considered *Incomplete* may provide some of the necessary data, but may be missing some critical information, such as the use of the test scores across the full range of students, and across all test interpretations. Evidence that may be considered *Lacking* provides little or no evidence and does not provide sufficient data to support any of the intended test score interpretations.

If you believe it to be appropriate, the worksheets provided earlier in this protocol could also be repurposed for the review of these materials. However, whether the worksheets or the formal scoring that was provided are followed or not, it will be important to evaluate whether the data and evidence should be considered as providing Low, Moderate, or Strong support for the intended uses of the materials.