Comprehensive Assessment System: Rhode Island Criteria & Guidance

Rhode Island
Department of Elementary and Secondary Education
Office of Instruction, Assessment, & Curriculum
255 Westminster Street, 4th Floor
Providence, RI 02903
Phone: 401-222-4600
www.ride.ri.gov

This page intentionally left blank

Comprehensive Assessment System: Rhode Island Criteria and Guidance

Table of Contents

Comprehensive Assessment System: Rhode Island Criteria and Guidance

I. Background and Rationale

On January 7, 2010, The Rhode Island Board of Regents approved *Transforming Education in Rhode Island,* the strategic plan for 2010-2015. Deborah A. Gist, Commissioner of Education, guided the development of this strategic plan. With input from the Board of Regents, Rhode Island Department of Education (RIDE) staff, educators, parents, community members, civic leaders, and youth, five priorities were set. One of the five priorities, Establish World-Class Standards and Assessments, mirrors the expectations in the Basic Education Program (BEP). The BEP states that by 2015 all Local Education Agencies (LEAs) will have comprehensive curriculum, instruction, and assessment systems that are internationally benchmarked. Supporting this strategic objective is another objective: Monitor and support LEA implementation of comprehensive local assessment and reporting systems based on internationally benchmarked standards (WCS 3).

On July 1, 2010, the Basic Education Program (BEP) that was adopted by the Board of Regents went into effect. It details expectations for implementation of a comprehensive assessment system. An excerpt from Section G-13-3, Comprehensive Local Assessment and Reporting Systems, follows:

> *"Each LEA shall develop a Comprehensive Assessment System that includes measures of student performance for the purposes of formative, interim, and summative evaluations of all students in each core content area."*

A comprehensive assessment system is a coordinated plan for monitoring the academic achievement of students from Pre-Kindergarten through Grade 12. The goal of the comprehensive assessment system is to increase student learning by producing actionable data*, evaluate the effectiveness of programs, and ensure that all students are making progress toward achieving learning goals. Research has shown that data-informed decision-making on the part of educators leads to greater student achievement.[1] In addition, students benefit when they understand the criteria for success and receive regular, descriptive feedback on their progress toward their goals.[2] The statewide adoption of the Response to Intervention (RTI) Framework necessitates that educators be well-versed in how to collect and interpret student data. Though the BEP requires a comprehensive assessment plan in the core content areas, the best practices and expected assessment literacy addressed in this document are applicable to all content areas, grades, and groups of students.

When properly designed and implemented, a comprehensive assessment system provides multiple perspectives and sources of data to help educators understand the full range of student achievement. This information can be used to evaluate educational programs and practices and make informed decisions related to curriculum and instruction, professional development, and the allocation of resources to better meet students' needs. The data inform educators and families regarding student performance on state, LEA, school, and classroom assessments and their relationship to ongoing instructional practice. Various types of assessments are required because

* For the purpose of this document, *data* refers to information about or measures of student behavior, performance, or learning. For example, attendance rates, spelling quiz averages, NECAP scores, graduation rates, and grade point averages are all pieces of data.

they provide different types of information regarding performance. A comprehensive assessment system must be appropriate for the student population and address the assessment needs of all students, including students with disabilities, culturally and linguistically diverse students, and students in early childhood programs.

Defining a process for how assessments are used to make educational decisions is critical to ensure there is consistency of rigor and expectations across all buildings and levels within an LEA. LEAs should have a well-established and documented system with reliable assessments that shows how data are used to make timely decisions about when and how to provide additional support or extend student learning.

The following information must be documented for each assessment in the comprehensive assessment system:

1. The name of the assessment
2. The purpose and use of data
3. The type of assessment (e.g., formative, interim, summative)
4. The scoring procedures along with the expected turnaround time for providing feedback to students
5. The implementation schedule
6. The allowable accommodations and/or modifications for specific students.

The above information should be kept on file and used as evidence of the LEA's comprehensive assessment system work, a foundation for conversations about changes to the assessment system, and guidance for future decisions regarding the assessment system. LEAs can review their assessment system using the tools and guidance provided in this document.

The purpose of this document is to outline the elements and features of a comprehensive assessment system, primarily as they apply to the roles and responsibilities of LEA leadership. However, the definitions, habits of thinking, and tools contained in the guidance may also be of use to school-level administrators and teachers. It provides a framework that LEAs should use to take inventory of existing assessments so as to determine any possible redundancy or gaps. Ideally, this work should be completed by teams of LEA and school leaders as well as content and grade-level experts who have a solid understanding of what data are needed and which assessments are best suited to provide it. Special educators and teachers of English Learners should also contribute to this analysis.

In some cases, LEAs may find that a fairly comprehensive assessment system is already in place. In others, LEAs may find that existing assessments are being used inappropriately or that more assessments are being employed for a given purpose than are needed. Or, LEAs may find that additional assessments are needed. Thoroughly evaluating the assessment systems in place to ensure that they are comprehensive will enable LEAs to introduce more efficiency, rather than additional burdens. Furthermore, data produced by a comprehensive assessment system will serve definable and significant purposes that, taken together, will enhance the educational outcomes for all students.

There are numerous ways to categorize and label the variety of assessments that are used in Rhode Island schools. For the purposes of this document, assessments are described in terms of purpose (to inform instruction, to screen/identify, and to measure outcomes) and type (summative, formative, interim). Students with disabilities and English learners are not addressed specifically in

any one section of the document. This is because, in most cases, good assessment practices for general education students are good assessment practices for diverse learners. Information about modifications and accommodations is contained in the "Consider Quality: Validity, Reliability, & Fairness" section of this document.

Current Efforts

RIDE, in partnership with local educators, has a multi-pronged strategy for enhancing existing assessment infrastructure, increasing assessment literacy, and assisting with the development of comprehensive assessment systems across the state. The instructional management system (IMS), which will launch in 2012, will be a single sign-on, web-based platform that will house curriculum, instruction, and assessment material and data. Through the IMS, educators will be able to access reports and query data at the student, classroom, school, and LEA level. The IMS will support an interim assessment item bank and test-development engine, which LEAs may use to design, generate, and score interim assessments. Also in development is a series of online formative assessment modules, which will be housed on the IMS, to familiarize educators with general assessment literacy and concrete formative assessment strategies. In addition, professional development will be offered to leadership teams to increase capacity in understanding and using data.

II. Purposes of Assessment

Assessment has an important and varied role in public education. Assessments are used to inform parents about their children's progress and overall achievement. They are used by teachers to make decisions about instruction, assign grades, and determine eligibility for special services and program placement. They are used by evaluators to measure program and educator effectiveness. Assessments are used to track progress toward school and LEA goals set by the state in accordance with federal regulations.

When it comes to assessment of student learning, the *why* should precede the *how*. Often the emphasis on measuring student learning creates very real pressure to purchase and implement programs and assessments that may not accurately assess the content and skills that need measuring. This pressure is felt at all levels of education and underscores the need to make thoughtful assessment choices that are not often amenable to quick solutions.

The vast majority of assessments are used for one of three general purposes: *to inform and improve instruction, to screen/identify (for interventions), and to measure outcomes (as part of an accountability system, for school improvement planning, or for evaluation)*.

When assessments are used *to inform instruction*, the data typically remain internal to the classroom. They are used to provide specific and ongoing information on a student's progress, strengths, and weaknesses, which can be used by teachers to plan and/or differentiate daily instruction. This is most typically referred to as Formative Assessment. However, interim and summative assessments can also be used to impact instructional decision-making, though not in the "short cycle" timeline that characterizes formative assessments. Assessments such as unit tests and even state assessment data can be used to reflect on and inform future instructional decisions.

When assessments are used *to screen/identify*, the data typically remain internal to the school or LEA. Assessments that are used primarily to screen are administered to the total population of students and generally assess key skills that are indicators of students' larger skill set, rather than an in-depth

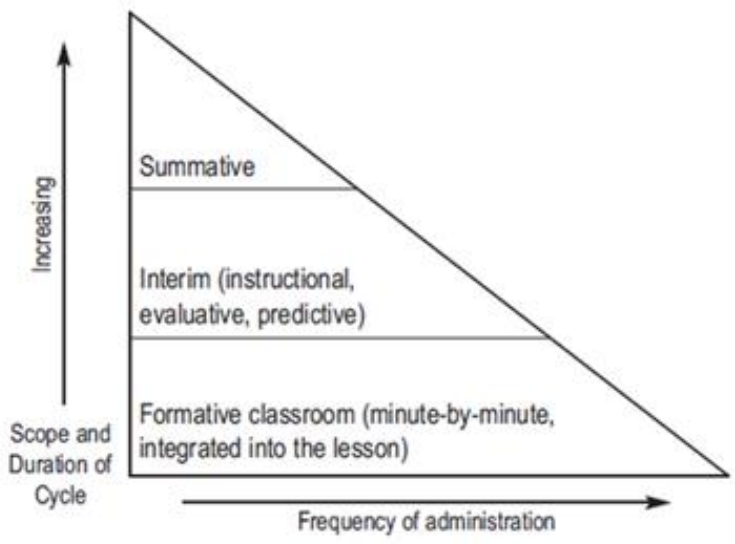
analysis of the standards. They should be relatively quick to administer and easy to score. Assessments used for screening purposes can inform decisions about the placement of groups of students within an academic program structure or individual students' needs for academic interventions or special programs. When needed, screening assessments are followed by diagnostic assessments to determine if more targeted intervention is necessary or if a student has a disability.

Finally, when assessments are used *to measure outcomes*, data are communicated to parties external to the classroom. Whether it is a unit test that is entered into a grade book and communicated to parents or a standardized test that is reported to the SEA, assessments used to measure outcomes attempt to measure what has been learned so that it can be quantified and reported. Some assessments that are used to measure outcomes may also be used to serve accountability requirements. These requirements are determined by state or federal regulations and corresponding state policy. In all cases, the particular type of assessment that is used is dependent on the claims that will be made about student learning, how the data will be used, and with whom it will be shared. No single type of assessment, and certainly no single assessment, can serve all purposes.

III. Types of Assessments

From informal questioning to final exams, there are countless ways teachers may determine what students know, understand, and are able to do. The instruction cycle generally follows a pattern of determining where students are with respect to the standards being taught before instruction begins, monitoring their progress as the instruction unfolds, and then determining what knowledge and skills are learned as a result of instruction. Assessments, based on when they are administered relative to instruction, can be categorized as formative, summative, or interim. Figure 1 and Table 1 illustrate how these types of assessments compare in terms of scope and use/purpose.

Figure 1. Tiers of Assessment



Source: Policy brief by Aspen/Achieve/Center for Assessment

Table 1: Intersections between Purposes and Types of Assessment

	Inform Instruction	**Screen/Identify**	**Measure Outcomes**
Summative	Generally not used as the primary source of data to inform instruction. May be useful in examining program effectiveness	Generally not used as the primary source of data to screen/identify students. May be one of multiple sources used	Primary purpose is to measure outcomes (at classroom, school, LEA, or state level). Can be used for accountability, school improvement planning, evaluation, and research.
Formative	Primary purpose is to inform instruction	Generally not used to screen/identify students	Generally not used to measure long term outcomes; rather, it is used to measure whether students learned what was just taught before moving on to instructional "next steps"
Interim	May be used to inform instruction	May be used to screen/identify students	May be used to measure outcomes in a longer instructional sequence (e.g., end of a unit of study or quarter, semester).

Summative Assessment: *Formal assessments that are given at the end of a unit, term, course, or academic year.*

These assessments are designed to judge the extent of student learning for the purpose of grading, certification, or evaluating the effectiveness of a curriculum. They are retrospective assessments of what students have learned, know, or are able to do. Given that common purpose, summative assessment items may take the form of anything from a persuasive essay to a geometry proof. Summative assessments typically have the most robust technical merit, allowing for more comparison and analysis of data, particularly on developing trends. These are the assessments most appropriately used to answer big questions such as "How are a group of students performing with respect to a body of standards or to their peers?" and "How well is the school/LEA/state serving its students?"

While a formative assessment might ascertain what students understand (and do not understand) at the end of a mini-lesson, a summative assessment measures what students can demonstrate they have learned at the end of a unit of study. Summative assessments typically have a less frequent cycle of administration than formative assessments and, as a

result, include more content. Because of the less frequent cycle of administration and more cumulative content, summative assessments are not typically used to inform instruction. Often, by the time assessments have been scored and results reported, the teacher has moved on to different material or group of students. The data produced are not particularly useful to teachers for creating student groupings or re-teaching material. However, it can be useful for informing *future* instruction. As teachers rethink the structure of a class after it has ended, they might review summative assessment data to determine what content or concepts were most challenging to students and, therefore, may warrant more time and attention next semester or next year. In some cases, summative assessments may signal whether a student should be more closely evaluated to determine if there is a need for additional supports or additional challenges.

Finally, summative assessment data can sometimes be used to evaluate the effectiveness of a particular program, curriculum, or instructional strategy. For example, if two similar elementary schools within an LEA are using two very different science curriculums (one project-based and the other more traditional), a common summative assessment might provide interesting data for comparing the effectiveness of the two programs, thus informing school-improvement planning. Additionally, summative assessments can be used for determining whether or not a student has met graduation requirements and for evidence of Student Learning Objectives* in the Rhode Island Model for Educator Evaluation.

Formative Assessment: *A process and/or a set of strategies that teachers and students use to gather information during (as opposed to after) the learning process and to make adjustments accordingly.*[3]

At the other end of the assessment spectrum is *formative assessment*. A teacher using formative assessment strategies knows where students need to end up and regularly checks in with students, using a wide variety of methods and strategies, to determine where they are in the learning process. Once the teacher clearly understands where students are, instruction is adjusted to accommodate the needs of the students in order to get them to where they need to be.

In contrast with summative assessment, formative assessments (such as quizzes, assignments, or quick verbal or non-verbal checks for understanding) are not used to grade in a formal sense. Rather, they are an exchange between the student and the teacher to determine what the student understands (and is therefore ready to move on from) and what may need to be re-taught or reinforced. A useful component of formative assessment may include teacher-student conferences and student reflections on their learning and skill development. Students must be actively involved in the formative assessment process, reflecting on their work and conducting self-evaluations of their learning. Students must be equal partners in the process in order to gain an awareness of where they are, where they are going, and what they need to keep moving forward toward their learning targets.

Formative assessment encompasses a variety of strategies. Teachers may require students to summarize the main idea of a story on an exit ticket before leaving class or to vote for which multiple choice selection they think is correct and defend their choice. They might give every student a whiteboard and require each one to solve a mathematics problem and hold up his or her work. Wiliam (2009) explains that formative assessment is

* Student Learning Objectives are long-term, measureable academic goals for students and are one measure of student learning in the Rhode Island Model for Educator Evaluation.

effective because it utilizes *pedagogies of engagement* and *pedagogies of contingency*.[4] By pedagogy of engagement, he means that effective formative assessment strategies require that 100% of the students in a classroom participate. All students must demonstrate their understanding (or lack thereof), thereby avoiding a scenario in which the teacher is directing most of his or her attention to those students whose hands are raised while neglecting those who choose not to participate. There is no opting out. By pedagogy of contingency, he means that formative assessment strategies require the teacher to adjust his or her instruction based on the data produced by these informal assessments.

For example, if a teacher administers a formative assessment and finds that all of the students are able to demonstrate understanding of a particular concept, he or she may adjust the lesson plan and move forward to match the pace of student learning. If the teacher finds that some students are able to demonstrate understanding while others are not, he or she may choose to create a small group for re-teaching or to create heterogeneous partnerships so that those students who can demonstrate competency can re-teach those who cannot. Or, in a third scenario, the teacher may find that few or no students are able to demonstrate understanding of a particular concept, in which case, he or she may decide to alter the next day's lesson plan in order to re-teach the concept in a different way or with greater detail. The key point is that formative assessment involves a short cycle of collecting data and using that data to keep instruction at pace with student needs and learning styles.

Shavelson (2006) describes three types of formative assessment: a) "on-the-fly," (b) planned-for-interaction, and (c) formal and embedded in curriculum. On-the-fly formative assessment occurs during "teachable moments" within the class. They are not planned for, yet they are an important opportunity to redirect misconceptions or flawed understanding.

During a planned-for-interaction assessment, a teacher may identify areas in the lesson plan to stop and assess understanding using response cards, one-sentence summaries, or purposeful questioning. This requires the teacher to plan questions ahead of time to be posed strategically throughout the lesson and the unit. The in-class questions as well as the delivery of the questions (using wait time to allow students appropriate time to think and respond) are key to advancing student learning.[5]

Finally, formal embedded-in-the-curriculum formative assessments may be administered every few lessons to determine student progress on sub-goals needed to meet the goals of the unit. For example, a teacher might administer a quiz that isn't factored into students' averages but is used to determine groupings or inform review for a summative assessment. These activities provide opportunities to teach to the students' areas of need.[6] In addition, formative assessment should provide opportunities for students to gain experience with and skills for self- and peer-evaluation. By setting clear learning targets and criteria for success, and providing multiple, low-stakes opportunities for assessment, teachers can help students become more independent, self-regulated learners.

Imagine a middle school writing class in which the teacher, unskilled in the strategies of formative assessment, is working to get her students to write informational essays with proficiency. She gives the assignment, which requires students to write an essay on an informational topic of their choice, sets a deadline, and provides work time in class and as homework. A few days or weeks later, the students

turn in their essays and the teacher grades and returns them. Work has been done, but has learning taken place? Have the students' writing skills been developed or just measured?

Now consider the same objective in the classroom of a teacher who has been trained in the formative assessment process. The teacher might begin with the same assignment. However, she also shows the students an exemplary essay, pointing out its features, and takes time to discuss what makes it a strong piece of work. Then, she has the class help create the rubric on which their essays will be scored. These activities clarify for students the *criteria for success*—what they need to incorporate in their writing in order to score highly. After writing their thesis statements and outlines, students are required to score each other's work and provide commentary on areas for improvement. During in-class writing time, the teacher conferences with students and asks them to assess their pieces against the rubric. After making careful observations to *identify gaps in learning*, she convenes strategy groups of students who are all struggling with the same concept, such as thesis sentences or paragraphing. This targeted intervention assists those who need it without slowing down those who don't. When rough drafts are submitted, the teacher provides *descriptive feedback*, which the students may use to revise their final draft. In the second scenario, students are required to be more engaged in and reflective about the writing process. The teacher assumes the role of a coach, assessing and guiding students during the writing process, not simply evaluating after the writing has been completed.

Formative assessment, in all forms, enables teachers to extract prior knowledge, identify concepts that students struggle with, and tailor instruction to meet the unique needs of a particular group of students. It enables students to strategically reflect upon their learning and become more aware of what they need to do to progress. Because it requires full participation of students and leads to more personalized, responsive teaching, formative assessment is a powerful tool for raising student achievement.

Interim Assessment: *Assessments administered during instruction that are designed to evaluate students' knowledge and skills relative to a specific set of goals to inform decisions in the classroom and beyond.*

As the name suggests, interim assessments fall between formative and summative assessments. They are typically administered every 6 to 8 weeks at the school or LEA level. Their purposes may include predicting a student's ability to succeed on a large-scale summative assessment, evaluating a particular educational program or pedagogy, or diagnosing gaps in students' learning.[7] As with any quality assessment, the specific interim assessment used is driven by the purpose and intended use of the data, but the results of an interim assessment must be reported in a manner that allows for aggregating across students, occasions, or concepts.[8] For example, LEAs might administer interim assessments to all the Algebra II classes in its high schools, half of which are using a new piece of technology, in order to evaluate the effectiveness of that tool. An LEA might administer interim reading assessments in order to identify or verify students for Personal Literacy Plans (PLPs). Or, in implementing RTI, an LEA may use interim assessments for progress monitoring, which may be administered at more frequent intervals, depending upon the intensity of the instruction or intervention. Many common assessments can be used as interims, including the Group Reading Assessment and Diagnostic Evaluation (GRADE) and the Developmental Reading Assessment (DRA), as well as quick tools, such as curriculum-based measurements.

Given their various purposes, interim assessments may be used both to inform instruction and to measure and document what has been learned.[9] Like formative assessments, interim assessments should inform classroom practice, though not with the same frequency and immediacy. Similarly, like summative assessments, interim assessments can be used for accountability purposes, though they don't typically carry the same high-stakes weighting. Interim assessments can be administered at the classroom level to track individual student progress. Common school or LEA interim assessments allow for comparisons across classrooms or schools. As a result, the line between interim and summative and interim and formative is not as distinct as the line between summative and formative.

In sum, each type of assessment has a role in a comprehensive assessment system. The goal is not to have "some" or "enough" of each type; rather it is to understand that each type of assessment has a purpose and, when used effectively, can provide important information to further student learning.

IV. Developing and Selecting Assessments

LEAs will not need to build a comprehensive assessment system from scratch. Rather, the process is one of revising the current system to make it more comprehensive and efficient. This involves identifying data needs, analyzing the quality of available assessments, and considering the capacity of the LEA to create, administer, score, and report on assessments. Once appropriate assessments are chosen, LEAs should document their comprehensive assessment systems and carefully review them for gaps and redundancies. Note that in the case of formative assessment, LEAs should identify the formative assessment practices that are widely used among their teachers. Documentation may include the formative assessment training that has been provided to teachers, the LEA's process for systematically implementing formative assessment strategies, and protocols for observing the use of formative assessment practices and sharing best practices/exemplars.

Consider Needs: Purpose, Alignment, and Form

Building or refining a comprehensive assessment system begins by agreeing upon the purposes of the assessments the LEA will administer. Decision-makers must first ask: "What claims do we want to make about student learning?", "What do we want to learn about students' skills and knowledge?" and "What data do we need?" Once claims and needs are identified, the appropriate assessments are selected to fulfill those data needs by asking: "Which assessment best serves our purpose?" Therefore, the LEA should not be concerned with having a sufficient number of each type of assessment but should select assessments that deliver the needed data for the intended purpose.

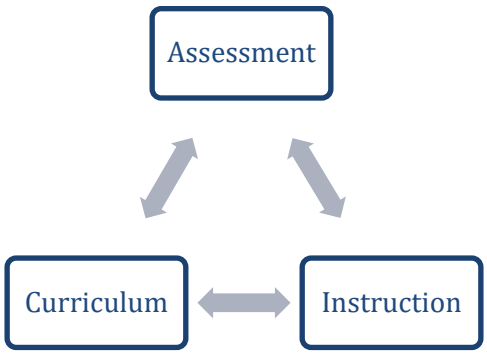
Consider Purpose First: Example 1

A 3rd grade teacher who wants to assess the reading skills and strategies a student uses for the purpose of informing instruction might administer a Record of Reading Behaviors.

Consider Purpose First: Example 2

A 7th grade mathematics teacher wants to know if any of his students may be in need of Tier 2 or Tier 3 interventions for mathematics computational skills. He administers the AIMSweb probes for computation and concepts and applications throughout the year.

In addition to considering what purpose an assessment will serve, attention must be paid to the alignment of the assessment with the curriculum and instruction within the school or LEA. The Board of Regents adopted the Common Core State Standards on July 1, 2010. As a result, LEAs will align their curriculum and assessments with these comprehensive standards for college and career readiness. Assessments that are not adequately aligned with what is taught are not accurate indicators of student learning. This is especially important when assessment data are used in high-stakes decision-making, such as student promotion, graduation, or educator evaluation. Because every assessment has its limitations and it is difficult to prove that any assessment is perfectly aligned with standards, it is preferable to use multiple measures when data are used in high-stakes decisions. By collecting a body of evidence, which hopefully indicates an overall conclusion, one can feel more confident in inferences drawn from such data. When curriculum, instruction, and assessment are carefully aligned and working together, student learning is maximized.



Finally, when developing or selecting assessments, knowing whether an assessment is a good fit for your needs requires a basic understanding of item types and assessment methods and their respective features, advantages, and disadvantages. Though this is certainly not an exhaustive list, a few of the most common item types and assessments methods are outlined here.

Selected Response – Selected response items present a question and a list of possible answers that a student can choose from. These can take the form of multiple choice questions, true/false questions, or matching questions. Selected response items often contain distractors, which are plausible incorrect answers intended to obscure the correct answer. They are generally used to assess recall knowledge and for questions for which it is easy to identify one correct answer. This item type can sometimes be used to assess higher-order thinking skills, though writing selected response items for this purpose is much more difficult.

Advantages: They allow for quick, inexpensive, and objective scoring. Because they usually take less time for students to complete, an assessment can contain a much higher number of these items than other item types, which increases the validity (see p. 14 for more information about validity) of inferences made on their basis.

Disadvantages: By definition, selected response items are fairly limited in form. Because the response options are provided, students' memories may be triggered, making it more difficult to accurately assess their knowledge and determine if they are able to generate authentic representations of learning.

Constructed Response – Constructed response items are open-ended questions that require students to produce a written response to a prompt or question. It may involve fill-in-the-blank, a short written paragraph, an extended response, working out a problem, or some other short, written activity.

Constructed response items are typically scored using a rubric or on a scale ranging from no credit to partial credit to full credit.

Advantages: Students must recall or produce a response without being prompted or reminded by options. Constructed response items are considered a more "authentic" assessment of certain skills, particularly writing.

Disadvantages: Constructed response items are more difficult to score because students can answer them in innumerable ways, usually necessitating human scoring. This makes scoring more time-consuming, expensive, and potentially open to subjectivity in the absence of strong scoring guides. Additionally, because these items usually take longer for students to complete, assessments usually contain fewer constructed response items, decreasing the validity of inferences made on their basis. Finally, because constructed response items typically require a written response, these items can conflate the skills being assessed. For example, a student's ability to express his understanding of the causes of the American Revolution may be limited by his ability to organize ideas in writing or express himself clearly in written English.

Selected response and constructed response items make up the majority of item types found on both locally developed and standardized assessments. On traditional assessments, either paper-and-pencil or computer-based, students answer the same core set of items (though they may appear in different forms) and their score is calculated based on the number of points earned out of the total number of possible points. On **computer-adaptive assessments** the items presented to a student are dependent upon his or her previous responses. For example, if a student consistently answers items correctly, the computer-adaptive program will select progressively more difficult items for that student. If the student answers incorrectly, the computer will select and present a less difficult item. The score is calculated automatically as the student completes the assessment. Computer-adaptive assessments might also contain a small number of constructed response items, which are either scored automatically by the computer or scored separately by human scorers and added into the overall score at a later time. In most cases, the overall score is calculated and ready to be reported by the time the student completes the assessment.

Performance Tasks – These are items or assessments that require students to apply their understanding to complete a demonstration, performance, or product that can be judged on clear performance criteria. For example, an essay might be considered a performance task if the skill being assessed is essay writing. However, an extended response on how to measure pH levels would not be a performance task if the skill being assessed is the ability to measure pH levels. In that case, having students use lab equipment to *actually measure* the pH levels of different substances may be considered a performance task. Strong performance tasks require students to apply and demonstrate their understanding, knowledge, skill, or ability. Performance tasks are often included as one type of assessment in portfolios and exhibitions, such as those used as part of Rhode Island's Proficiency Based Graduation Requirements. They could also be used as one type of evidence of progress or mastery for Student Learning Objectives, as part of the Rhode Island Model for Educator Evaluation.

Advantages: Because of their broad range of forms and contexts, performance tasks allow for richer, more "authentic" assessment of skills. In addition, depending upon the quality of the performance task, they can require higher-order

thinking and the application of multiple skills. Strong performance tasks require students to *apply* their understanding.

Disadvantages: Given their formats, forms, and contexts, performance tasks can be difficult and expensive to develop and score. They usually require human scorers. Ensuring consistency in the evaluation of performance tasks requires training of scorers. Performance tasks can be difficult to administer in a controlled and consistent manner. As they often require significantly more time than other item types, assessments usually only include one or a small number of performance tasks. This decreases the validity of the inferences made on their basis. Additionally, performance tasks can also conflate the skills being assessed. For example, a laboratory experiment designed to assess students' understanding of how energy is transferred may also assess students' ability to properly use laboratory equipment.

Observations/Interviews – This form of assessment includes actually watching students perform a task in order to determine if they are able to do it properly or having a formalized discussion with a student about the knowledge or skill being assessed. Observations and interviews are commonly used in early childhood education and as alternate assessments when students have difficulty expressing their knowledge or understanding on a written assessment.

Advantages: Observations and interviews are considered authentic assessments because they allow students to demonstrate their knowledge/understanding/skill firsthand and in a natural setting.

Disadvantages: This assessment method is very time-consuming and, therefore, can be very expensive to use and score. For this reason, it is often difficult to conduct more than a few of observations/interviews per student. This limits the validity of inferences drawn on their basis. In addition, observers and interviewers must be trained to know what to look for, how to avoid influencing the child during the assessment, and how to score consistently.

Consider Quality: Validity, Reliability, & Fairness

LEAs have discretion in deciding which assessments to use to meet their various needs. However, they should always seek to create or purchase assessments of high quality. Assessments of poor quality are of limited utility as the information they produce does not represent student learning well enough to properly inform decision-makers about the changes that are needed. There are three major indicators of assessment quality: validity, reliability, and fairness.

Validity refers to the accuracy of inferences drawn from an assessment, or the degree to which the assessment measures what it is supposed to measure. Valid interpretations provide an accurate picture of what students know, understand, and are able to do at different levels of application and understanding (i.e., cognitive complexity). How do you determine if the interpretation of a particular assessment is valid? Because validity is closely tied to the purpose or use of an assessment, the appropriate question is not "Is this assessment valid?" but "Is the interpretation of this assessment valid *for my purpose*?" For example, if a student's weight is 100 pounds, and the nurse's scale indicates that the student weighs 100 pounds, the scale has provided a valid assessment of the student's weight. However, it would not be valid to interpret this as an assessment of the student's height.

As described in the *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 1999)

below, the process of validation requires the collection of various sources of evidence:

> Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Validity is, therefore, the most fundamental consideration in developing and evaluating a test. The process of validation involves accumulating evidence to provide a sound scientific basis for the proposed score interpretations. It is the interpretations of test scores required by proposed uses that are evaluated, not the test itself. When the tests are used or interpreted in more than one way, each intended interpretation must be validated (p. 9).

It is also helpful to have a basic understanding of various types of validity, including *construct validity*, *concurrent validity*, and *predictive validity*.

Every assessment is designed to measure something. For interpretations of an assessment to have c*onstruct validity*, the assessment must actually measure what it is designed to measure and not contain features that would convolute interpretations. For example, a teacher finishes a unit on multi-digit multiplication and he wants to assess his students' understanding of said skill. He needs to administer an assessment that truly measures students' understanding of how to do multi-digit multiplication, not their understanding of multi-digit addition *or* their ability to memorize common multi-digit multiplication problems.

Construct validity depends not only on alignment to content but also on the level of cognitive demand. Assessments must ask students to engage in the content at different levels of understanding, depending on where they are in their learning. When students are learning a new concept or skill, an assessment should be of a sufficient cognitive demand to allow them to demonstrate where they are and then require them to apply those concepts at increasing levels of complexity.

There are many frameworks for measuring cognitive demand. This document refers to Webb's Depth of Knowledge Framework (2002), which outlines four levels of cognitive demand that are applicable to all content levels:

1. Level 1 is Recall and is characterized by simple retelling or recitation of facts or a procedure.
2. Level 2 is Skill/Concept and necessitates some type of decision-making. The response to a prompt will not be automatic and will require more than one step for the student to arrive at the answer.
3. Level 3 is Strategic Thinking. This is where reasoning becomes more complex and demanding. Tasks of this variety require greater planning, abstraction, evidence, and justification from the student. A student engaged in Level 3 is often required to form a hypothesis or conjecture.
4. Level 4 is Extended Thinking and manifests itself in tasks that require an extended period of time utilizing complex thinking and planning. Level 4 tasks compel students to make connections within a discipline and/or to other disciplines. More than likely, there are multiple solutions to a problem and multiple pathways for attaining a solution. Level 4 tasks are not typically found in large-scale assessments as they usually require multiple days of thought and consideration by the student. Students should be applying what they know to new situations to come up with complex answers and justifications.

It is important to note that Depth of Knowledge levels are not discrete but rather they are on a continuum. For this reason, it is important to

discuss test items and be familiar with DOK levels in order to ensure that students apply their skills and knowledge in the ways that encourage creativity, proficiency, and independence. Furthermore, DOK levels do not necessarily involve steps to solving a problem but rather how the students are being asked to apply their skills and knowledge. So while multi-digit multiplication involves more than one step, it is not necessarily a level 2 DOK because students are still applying a procedure.

Concurrent validity is an indicator of how well an assessment correlates with other assessments that measure the same skill/content. For example, a student who scored highly on the AP Biology exam is expected to also score highly on the SAT II Biology Subject Test. In the aforementioned mathematics teacher example, if the data from the multi-digit multiplication test were similar to the LEA interim assessment on multi-digit multiplication administered one week later, the teacher can assume that concurrent validity has been established.

On the other hand, consider a scenario in which an LEA has purchased a reading fluency intervention program and its accompanying assessments. That LEA needs to ensure that concurrent validity exists among program assessments by using multiple measures. If students who receive the intervention show increased scores on both the program-supplied assessment and on other measures of reading fluency, the LEA might infer that the program is effective for improving reading fluency and that interpretations based on program-supplied assessments are valid. However, if the students show improved scores on the program-supplied assessment but not on other measures of reading fluency, the program-supplied assessment might not be a valid measure of student reading fluency or the fluency intervention program might not be sound. This example underscores the importance of using multiple sources of data, when possible.

Predictive validity is an indicator of how accurately an assessment can predict performance on a future assessment. For example, college admissions officers use SAT scores to predict how a student will perform in college. If the mathematics teacher's multi-digit multiplication test data are highly correlated with students' scores on the end-of-the-year mathematics assessment, which is heavily based on multi-digit multiplication, it can be inferred that predictive validity has been established.

An assessment that is highly reliable is not necessarily valid. However, for an assessment to be valid, it must also be reliable.

Reliability refers to the consistency of an assessment. A reliable assessment provides a consistent picture of what students know, understand, and are able to do. For example, if the nurse's scale reports that a student weighs 100 pounds every time he steps on it, that scale provides a reliable assessment of the student's weight. If his true weight is104 pounds, however, the scale does not provide an accurate assessment of his weight.

Understanding reliability measures in large scale or purchased assessments and programs is important. It is also important to note that reliability measures will be available for the stated purpose of the test, not for any imagined or alternative purpose. This is another reason why it is important to use the programs and assessments for their stated purposes and be wary of alternative uses.

Determining reliability in teacher-developed assessments is a little more difficult given the small scale of the administration and the multiple purposes of assessments. It is useful to compare the results of a teacher-developed assessment with other assessment results. Did the students who are doing poorly on other assessments and classroom work pass this test? Did generally high-performing students do poorly on the test? If the test results indicate that struggling students are doing well, then the test is probably not reliable. This is one instance where gathering multiple sets of data is useful. It can help teachers evaluate the effectiveness of their own assessments.

How do you know if an assessment is reliable? A reliable assessment should yield similar results for the same student if administered more than once. All commercially available assessments should provide reliability information within their technical manuals. Reliability information can be reported in different ways, including, but not limited to, standard errors of measurement, confidence intervals, reliability coefficients, generalizability coefficients, and/or IRT-based (Item Response Theory) test-information curves.[10] Ideally, assessment should have high reliability and generalizability coeffiecients, low standard errors, and small confidence intervals. For IRT-based test-information curves, the test information (i.e., a numerical value indicating the precision of measurement) should be high at cut scores (e.g., between below proficient and proficient).

How reliable does an assessment need to be? The answer depends on the purpose of the assessment. When data are used to make high-stakes decisions (regarding student promotion, graduation, or educator evaluation, for example), they need to be highly reliable, in addition to being valid. Teachers, students, and parents need to feel confident that the assessments results are an honest representation of what students know and are able to do.

To understand how validity and reliability are linked, consider the target analogy. The center of the target is what you are trying to measure (student learning). Imagine that each dot on the target represents one measure of student learning. If the inferences based on that assessment are perfectly valid, the center of the target would be consistently hit, as in Figure 1. If the inferences are not valid, the dots would miss the center. If each of the dots hits the target at about the same spot, the assessment is reliable. However, as Figure 2 illustrates, a reliable assessment is not necessarily valid. The goal is to administer assessments that accurately reflect student learning (hitting the center of the target) and produce consistent data (dots are closely grouped). [11]

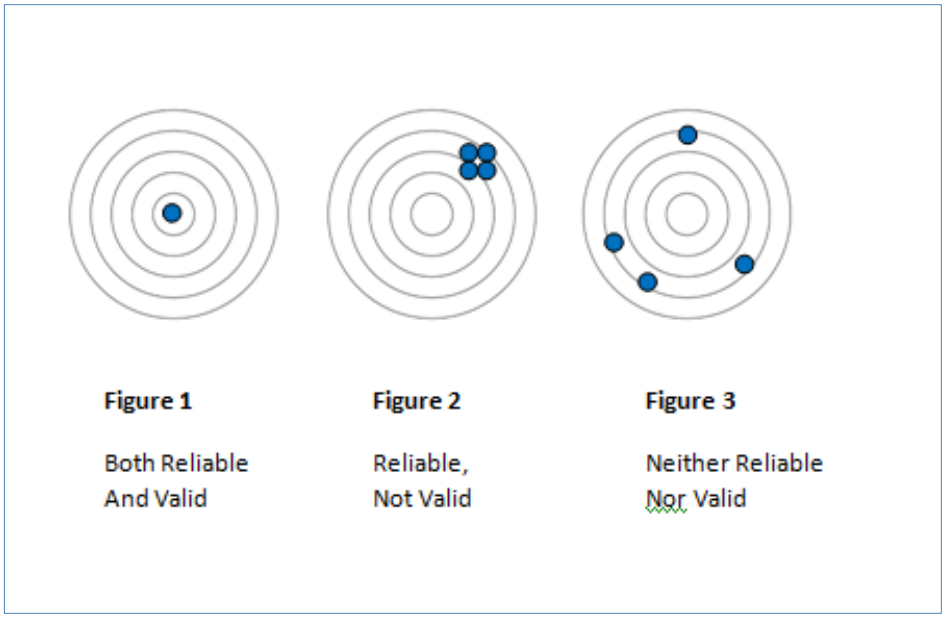


Figure 1

Both Reliable And Valid

Figure 2

Reliable, Not Valid

Figure 3

Neither Reliable Nor Valid

Fairness entails a lack of bias, the accessibility of the assessment, and the equity with which the assessment is administered.[12] A fair assessment is one that distinguishes only between students on the basis of the skills or content being measured. Therefore, on an unbiased assessment, a student will not have an advantage or disadvantage based on cultural background or gender. In designing an assessment, it is critical to not include terminology or examples that favor the background knowledge or experience of one group of students over another.

Accessibility refers to the ability of all students to participate in the assessment and may be ensured by offering assessments in various modalities (Braille, oral) and languages. If accessibility is not considered, an assessment of a non-native English speaker's content knowledge may be highly influenced by his or her language skills. Nonetheless, an assessment administered with accommodations must still measure the construct it is designed to measure. For example, it might be appropriate to provide a scribe to type a student's response on a final exam in American history, but it would not be appropriate to provide a scribe to type a student's final exam in typing.

Equity of test administration means that all students took the assessment under equitable conditions that were appropriate to produce the best working environment for the student (i.e., they were allowed the appropriate amount of time, they were provided with the materials they needed, they took the assessment under appropriate testing conditions). Ensuring equitable test administration may require the use of alternative ways to administer the test and/or the use of tools that enable students to engage in the test content.

The New England Common Assessment Program (NECAP) Accommodations Guide states:

Test accommodations are changes in setting, timing (including scheduling), presentation format, or response format that do not alter in any significant way what the test measures, or the comparability of the results. When used properly, appropriate test accommodations remove barriers to participation in the assessment and provide students with diverse learning needs an equitable opportunity to demonstrate their knowledge and skills (p.14).

Accommodations may include small group testing to reduce distractions, Braille or large-print materials, extended time, or access to a word processor. Conversely, assessment *modifications* may include focusing the assessment on some standards (versus all), reducing the complexity of a performance task (i.e., eliminating steps), or using alternative scoring rubrics. Accommodations are typically an adjustment in *how* a student is assessed while modifications are an adjustment in *what* is assessed. Modifications should be used only when available accommodations have been used and the assessment is still prohibitive.

The decision of what, if any, accommodations and/or modifications to use depends on the purpose of the assessment. For example, if the purpose is to screen/identify or to measure outcomes, the same assessment must be administered to all students in order to meaningfully compare the data. However, if the purpose is solely to inform instruction, a modification might be useful in order to assess a particular student's appropriate level of instruction.

Ensuring equity of administration also requires LEAs to consider the security and implementation schedule of their assessments. They must establish procedures for how teachers and other test administrators receive and return their materials, so as to standardize access to the materials and protect the comparability of results.

Some assessments, such as the Northwest Evaluation Association (NWEA), require careful planning to reserve computer lab space and produce alternate schedules. For LEA-wide assessments, common schedules should be articulated to ensure that differences in data reflect differences in student achievement, not differences in access to the test.

Similarly, an established procedure for moving from a screening assessment to a diagnostic or identification assessment should be in place within an LEA. Without one, schools may have dramatically different steps and timeframes for administering the assessments, therefore rendering the results less comparable across schools.

LEAs should make every effort to ensure that the assessments their students encounter are valid, reliable, and fair, particularly for high-stakes testing and decision-making. When common or highly validated assessments are not available, multiple measures must be used. For example, a teacher may not have a common assessment to measure a Student Learning Objective. In this case, the teacher should use more than one measure of student learning. By triangulating data sources, the teacher can determine if each measure is reporting the same or similar results, therefore allowing for more confidence in the validity of the inferences.

Formative assessments should also be held to high standards of validity, reliability, and fairness. They are not typically subjected to external validation but can be validated by multiple measures. Generally, however, the best way to ensure quality formative assessment is to provide comprehensive training to teachers in formative assessment strategies and techniques and conduct regular observations to ensure that they are utilizing them properly.

Table 2: Ensuring validity, reliability, fairness

Purpose:	**To inform instruction**	**To screen/diagnose/ascertain outcomes**
Validity	Ask questions based on taught curricula Ask questions in various modes (paper and pencil, orally, in groups) Allow students to demonstrate knowledge/skills in multiple ways Ask questions at varying Depth of Knowledge levels	Ensure alignment with standards Ensure a variety of Depth of Knowledge levels Ensure a variety of item types (multiple-choice, constructed response) Ensure accurate test delivery Ensure high correlation with outcome variable
Reliability	Ask the same question more than once (to a different student, to the same student at a later time) or in different ways Ask questions randomly/call on students who have not raised their hands	Review rubrics to ensure alignment and clarity Review internal consistency of assessment (Published in technical reports) Review scorer reliability, when necessary
Fairness	Provide multiple ways for students to demonstrate what they know Expect participation from 100% of students	Ensure equitable testing conditions (time, materials, tools, conditions) Provide appropriate accommodations Ensure items have been reviewed for potential bias (statistically and/or via bias committees)

Note: This is a sample of strategies, not an exhaustive list.

Consider Capacity: Administration, Scoring, & Reporting

The purpose and quality of an assessment are not the only considerations when building a comprehensive assessment system. An assessment might be perfectly suited to an LEA's purpose and of the highest quality and still not be an appropriate addition to the comprehensive assessment system. Decision-makers must also consider the professional development, funding, and personnel capacity necessary and available to appropriately administer, score, and interpret the results.

For example, in order to administer an assessment in a valid and reliable manner, appropriate procedures must be followed. Thus, LEAs should ask: "Do we have the technological capacity to properly administer this assessment?" and "What support will teachers need to use the data effectively?" This may include ongoing professional development to develop and administer assessments or to administer commercially developed assessments. Additionally, data that are reported in a manner that teachers cannot understand or interpret are ultimately not useful. LEAs, therefore, must provide assessment literacy professional development to teachers on how to interpret the score reports and act upon the data. Assessments that require computer administration or teacher scoring may necessitate additional training and will certainly require an investment of time and funding. These considerations are mentioned to promote discussion and careful thought, not to discourage the use of assessments that require significant time, resources, or training. Assessments should not be chosen on the basis that they are inexpensive, quick, and easy to administer, score, and report. However, an assessment that is not (or cannot be) used properly is probably not the best use of LEA resources or students' time.

V. Interpreting and Communicating Data

Administering a rich repertoire of appropriate assessments generates meaningful data for schools and LEAs—but it cannot be the end of the story. In order to truly have a comprehensive assessment system, LEAs need to close the loop by effectively *using* the data their assessments generate. To do so, it is critical that teachers, students, school administrators, parents, and LEA administrators have a level of assessment literacy that enables each group to communicate and understand the information disseminated from assessments commensurate with their roles and responsibilities. Each group must understand what the various types of scores mean and how to properly interpret them. They must understand what the data show and, just as important, what the data do not show. LEAs must also consider how they are converting data into actionable information and then communicating this information in a manner that makes it not only available, but also salient and accessible to a variety of stakeholders.

Interpreting Scores & Scales

In order to properly interpret assessment data produced by a comprehensive assessment system, it is necessary to have a basic understanding of common score types and scales. Knowing what these scores and scales are—and are not—will limit misunderstanding and misuse of assessment data.

A common source of confusion is the difference between criterion-referenced assessments and norm-referenced assessments. **Criterion-referenced assessments** measure a student's level of mastery on a set of criteria such as the Rhode Island state standards on the NECAP or WIDA standards on the ACCESS. **Norm-referenced assessments** compare a student's performance with the performance of a group. Percentile rank scores are used exclusively with norm-referenced

assessments. Raw and scaled scores are used for both norm-referenced and criterion-referenced assessments.

Raw scores are the most straightforward type of score. They typically represent the number of multiple-choice or short-answer items that a student answered correctly, plus any additional points earned on extended-response questions (if available). Raw scores are often converted to derived scores, which allow for easier comparison and interpretation. Common types of derived scores include scaled scores, percentile rankings, and grade-equivalent scores.

Scaled scores convert raw scores into scores on a different scale. For example, students' NECAP scores are converted from a raw score on the test (the number answered correctly out of the total number of items on the test) into a score on the 80-point NECAP scale. This allows for comparisons between the tests across years, subject areas, and grade levels.

Cut scores are those scores at which score categories are divided (e.g., the point at which Proficient scores are separated from Proficient with Distinction scores). Typically, cut scores are represented as scaled scores. For example, the NECAP cut score between Partially Proficient and Proficient is 40 for all tested grades and subjects.

Percentile rankings are generally easy to understand and communicate to various stakeholders such as parents and students. A percentile score is measured on a 100-point scale. A student's performance is typically measured in relation to a norm group—a sample of the intended audience of the assessment that represents the demographic composition of the larger population. Large-scale assessments use norm groups to control for slight variation from administration to administration. A percentile score represents the percentage of students scoring at or below the student's raw score. For example, a raw score of 120 that converts to a percentile ranking of 64 would indicate that 64% of students in that normative group scored equal to or less than 120.

Grade-equivalent scores are another type of derived score. They are most commonly used at the elementary level and are expressed in terms of the grade level and the month of the academic year. For example, a score of 3.6 would indicate the sixth month of grade 3. These scores are often misunderstood as the grade level work that a student is capable of completing. That is not an accurate interpretation of this type of score. Consider, for example, a fifth grade student who receives the following grade equivalent scores.

Mathematics	5.4
Reading	8.1

Many people misunderstand this data to mean that this student is reading at an 8th grade level. The score actually indicates that the student read the test as quickly as, and made as few errors as, an average 8th grader in his or her first month of school might have on the 5th grade test. It cannot be inferred that he or she can read 8th grade texts because he or she has not been tested on 8th grade material.

Stanine scores (short for standard nine) are based on a scale of 1 to 9. Typically, a stanine score of 1, 2, or 3 indicates below-average performance, a score of 4, 5, or 6 indicates average performance, and a score of 7, 8, or 9 indicates above-average performance, as compared with other students who took the test.

Normal Curve Equivalent (NCE) scores indicate where a student falls along a normal curve using a scale of 1-99. The benefits of using NCEs is that

under certain conditions (normally distributed populations, nationally representative norming groups) NCEs are based on an equal-interval scale and, therefore, can be averaged and used to compare student achievement from year to year. For example, in a normally distributed population, if a student made exactly one year of gains, his or her NCE score would remain the same and their NCE gain would be zero (though they *have* progressed). A student with a net loss in NCE score has made less progress on the skills/content assessed than the general population, while a student with a net gain in NCE score has made more. Caution should be taken when comparing NCE results from different assessments. If a student receives an NCE score of 40 on a reading test and an NCE score of 30 on a mathematics test, it does *not* necessarily mean that the student is doing better in reading than in mathematics. The scores represent different content areas that have been assessed in different ways and are therefore not comparable.

Standard scores (z-scores or t-scores) also allow for comparison between various assessments because they are "standardized" to the same numerical scale. The scores represent raw scores converted to standard scores, which indicate how far above or below the average (i.e., mean) an individual score falls when using a common scale such as a t-scale with a mean of 50 and standard deviation of 10.

Though the aforementioned score types are the most commonly reported by commercial assessments, this is certainly not an exhaustive list. The important take away from this section is that whenever educators use a type of score to make programmatic or instructional decisions, they should have a solid, common, and *accurate* understanding of what those scores represent and how they are intended to be used.

Another common confusion stems from interpreting data based on ordinal scales and interval scales. On an **ordinal scale**, numbers are ordered such that higher numbers represent higher values, but the intervals between the numbers on the scale are not necessarily equal. For example, consider the Fountas & Pinnell reading level scale, which identifies 26 reading levels labeled with letters of the alphabet. A student reading at a level E is certainly a stronger reader than one reading at a level C. However, we cannot accurately quantify the differential between these two readers because we cannot know that the difference between a level C text and a level D text is the same as the difference between a level D text and a level E text. Other examples of ordinal scales are ranks and percentile scores. Because the intervals between the numbers on an ordinal scale are not necessarily equal, it is inappropriate to calculate averages or subtract scores with ordinal scales. However, in practice this misuse of ordinal-scale data occurs often.

On an **equal-interval scale**, the difference between any two consecutive points on the scale is always the same, as on a thermometer (the difference between 14°and 15° is the same as the difference between 15° and 16°). This type of scale allows for more manipulation of data, such as subtracting scores to calculate differences and calculating averages. Common examples of interval scales include the NECAP and SAT. One limitation of this and ordinal-scale data is that these scales do not have a "true" zero point; rather zero points are arbitrarily set (0°F does not actually represent the absence of temperature). Therefore it is not possible to make statements about how many times higher one value or score is than another (it is not valid to say that 50°F is twice warm as 25°F). These types of comparisons can only be made using a **ratio scale**, such as the Kelvin scale of temperature, which are uncommon in educational testing. It is important to understand the type of score and scale being used

before attempting to calculate averages or otherwise manipulate or graph data. One way to do so is by considering the following:

Are the data simply ordered from highest to lowest, or do increases (or decreases) in the scale represent equal intervals? An affirmative answer to the former statement would indicate an ordinal scale, while an affirmative answer to the latter would indicate an interval scale.

A **vertical scale** is one that allows a student's score in one grade to be compared with his or her scaled score in another grade (provided the scores are in the same language and subject). In order to allow for this, the assessment contains spiraled content from the previous grade's assessment. The ACCESS test for English Learners is an example of an assessment that uses a vertical scale. It is important to note that the NECAP does *not* have a vertical scale. It may appear, for example, that the fourth grade scale (which ranges from 400-480) is a continuation of the third grade scale (which ranges from 300-380), but it is not. The grade level included as the first digit of the score is for informational and organizational purposes only. Therefore, it is *not* appropriate to calculate a growth score, for example, by subtracting a student's third grade NECAP score from their fourth grade NECAP score. However, growth scores can be calculated on assessments, like the NECAP, that are not vertically scaled using other methods like those used in the Rhode Island Growth Model[†].

Of course, only a portion of the assessments administered LEA-wide use these types of standardized scores and scales. LEAs should also consider what types of scores and scales are used on local assessments and other measures of student learning, such as grades. For example, does the LEA have a grading policy that requires the use of a common scale? Are grades allowed to be curved and, are therefore, norm-referenced? Are there guidelines available to direct teachers as to what distinguishes a B- from a C+? When using local assessments that do not have standardized scores and scales, it is important to think about and discuss issues such as what qualifies as proficient and what the cut scores are between letter grades. In addition, LEAs should examine the consistency of policies for allowable accommodations and modifications, as inconsistencies may limit the degree to which scores can be compared across classrooms and schools. These discussions lead to common understandings and, ultimately, more appropriate interpretation and use of assessment data.

Considerations for Non-Standardized Assessments:

- **What are the cut points between letter grades?**
- **Is there a common grading scale in the LEA?**
- **Is the common grading scale adhered to consistently?**
- **Is there a policy for accommodations and modifications?**
- **What is the cut score for proficiency?**

Understanding the Limitations of Data

Data-informed decision making has become a best practice among educators. Allowing data to guide the allocation of resources leads to a more strategic use of funds and more targeted interventions. However, while data provide a wealth of important

[†] The Rhode Island Growth Model is one measure of student learning in the Rhode Island Model for Educator Evaluation. For more information on the model, please visit https://ride.ri.gov/instruction-assessment/assessment/rhode-island-growth-model

information, it is critical that decision-makers are clear about its limitations.

State assessment results, likely to be many LEAs' largest data set, are very useful for providing descriptive information on students' performance and identifying general areas of improvement or need. For example, when the results signal an improvement, they can be used as one indicator that a new reading curriculum is having a positive effect. When the results signal a need, they can be used as part of the basis for a decision to reallocate a coach from one school to another. However, results on a single state assessment should not be used to make programmatic, curricular, or instructional decisions; rather a body of evidence should be used from various sources to mitigate some of the limitations of educational assessment. By triangulating data sources, educators either gain confidence in the interpretations of the data or have reason to question the significance of any one piece.

At its core, educational assessment is about making inferences concerning students' knowledge, skills, and accomplishments. Yet educational assessment is limited because data are never completely comprehensive or unequivocal. In fact, educational assessments represent just a sampling of items measuring all possible aspects of a construct, such as mathematical ability. Thus, it is inappropriate to conclude that a student is or is not proficient in regard to a mathematics standard based on their performance on only a very small number of test items measuring that standard, for example. Such conclusions are only warranted using a body of evidence from a comprehensive assessment system.

Furthermore, as in any assessment situation, there is error in educational assessment due to various sources relating to the task at hand, the rater/scorer, or the occasion. These may include the characteristics of assessment itself (i.e., task) such as ambiguous questions and confusing directions; rater characteristics such as inconsistent scoring or a weak adherence to the rubric; and student characteristics such as test anxiety, guessing, or mood on testing day.

Despite this inevitable uncertainty, we must interpret the data in order to reach accurate conclusions about students. This involves understanding what evidence the data provide. The same data can prove conclusive for some inferences about student performance, but barely suggestive for others. It is important to understand why certain data is being collected, and in turn, use this evidence to reach appropriate conclusions. Part of this process involves understanding the purpose that the assessment was designed to serve. Summative assessments are typically not designed to inform instruction. Formative assessments are not designed to measure outcomes for high-stakes decisions. LEA leadership must be clear about what data the assessment was designed to produce and ensure that they are using that data accordingly. When using assessments for a different purpose than that for which it was originally designed, it is important to validate the assessment for the new purpose.

Similarly, attention should be paid to the type of score that is being reported. Norm-referenced scores compare student performance against the performance of the norm group, not against the standards. This type of score might be very useful in some scenarios, but may not explicitly reveal a student's level of proficiency. Other types of scores do measure students' proficiency with specific standards or curricular domains. However, it is important to be aware of the number of items that are used to calculate any type of score. A low number of items might encourage the interpreter of the scores to be cautious as they likely do not represent the broad spectrum of the construct being measured, but rather a small sample.

Communicating Assessment Data

Assessment data needs to be analyzed and converted into usable, actionable information if it is to be used to inform decision making. In order to package the information in a way to maximize use, an LEA should consider the target audience, from teachers and administrators to parents, students, and community partners (such as after school tutoring programs). Different stakeholders may require different types of data in different formats (data briefs, score reports, report cards, etc.).

First, consider what is being reported. Perhaps parents are being excluded from the conversation because assessment data are not shared with them. Or, perhaps parents are being inundated with scores and reports that they do not understand and cannot interpret. It is the responsibility of the LEA to ensure that students and their families are receiving sufficient and clear communication about the assessment data that is collected and what it can and cannot tell them. LEAs should look critically at the reports that are distributed and reach out to parents to ask them if their needs are being met and if they understand what is being shared with them. If not, the LEA might consider hosting an information session about assessment data or simply including a "How to Read this Score Report" memo when the data are sent home.

Students, when old enough to properly understand, should be encouraged to look at their assessment data. If students understand the purposes for which they are being assessed, they may be more motivated to perform and more engaged in their learning. Educators and parents should help them to understand what the data say and what the limitations of that data are. The goal is to equip all parties with the available information to lead to the best questions, the richest discussions, and the most appropriate decisions.

VI. Suggested Next Steps

Establishing an assessment system that monitors the academic achievement of students from Pre-Kindergarten through Grade 12 and produces actionable information to inform the learning process will take time. Not only must it provide all of the necessary information, but it must be of high quality and function smoothly. Revisions will be needed as curricula change, student learning improves, or new data needs arise. Certainly, this process requires a significant investment of time, energy, and resources. However, investing in a comprehensive assessment system will promote efficiency and produce programs that are tailored to local needs and more effective for promoting student achievement.

The Steps of Evaluating an Assessment System

Step 1: Inventory the assessments used across the LEA, at all grades, for all purposes, and in all content areas. The Assessment Map documents (Appendix A) will help LEA teams gather information from across the LEA and present it in a format where it provides an overview of what assessments are being used for which purposes. These tools will highlight areas where LEAs are not collecting data where they should be, and areas where they are administering assessments that produce redundant data. This step may be organized by the LEA team in one of two ways: have the schools complete the inventory on their own and then aggregate the information at the LEA level or have the LEA team complete the table on behalf of the schools.

Step 2: The LEA assessment team discusses the populated maps to understand the number and purposes of the assessments being used. It is important to understand if the intent of the assessments and their application is understood across all of the schools using that assessment.

Step 3: Are the assessments being used for their intended purpose? To help LEAs more clearly understand the information better, below are key questions to ask of each other and the schools:

Purchased Assessments and Programs

1. Are the assessments listed being used for their intended purpose? For example: if a screening assessment is being used for progress monitoring, this may not be appropriate given the design of the test.
2. Are the assessments being used to the full extent possible? Why or why not? For example: many purchased programs have different types of assessments built into them that may or may not be useful for teachers. The vocabulary component of a reading assessment may not be as thorough as a different assessment or it may not serve a particular set of students adequately so an additional assessment may have been purchased or developed to augment or supplement that component.

LEA and Teacher-Developed Common Assessments (e.g., PBGR and common tasks)

1. Are these assessments being used at the appropriate curricular time during the school year?
2. Were assessments validated, benchmarked, and scored according to a standard protocol?
3. Are assessments being used by all teachers in the necessary grade/content area?
4. Do the assessments address the needs of students both at low and high levels of achievement?

Step 4: Now that the assessments have been identified and their purposes and uses are understood, it is important to ask questions about the number of assessments used in a given area. Are there too many or two few in any area? Reading is a clear example as there are many purchased assessments available that address the various components of reading (see Appendix C) as well as RTI models. When determining whether or not there are redundancies, it is helpful to consider the finer points of the assessment design.

Purchased Assessments

1. What grade levels do the assessments serve? If there are two screening assessments used, each at different grades, does the information generated by the results "match" or "complement" the results from the first assessment? In other words, if reading assessment 1 provides a benchmark of reading comprehension that involves retelling, does reading assessment 2's benchmark of reading comprehension also involve a type of retelling? In this way, results may be complementary across assessments because they are measuring a skill or concept in a similar way. It is important to note that differences in assessments from one grade to another are necessary because of the depth of the skill being measured. It is important to have an understanding of *why* and *how* each assessment measures the content and skill in question. This ensures that results are used appropriately and avoids improper inferences.
2. It is important to talk with teachers about these assessments and programs to understand why the assessments are or are not needed and what they find valuable about each component.

LEA and Teacher Developed Common Assessments

1. Is there a particular strand or domain in a content area that has too many assessments developed for it?

2. Are there assessments across the various strands and domains that stretch high achieving and low achieving students appropriately?
3. Teacher-developed assessments and common tasks have a unique place in educational assessment in that they can be complex, dynamic, and incorporate many instructional strategies that other assessments cannot. The creativity employed by teachers in developing tasks and common assessments is wide; do enough common tasks incorporate various ways students work: with technology, research, self-direction, etc.

Step 5: Outline changes and alterations that need to be made and develop a timeline.

Step 6: Repeat Step 1. Assessments and the systems that use them should be constantly evolving. LEA and school staff should be continuously improving their assessment literacy skills so they can evaluate and discuss new developments in assessment. This ensures that everyone has a stake in gathering data that improve instruction and student learning and that cutting-edge research and assessment designs are used well and appropriately. Assessments are tools, not ends in themselves, and better, more accurate tools provide better data from which to make decisions.

The BEP only requires a comprehensive assessment system for the core content areas. However, LEAs should extend this work across all content areas. Such careful reflection and analysis leads to improved quality of assessment by encouraging alignment to state or national content standards, raising expectations for the assessment literacy of all content educators, and providing consistency in expectations and language across the curricula.

The second tool—Considerations for Interim and Summative Assessments (Appendix B)—provides a set of prompts to guide LEA leadership as they determine whether or not an assessment is a good match for their purpose, is of high quality, and fits within the LEA's capacity for administration, scoring, and reporting. This tool can be used to determine the appropriateness of an assessment that the LEA has been using or an assessment under consideration. The tool can be applied to assessments developed at the LEA level and those that are purchased. In addition to these two general tools, which can be used for any content area, you will also find a comprehensive Reading Needs Assessment Worksheet (Appendix C). This worksheet determines what assessments are being used and documents the reading assessment system within the LEA.

The data culled from of these tools provides a fairly complete picture of the assessment system currently in place within the LEA. As a result, LEAs should begin asking questions. What additional assessments appear to be necessary? What, if any, assessments are redundant and unnecessary? A good practice for evaluating the need for adjustments and revisions to the comprehensive assessment system is to ask if the needs of the LEA, schools, teachers, parents, and students are being met.

At the LEA level, are sufficient data available to analyze the academic achievement of subgroups? Can the LEA identify gaps between populations of students? Do the data allow for the identification of trends over time?

At the school level, are data available to analyze the effectiveness of programs and curriculums? Can school leaders use data to get a picture of what is going on in particular classrooms? Can they use data to track at-risk students? The best way to determine

if a school's data needs are being met is to ask leaders, either in face-to-face meetings or in surveys.

Similarly, LEAs should inquire as to whether teachers' data needs are being met. At the classroom level, do they have assessments for producing the data they want? Do they know how to read and interpret the data? Do they have the knowledge of how to use the data? LEAs should think deeply about the capacity of its educators to properly utilize the data produced by the assessment system. After all, if the data cannot be properly interpreted and utilized, the system will not wield a significant impact on student achievement. A crucial final step, therefore, is determining the professional development needs that exist within the LEA. Some areas of need may include interpreting multiple pieces of data, translating data into instruction, and communicating data to students and parents.

Finally, LEAs should ask families and students if they are satisfied with the amount and quality of the data that are being collected and communicated by the LEA. Do they have questions that aren't being answered or needs that aren't being met?

Ensuring a comprehensive assessment system at the LEA level is not a simple process. It must be artfully pieced together through collaboration, reflection, discussion, and analysis. It cannot be dashed together, hired, or purchased. It is RIDE's belief that the tools and considerations in this guidance help facilitate that process. Carefully thinking about the assessment system as a whole will promote alignment between standards and assessments. It will reduce redundancies, inefficiencies, gaps in data, and misuse of assessments. The result will be a comprehensive assessment system that yields meaningful data for educators who are equipped to utilize it to promote student achievement. RIDE believes that taking the steps outlined in this guidance to create comprehensive assessment systems across the state will move Rhode Island closer to the goal of college and career readiness for every student.

Sources

[1] Trimble, S. (2003). NMSA research summary #20: What works to improve student achievement. Westerville, Ohio: National Middle School Association.

[2] Hattie, J. and Timperley, H. (2007). The power of feedback. *Review of Educational Research,* 77(1), 81-112. doi: 10.3102/003465430298487

[3] Popham, W.J. (2008) Transformative Assessment. Alexandria, VA: ASCD.

[4] Wiliam, D. (2009). *Assessment for Learning: Why, what, and how?* London: Institute of Education University of London.

[5] Rowe, M. B. (1974). Wait time and rewards as instructional variables, their influence on language, logic and fate control: Part one-wait-time. *Journal of Research in Science Teaching*, 11, 81-94. doi: 10.1002/tea.3660110202

[6] Shavelson, R. J. (2006). *On the integration of formative assessment in teaching and learning with implications for teacher education.* Paper prepared for the Stanford Education Assessment Laboratory and the University of Hawaii Curriculum Research and Development Group.

[7] Perie, M., Marion, S., Gong, B., & Wurtzel, J. (2007). The role of interim assessments in a comprehensive assessment system: A policy brief. Retrieved from http://www.achieve.org/files/TheRoleofInterimAssessments.pdf

[8] Perie, M., Marion, S., & Gong, B. (2007). A framework for considering interim assessments. Washington, D.C.: National Center for the Improvement of Educational Assessment.

[9] The Council of Chief State School Officers (2008). Interim assessment practices and avenues for state involvement TILSA SCASS interim assessment subcommittee.

[10] American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing.* Washington, D.C.: Author.

[11] Trochim, W. (2006). Research methods knowledge base: Reliability & validity. Retrieved from http://www.socialresearchmethods.net/kb/relandval.php

[12]National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, D.C.: National Academy Press.

[13] Edwards, P.A., Turner, J.D., & Mokhtari, K. (2008). Balancing the assessment *of* learning and *for* learning in support of student literacy achievement. *The Reading Teacher*, 61(8), 682-684. doi: 10.1598/RT.61.8.12

[14] Rhodes , L. K., & Shankin, N. L. (1993). *Windows into Literacy: Assessing learners K-8*. Heinemann: Portsmouth, NH.

[15] Torgesen, J.K. (2006). A comprehensive K-3 reading assessment plan: Guidance for school leaders. Portsmouth, NH. RMC Research Corporation, Center on Instruction.

[16] Webb, N.K. (2002) Depth-of-Knowledge Levels for Four Content Areas. Wisconsin Center for Educational Research.

[17] Olinghouse, N. G. (2009). Writing assessment for struggling learners. *Perspectives in Language and Literacy*, Summer, 15-18.

[18] NCTE Writing Assessment: A Position Statement (2009).

[19] NCTE Writing Assessment: A Position Statement (2009).

[20] Routman, R. (2005). *Writing Essentials: Raising Expectations and Results While Simplifying Teaching.* Portsmouth, NH: Heinemann.

[21] Strickland, K. & Strickland, R. (2000). Making Assessment Elementary. Portsmouth, NH: Heinemann.

[22] Newkirk, T. & Kent, R. (2007). *Teaching the Neglected "R": Rethinking Writing Instruction in the Secondary Classrooms.* Portsmouth, NH: Heinemann.

[23] National Council of Teachers of Mathematics. (1995). *Assessment standards for school mathematics.* Reston, VA: Author.

[24] American Federation of Teachers. (2008). Making classroom assessments work for you: Training Materials. Washington, DC: Author.

[25] Webb, N.& Collins, M. (2008). *Depth of knowledge: Descriptors, examples and question stems for increasing depth of knowledge in the classroom.* Kirksville, MO: Truman State University.

[26]Kentucky Department of Education (2007). Support materials for core content for assessment version 4.1: Mathematics.

[27] Bush, W.S. & Leinwand, S.(Eds.). (2000). *Mathematics assessment: A practical handbook for grades 6-8*. Reston, VA: National Council of Teachers of Mathematics.

[28] Sutton, J. & Krueger, A. (Eds.). (2002). *ED thoughts: What we know about mathematics teaching and learning.* Aurora, CO: Mid-continent Research for Education and Learning.

[29] Ibid.

[30] Bush, W.S. & Leinwand, S.(Eds.). (2000). *Mathematics assessment: A practical handbook for grades 6-8*. Reston, VA: National Council of Teachers of *Mathematics*.

[31] Lesh, R. & Lamon, S.J. (1992). *Assessment of authentic performance in school mathematics.* Washington, DC: American Association for the Advancement of Science.

[32] National Science Teachers Association. (2010). Science education assessments - NSTA's official position on assessment. Arlington, VA: Author.

[33] Keeley, P. (2008). *Science formative assessment: 75 practical strategies for linking assessment, instruction, and learning.* Thousand Oaks, CA: Corwin Press.

[34] Black, P. & Wiliam, D. (1998). Assessment and classroom learning. *Education: Principles, Policy and Practice*, 5, (1), 7-71.

[35] Porter, A. (1993). *Brief to policymakers: Opportunity to learn*. Center on Organization and Restructuring of Schools. Retrieved from http://www.wcer.wisc.edu/archive/cors/brief_to_principals/BRIEF_NO_7_FALL_1993.pdf

[36] National Council for the Social Studies. (2008). A vision of powerful teaching and learning in the social studies: Building effective citizens. *Social Education*, 72 (5), 277-280.

[37] Ibid.

[38] Alleman, J. & Brophy, J. (1999). The changing nature and purpose of assessment in the social studies classroom. *Social Education,* 65 (6), 334-337.

[39] Ibid.