Technical Guide A: Considerations Regarding District-Determined Measures

May 2013

Massachusetts Department of Elementary and Secondary Education
75 Pleasant Street, Malden, MA 02148-4906
Phone: 781-338-3000 TTY: N.E.T. Relay 800-439-2370
www.doe.mass.edu

Contents

Section 1. Introduction 1

Overview of This Guide 1

Background 2

Section 2. Core Assessment Concepts Related to DDMs 3

Assessment Terminology 3

“Assessments” Versus “Instruments” 3

Assessment Domains 3

Components of Instruments 4

Item Types 5

Defining Characteristics of Assessments 7

Direct and Indirect Measures of Student Learning 8

Assessment Types 8

Commercial Assessments 10

Standardization 11

Alignment of Content to the Curriculum 11

Introduction to Reliability and Validity 14

Reliability 14

Validity 17

Fairness 19

Bias 19

Documentation 20

Section 3. Selecting an Existing Assessment as a DDM 22

Identifying Coverage Areas for DDMs 22

Identifying Potential Existing Assessments 22

Evaluating Potential Existing Assessments 23

Step 1: Gather Documentation Regarding the Quality of the DDM (Reviews, Critiques). 24

Step 2: Use the Documentation to Evaluate the Assessment. 24

Finalizing a Decision Regarding DDMs 26

Monitoring the Assessment’s Use 27

Section 4. Developing a New Assessment 27

Determining What Needs to Be Measured 27

Selecting an Assessment Type and an Administration Method 27

Building an Initial Draft of the Assessment 29
- Create a Table of Test Specifications 29
- Select or Write Items 30
- Build a Scoring Key or Rubric 31
- Review Items for Bias and Sensitivity 32
- Construct Draft Versions of the Instrument and Administration Manual 32

Conducting a Pilot of the Draft Assessment 33
- Administer the Draft Assessment to a Representative Sample of Students 33
- Evaluate Items for Quality 34
- Evaluate the Scoring Method 34
- Evaluate Reliability 35
- Evaluate Validity 35
- Revise the Assessment and Construct the Final Version 35

Administering the Final Assessment 36
- Establish Local Policies 36
- Communicate Results 36
- Monitor Assessment's Use 36

Appendix A. Assessment Quality Checklist and Tracking Tool

Appendix B. DDM Resources

Section 1. Introduction

Overview of This Guide

In August 2012, the Massachusetts Department of Elementary and Secondary Education (ESE) published Part VII of the Massachusetts Model System for Educator Evaluation titled *Rating Educator Impact on Student Learning Using District-Determined Measures of Student Learning, Growth, and Achievement,*[1] which describes the Commonwealth's approach to incorporating measures of student growth into educator evaluation. The following document, *Technical Guide A,* is a supplement to Part VII and is intended to provide guidance to districts on how to evaluate and identify district-determined measures (DDMs). *Technical Guide B* will be released in the summer of 2013 and will provide more information on how districts can use data from DDMs to assess educator impact on student learning, growth, and achievement.

This technical guide is organized into four sections, with two supporting appendices:

- Section 1 offers an overview and background information about the use of measures of student learning, growth, and achievement in the Massachusetts Framework for Educator Evaluation.
- Section 2 provides explanations of key concepts in assessment and describes quality indicators of potential DDMs.
- Section 3 guides districts in identifying areas where DDMs may be required and presents considerations when selecting an existing assessment for use as a DDM.
- Section 4 shares information regarding the steps required for building a new measure if districts choose this course of action.
- Appendices:
  - A: Assessment Quality Checklist and Tracking Tool provides a mechanism for scoring and cataloguing assessments.
  - B: DDM Resources provides key resources for implementing DDMs.

ESE is available for consultation with districts to support implementation of DDMs. Districts may contact ESE via e-mail at EducatorEvaluation@doe.mass.edu.

[1] To access Part VII, see http://www.doe.mass.edu/edeval/model/

Background

On June 28, 2011, the ESE adopted new regulations to guide the evaluation of all licensed educators: superintendents, principals, other administrators, teachers, and specialized instructional support personnel. Under these regulations, every educator will receive two ratings: a summative performance rating and a rating of impact on student learning. The student impact rating will be based on trends and patterns in student learning, growth, and achievement using at least two years of data and at least two measures of student learning, each of which is comparable across the district for all educators in a grade or subject. These measures are referred to as **district-determined measures** (DDMs), and they are defined in the regulations[2] as follows:

> *. . . measures of student learning, growth, and achievement related to the Massachusetts Curriculum Frameworks, Massachusetts Vocational Technical Education Frameworks, or other relevant frameworks, that are comparable across grade or subject level district-wide. These measures may include, but shall not be limited to: portfolios, approved commercial assessments and district-developed pre and post unit and course assessments, and capstone projects.* (603 CMR 35.02)

Districts must be prepared to identify at least two measures for assessing educator impact on student learning, growth, and achievement for every educator who must be evaluated. The measures must cover all grade spans from Kindergarten through Grade 12 and all subject areas, including English language arts, family and consumer science and industrial arts, fine and performing arts, foreign languages, history and social studies, mathematics, physical education and health, science and technology, vocational and business education, and others.

Districts may consider measures of academic, social-emotional, and psychomotor learning. The Department encourages districts to assess the application of skills and concepts embedded in the new English language arts and mathematics curriculum frameworks that cut across subjects, such as writing text, nonfiction reading, reasoning and analysis, and public presentation. Districts are encouraged by the regulations and ESE to look beyond traditional standardized, year-end assessments to performance assessments and capstone projects scored against district rubrics and scoring guides, as well as interim and unit assessments with pre- and post- measures of learning.

Note that state regulations require that student growth percentile (SGP) scores on the Massachusetts Comprehensive Assessment System (MCAS) be used as at least one measure of student learning where available. Gain scores on the Assessing Comprehension and Communication in English State-to-State for English Language Learners (ACCESS for ELLs[3]) must be used where available. However, even for educators with a measure of growth available from a state assessment (MCAS SGP or ACCESS for ELLs gain score), districts may still need to identify at least one other measure to determine the educators' student impact rating.

[2] For the full text of the regulations, see http://www.doe.mass.edu/lawsregs/603cmr35.html.

[3] ACCESS has replaced the Massachusetts English Proficiency Assessment (MEPA) exams as our state's measure for the acquisition of English by ELLs.

Section 2. Core Assessment Concepts Related to DDMs

In this section, ESE defines important assessment concepts as they relate to DDMs. Through examples and other instructions, ESE identifies features of high-quality assessments. After reading this section, districts teams should have a basic understanding of key assessment concepts such as assessment domains, reliability, validity, and bias/fairness in testing. District teams with a strong background in assessment concepts may wish to scan this section and proceed to Section 3 for information on selecting an assessment for use as a DDM.

Assessment Terminology

“Assessments” Versus “Instruments”
Assessment is a general term that, as used in this guide, describes the systematic process of collecting, reviewing, and using information about students. The assessment process includes many steps:

- Identification of the purpose of the assessment and the target audience to be assessed
- Identification of the knowledge, skills, abilities, and/or behaviors to be measured
- Identification or development of a data collection tool (e.g., an instrument such as a test) that can be used to gather information about students
- Development of a method to score the instrument
- Administration and scoring of the instrument
- Development and application of policies associated with the reporting, interpretation, and use of the resulting scores
- Development of documentation to detail the assessment process

For the purposes of this guide, the term *assessment* will be used to refer to a complete process of measuring student growth in the cognitive (academic), affective/behavioral, and psychomotor domains. The term *instrument* will be used to refer specifically to the tool that collects information from the student, regardless of whether the tool being used is a test, survey, portfolio, performance task, observation, interview, paper, project, journal, discussion, or other type of data collection tool.

Assessment Domains
DDMs will be aligned with a range of educational roles in school districts. Most, but not all, of these roles will focus on learning associated with students’ cognitive (academic) gains (e.g., teachers in academic subjects). A small number of educator roles will be more closely aligned with student learning gains in affective/behavioral or psychomotor areas (e.g., teachers in nonacademic subjects and school counselors).

Although cognitive assessments, such as knowledge tests that use multiple-choice items, are common in K–12 education, the selection of an assessment for use as a DDM should be informed by the educator role.

The three learning domains are briefly described below:

- **Cognitive:** The cognitive domain is concerned with the acquisition of academic knowledge, content, and skills, such as the acquisition of math or history knowledge. These behaviors are often measured through knowledge tests or through an examination of student work products as in portfolios.
    - A cognitive assessment may be appropriate for measuring student growth for a history teacher or other educators teaching in an academic content area.
- **Affective/Behavioral:** The affective/behavioral domain is concerned with students' attitudes, interests, perceptions, and academic behaviors such as homework completion and attendance. Affective indicators are often elicited through attitudinal surveys or social inventories, such as a school climate survey. Student behavioral characteristics may be studied using measures like attendance, tardiness, or grade progression.
    - An affective assessment, such as an inventory that measures student attitudes toward school, may be an appropriate direct measure for school counselors.
- **Psychomotor:** The psychomotor domain is focused on students' abilities to perform small- or gross-motor (large-motor) tasks such as playing a musical instrument, creating a work of art (drawing, painting), or installing a window. Psychomotor behavior is typically measured by having the student perform one or more hands-on tasks or activities.
    - A performance assessment that includes a psychomotor component may be appropriate for measuring student growth in carpentry or music for teachers in those subject areas.

Components of Instruments

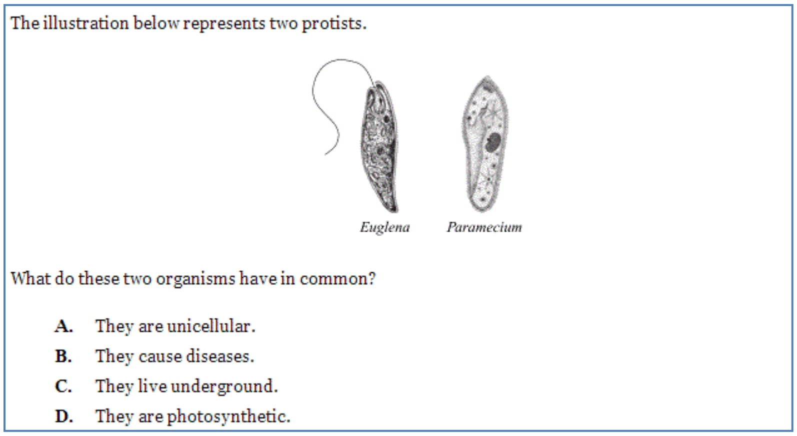
As described above, the assessment instrument is used to gather information about student learning. All instruments used as DDMs will include some common components, particularly student instructions, assessment items, and a method or methods for students to respond:

- *Instructions* provide sufficient directions to enable students to take the instrument (e.g., test). Instructions should be complete but concise and placed in a location that helps ensure students will read them.
- *Items* are statements, questions, or other prompts that elicit the knowledge, affect/behavior, or skills being assessed. Items are typically comprised of an *item stem,* which contains the statement, question, or other prompt to which students must respond, along with other stimulus materials, such as graphs, drawings, or narrative text.
- *Response methods* clearly indicate how examinees provide answers to the instrument. Response methods typically include directing examinees to (a) mark answers on the instrument itself, (b) use a bubble sheet to mark answers, (c) provide short or extended responses, or (d) complete a performance task in a public, small-group, or one-on-one situation.

The 8th grade MCAS science item presented below illustrates these components. The item stem asks examinees to identify what the two organisms pictured in Figure 1 have in common. The stem includes the drawings as a stimulus that also completes the item prompt. (In other words, examinees need both the stem and the stimulus to respond to the item.) The response method directs examinees to identify the answer choice via one of the answer options, A through D.

(Directions for selecting A through D are provided at the start of the item set in the test booklet.) Examinees are further directed to mark their correct response on the bubble sheet that is provided to them at the start of the testing session.)

Figure 1. Item Stem and Stimulus



The illustration below represents two protists.

What do these two organisms have in common?

A. They are unicellular.
B. They cause diseases.
C. They live underground.
D. They are photosynthetic.

Source: MCAS Item: Grade 8 Science and Technology Engineering Test, spring 2012
(Correct response is "A. They are unicellular.")

Item Types

While there are many item types, a simple distinction is whether students select or construct (create) a response. For *selected response* items, students choose the correct response to an item from among two or more options. Selection items include multiple-choice, true-false, rate/rank, matching, and paired comparison items, among others. Selection items are scored easily and objectively by using an answer key, which identifies the correct answer for each item. Objective scoring indicates that there is almost no error associated with the scoring (provided the correct answer is identified in the key).

For *constructed (or open) response* items, students create their own response rather than selecting one from a set of potential responses. Constructed response items are further differentiated by the length and parameters of the student's response. For example, constructed response items include the following:

- Short answer (requests a very short response, usually one word, a short phrase, or a number)
- Restricted constructed response (directs examinees to provide a fairly short response, usually a brief, targeted answer or explanation, or a directed math symbol, equation or scientific drawing is requested)
- Extended constructed response (requires examinees to provide a long, organized response, such as an essay, narrative, criticism, mathematical or scientific reasoning or proof, or other complex response)
- Performance item (directs examinees to demonstrate capability through a live performance or through the creation of a product, such as a completed lab experiment, research paper, art product, or performance piece)

- Portfolio item (guides examinees toward the development, vetting, and selection of portfolio pieces according to the criteria provided)
- Observational/interview item (provides directions for proctors to observe and catalog student behavior through a rating scale, checklist, or anecdotal records form). The observational method can be unobtrusive, in which the examinees are unaware that their behavior is being recorded or when the observer is sufficiently removed from the performance so as not to intrude, or it can require more participation from examinees, in which the proctor requests that examinees demonstrate the assessed behaviors.

Each item type has strengths and potential challenges. For example, selected response items take less time to administer, allowing examinees to take more of them during a shorter period of time (increasing content representation on an assessment). They also have more reliable scoring. However, constructed response items can assess student performance of complex constructed behaviors (such as the ability to research a topic or construct an organized response). A comparison of the advantages of three popular item types (selected response, constructed response, and performance/portfolio observational items) is presented in Table 1. Note that the table assumes the item types are well conceived, developed, and implemented.

Table 1. Comparative Advantages of Item Types

	Selected Response (Objective Items)	**Constructed Response (Subjective Items)**	**Performance/Portfolio/ Observational Items**
Sampling of Curriculum	Samples a lot of curriculum in a short period of time	Samples less curriculum than selected response items; takes longer examinee administration time	Samples less curriculum than selected response items; takes longer examinee administration time
Item Development	Requires the development of many items	Fewer items are needed	Fewer items are needed, but the items are written to break out the components of the task
Rigor	Can sample the range of Bloom's Revised Taxonomy from Remembering to Evaluating; takes skill to write items at the higher levels of rigor	Should be written for higher levels of rigor	Can range the levels of rigor, although some items should represent higher-level demands
Complexity	Low to moderate complexity	Can range from low to high complexity	Tasks should reflect moderate to high levels of complexity
Scoring	Objective scoring—efficient with a scoring key	Subjective scoring—requires the use of rubrics/scoring papers and scorer training	Subjective scoring—requires the use of rubrics; students can participate in scoring
Influence on Learning	Overuse of the selected response item format can encourage learner passivity; can encourage development of critical thinking skills when items align with higher levels of rigor	Good-quality constructed response items can encourage examinees to demonstrate creativity, organizational skills, topic development, critical thinking skills	Encourages the examinees to demonstrate what they know and can do. Depending on the item content, can encourage the development of critical thinking, organizational skills, and creativity

	Selected Response (Objective Items)	Constructed Response (Subjective Items)	Performance/Portfolio/ Observational Items
Reliability[4]	High internal consistency reliability is possible with the inclusion of 20+ high-quality items	Reliability is typically lower than with selected response items due to scorer differences, fewer number of items	Reliability is typically lower than with selected response items due to scorer differences, fewer number of items

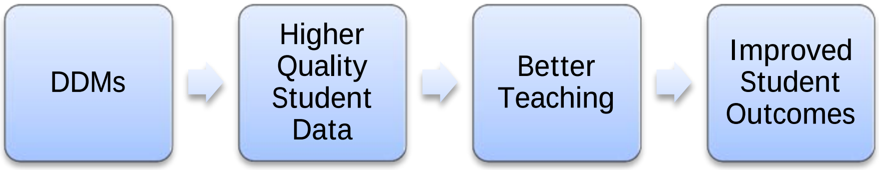
Adapted from: Linn & Gronlund (1995).

Other item qualities must also be considered when selecting items for DDMs. A high-quality assessment should be comprised of items that are not too difficult or too easy and should be populated with items that discriminate appropriately between strong and weak performers. For more information on the characteristics of high-quality items, see to Appendix B.

Defining Characteristics of Assessments

ESE encourages districts to employ assessment strategies that enhance instructional practice and promote high-quality learning in classrooms. High-quality assessment programs benefit students and educators alike by providing educators with robust student data they can use to improve their instructional practices, leading to better student outcomes (see Figure 2).

Figure 2. High-Quality Assessment Programs



The task of identifying assessments as DDMs provides districts the opportunity to scrutinize their assessment practices in order to answer important questions:

- How well is the district measuring student learning?
- What types of assessments is the district using and why?
- Are there other types of assessments the district could/should be using for certain educator roles?
- Does the district have systems in place to ensure that all students have a fair and equal opportunity to demonstrate what they know?
- Are the district’s assessments well aligned to its local curricula?

This section of the guide describes examples of defining characteristics of assessments that districts should consider as they identify potential DDMs.

[4] Reliability will be discussed in detail later in this section.

Direct and Indirect Measures of Student Learning

Direct measures of student learning assess student learning, growth, or achievement with respect to specific content represented by key standards and learning objectives identified by the district. Direct measures are strongly preferred for evaluation because they measure the most immediately relevant outcomes from the education process. Examples include MCAS scores and the Student Growth Percentiles (SGPs) in mathematics and English language arts that are based on those scores, other standardized assessments of student achievement in a subject, portfolios of student work, performance assessments, or survey results that represent key student behaviors that result from services delivered by educators who do not teach educational content.

Indirect measures of student learning, growth, or achievement provide information about students from means other than student work. These measures may include student record information (e.g., grades, attendance or tardiness records, or other data related to student growth or achievement such as high school graduation or college enrollment rates). To be considered for use as DDMs, a link (relationship) between indirect measures and student growth or achievement must be established. For some educators such as district administrators and guidance counselors, it may be appropriate to use one indirect measure of student learning along with other direct measures; ESE recommends that at least one of the measures used to determine each educator's student impact rating be a direct measure.

Technical Guide A describes the types and properties of assessments suitable for use as DDMs. *Technical Guide B* will describe the methods for identifying student growth, based on the DDM assessments. This distinction is important when considering the difference between measures and assessments to be used for DDMs. As defined earlier, instruments are part of the assessment process. They are the tools (e.g., the test, survey, portfolio, or other data collection tool) used to gather information from students. Results from the instruments can then be used to estimate student growth. These estimates of student learning, growth, or achievement are the measures. One readily available measure for growth is the SGP. In English language arts and math, SGPs are a measure of student growth derived from students' scores on the MCAS exams.

Assessment Types

Assessments administered at different times capture learning at different periods and intervals. End-of-year (EOY) and end-of-course (EOC) on-demand assessments may initially seem to be the most logical choice for DDMs, but they are not the only options. The regulations guiding districts' selection of DDMs are intentionally flexible to encourage districts to set their own criteria for selecting and developing assessments. This section described a variety of assessment types districts may wish to consider for use as DDMs.

EOY and EOC assessments are used to determine how well students have mastered the knowledge and skills included in content standards covered during a semester or a full school year. District-wide end-of-year exams typically cover a grade-specific subset of a content domain such as Grade 5 mathematics or Grade 10 English; end-of-course exams cover specific subjects such as world history or algebra I. (Advanced Placement exams are examples of EOC assessments.) The benefit of using EOY and EOC assessments is that they can evaluate a large subset of key curricula, skills, and concepts taught throughout the district.

Interim assessments are common assessments aligned with key educational standards and learning objectives (identified through the district's curriculum scope and sequence or curriculum map) and administered in a single grade level across all applicable schools, typically in a single subject. DDMs should be aligned with key curricular content; districts have the latitude to determine the scope of content coverage (e.g., a single instructional unit, multiple units, a semester, or a year). Interim assessments provide comparable data about student performance (at the individual student, class, school, and district levels). Some districts have invested in interim assessment systems, such as Galileo® (Assessment Technology, Incorporated), Measures of Academic Progress® (Northwest Evaluation Association), and Acuity® (CTB/McGraw-Hill), that provide districts with customized interim assessment and scoring and reporting services.[5]

Performance assessment requires examinees to perform a task, often an authentic or "real" task, rather than simply respond to a prompt. Well-constructed performance assessments are often engaging and meaningful for students. Performances may include demonstrations, explanations, conducting work, problem solving, etc. Examinees are then scored on their performances, which may or may not include products that may be components of the performance.

Groups of teachers often create good performance assessments over time by trying the performance tasks out first with students, then using an iterative process to gradually hone the tasks to improve their instructional and measurement qualities. After the tasks are administered, teachers or test developers review the responses to determine how well the task items and prompts elicited the targeted student behaviors. The review team then adjusts the tasks, items, and prompts in an effort to improve the students' demonstration of the learning objectives.

Performance Task Example: Tax Collector

The following performance task corresponds with the Massachusetts 4th grade content standard: Gain familiarity with factors and multiples. The teacher or test proctor begins the task by demonstrating how to play a game called **Tax Collector.**

The teacher writes the numbers 1 through 6 on the board. He or she explains that there are two players: the student and the "Tax Collector." The student goes first and selects an available number. The Tax Collector always receives all the factors of that number. However, the Tax Collector must be paid each round, so the student cannot take a number if there are not any factors of that number still on the list; it goes to the Tax Collector. Here is a simulated game of three rounds using the digits 1–6:

1 2 3 4 5 6		
	Student	Tax Collector
Round 1	4	2 and 1 (factors of 4)
Round 2	6	3 (the only factor of 6 left)
Round 3		5 (no factors of 5 left)
Score	10	11 – Tax Collector wins

When the students understand the game, they are provided the performance task:

Play several games of Tax Collector using these 10 numbers:

1 2 3 4 5 6 7 8 9 10

Make a record of your best game. Be sure to show which numbers you took and the order in which you took them, not just the final score. Then, answer the following questions:

1. Did you beat the Tax Collector?
2. What number did you choose first? Why?
3. Do you think anyone could ever play a better game than your best game? Explain why or why not.
4. Suppose you were going to play Tax Collector with the digits from 1 to 95. What number would you choose first? Why?

*This example is adapted from National Research Council. (1993). *Measuring up: Prototypes for mathematics assessment.* Washington, DC: National Academy Press. The National Academy Press describes methods to standardize the administration of performance items, including videotaping the initial introduction of the game so that all examinees are provided with the same set of conditions for taking the assessment.

Portfolio assessments involve the purposeful and systematic collection of student work over time. Consequently, they do not measure student

[5] This guide includes references to commercially available assessments. These references are provided for illustrative purposes only and do not constitute ESE endorsement of any commercial product.

performance at any single point in time; rather, they can capture performance at many time intervals. Portfolios are assembled in accordance with a protocol and scored using a well-defined rubric or scoring papers. When appropriately designed and implemented, portfolios provide an opportunity to conduct an “authentic assessment” (one that is intimately embedded in instruction and limits time spent away from instruction) and allow for the examination of students’ complete work products. To that end, they can demonstrate students’ complex thinking, organizational and problem-solving abilities. Scoring should consider the degree to which work products reflect independent versus educator- or peer-supported student effort. As is the case with constructed response items, considerable time will be spent developing scoring guidelines and materials, providing scorers with training, and establishing that the scoring of the student work is done reliably. See the section below on reliability. The MCAS Alternative Assessment (MCAS-Alt)[6] is an example of a well-designed portfolio assessment. The MCAS-Alt is designed to measure a student's knowledge of key concepts and skills outlined in the Massachusetts Curriculum Frameworks. A small number of students with the most significant disabilities who are unable to take the standard MCAS tests even with accommodations participate in the MCAS-Alt.

C*apstone projects* are projects conducted at the end of a course of study that are designed to allow students to demonstrate the breadth of accumulated knowledge, skills, and concepts acquired through the course. There are different approaches to conducting a capstone project, many of which are described in this document. Capstone projects can take the form of rigorous performance or portfolio assessments. Additionally, capstone projects may be presented as any of the following:

- Research projects
- Simulations with a performance or hybrid assessment
- Thesis papers

Hybrid assessments combine an on-demand assessment with a portfolio or performance assessment to achieve a more balanced type of assessment, one that provides a broad representation of content (in the on-demand assessment) and includes complex tasks (presented in the portfolio or performance assessment). Here is an example of a hybrid assessment:

- A hybrid science assessment in an introductory physics class includes 30 selected response items, 4 restricted constructed response items, and 4 performance assessment items. For the performance items, students are presented with a scenario about pendulums and are asked to (a) set up an experiment, (b) conduct the experiment and record the results, (c) describe what they learned from the experiment as it relates to their hypothesis, and (d) expand on what they learned by designing a follow-up experiment.

Commercial Assessments

Commercial assessments are tests purchased from commercial vendors for use as DDMs. There are two basic types of commercial assessments: criterion-referenced and norm-referenced.

Criterion-referenced assessments measure how well a student has learned a specific body of knowledge and skills (i.e., the “criterion” or “domain” of interest). The MCAS, Advanced Placement exams, and most tests and quizzes written by teachers are criterion-referenced assessments. The commercial interim assessment programs used in some schools and districts (such as Acuity® or MAP®) are examples of district-based criterion-referenced tests. These tests are designed to represent the district curriculum. In some cases, they are designed to predict

[6] For more information about MCAS-Alt, see http://www.doe.mass.edu/mcas/alt/

performance on the state’s summative assessment program (MCAS, in this instance) using the performance levels reported in the MCAS exam (Failing, Needs Improvement, etc.). Criterion-referenced assessments have the advantage of connecting examinee performance to a set of standards, grounding the score reporting in a description of how well a student has progressed in learning and development.

Norm-referenced assessments, on the other hand, provide an estimate of how an individual student performed on the assessment compared to a predefined group. The “norm group” is intended to be a representative group of students based on the country’s demographic characteristics at the time the test is developed. The commercial test developer first administers the assessment to the norm group and generates scaled scores and percentiles based on that group’s performance. Students taking the published test then receive scores that are compared to the norm group. Percentiles are one type of comparative score that report the percentage of students (in the norm group) that the examinee performed better than. (For example, a percentile of 75 indicates a much better score than a percentile of 25.) Well-known norm-referenced tests that employ a national-level norm group include the Stanford Achievement Tests, the Iowa Tests of Basic Skills, and the SAT exams.

Standardization

A *standardized assessment* is one that is administered and scored in the same way for all examinees. In high-quality assessments, the conditions for providing accommodations, administering the assessment, applying the scoring procedures, and reporting the scores will be conducted in a consistent or standard manner to ensure all examinees have an equal opportunity to demonstrate what they know and can do on the exam.

Alignment of Content to the Curriculum

An important characteristic of high-quality assessments is that items on the instrument are representative of the intended curriculum. The intended curriculum is usually described in the district’s curriculum scope and sequence or curriculum map. An abbreviated example of a curriculum map provided in Table 2 describes mathematics content to be taught at Grade 4. A curriculum map typically will present the following information:

- The time frame for each section (unit) of curriculum content
- The associated standards
- The curriculum unit connections
- Assessments associated with the content

Table 2. Grade 4 Curriculum Map

Time Frame	Standard	Curriculum Unit Connections	Assessments
↓			
January–March	4.OAT.4, 4.OAT.5	Unit: Factors, Multiples, & Arrays	PA: Tax Collector Unit Test
April–May	4.F.1, 4.F.2, 4.F.3 & 4	Unit: Parts & Wholes	PA: Candy Box Designs Unit Test

Assessments should include a Table of Test Specifications that provides a link between the assessment and the content identified in the curriculum map. When evaluating a potential DDM, district teams should review the Table of Test Specifications to ensure the assessment is well aligned to the local curriculum scope and sequence or curriculum map (see Section 3). District teams developing a new assessment will find that creating a Table of Test Specifications is a critical step in the development process (see Section 4). In some cases, the test items can be linked directly to the academic standards. In other cases, the standards will need to be broken out into observable (e.g., measurable) learning skills or objectives. In the following example, the Table of Test Specifications breaks each standard out into 2–3 learning outcomes. Two levels of cognitive rigor (taxonomy) are identified at the top of the table: lower-order (associated with Bloom's Revised Remembering and Understanding levels) and higher-order thinking (associated with Bloom's Revised Analyzing to Evaluating levels). Each cell indicates the percentage of test score points devoted to the associated learning outcome and level of rigor. These percentages should conform to the amount of emphasis given to each topic during instruction. When populated, a Table of Test Specifications demonstrates the alignment between the assessment's content and the performance domain by ensuring that:

- The standards and learning outcomes listed capture all the key content in that performance domain.
- The percent coverage matches the expected amount of coverage of that topic during instruction.

According to the Table of Test Specifications shown in Table 3, the DDM weights standard 4.OAT.4 more heavily than standard 4.OAT.5 (65% versus 35% score points) and weights higher-order thinking items more heavily than items tapping lower-order thinking (60% versus 40% score points).

Table 3. Table of Test Specifications for Grade 4 Math Test

Grade 4 Mid-Unit Math Test			
	Cognitive Taxonomy		
	Lower-Order	**Higher-Order**	**Subtotal**
Mathematics.4.OAT.4: Gain familiarity with factors and multiples.			
LO #1: Students will identify all factor pairs for integers from 1–100.	15%	20%	35%
LO #2: Students will identify multiples for all integers up to 100.	10%	10%	20%
LO #3: Given an integer from 1–100, students will determine whether it is a prime or a composite number.	10%	0%	10%
Subtotal	35%	30%	65%
Mathematics.4.OAT.5: Generate and analyze patterns.			
LO #1: Given a sequence of block shapes, students will identify the pattern and fill in the missing data.	0%	10%	10%
LO #2: Given a sequence of rectangles, students will identify the pattern rule(s) and draw the next rectangle in the series.	5%	20%	25%
Subtotal	5%	30%	35%
Total	**40%**	**60%**	**100%**

Table 3 is one (abbreviated) example of a Table of Test Specifications. Tables of Test Specifications can break items out into the various categories of rigor, and the cells can display the percentage of score points or the number of test items. Additionally, districts can generate more than one Table of Test Specifications, with additional tables breaking the number of score points or the number of items into item types.

Introduction to Reliability and Validity

Issues of the reliability and validity of educational assessment programs are integral to the use of tests for all educational purposes. Reliability and validity issues affect everyday use of assessment results in classrooms, but because the effects are usually confined to the classroom and because the decisions are so numerous and rapid, the effects are not often noticed. Issues related to reliability and validity become more noticeable when test scores are used to make decisions about students and teachers, as is the case with DDMs in making decisions about teacher impact on students.

This document next explores issues of reliability and validity in the context of DDMs and educational assessment.

Reliability

Reliability, in its broadest sense, refers to the consistency or stability of an assessment. It is an indication of the confidence one can have that differences in test scores reflect actual differences in the characteristic being measured (e.g., what a student knows and can do in math), as opposed to "error" (sometimes referred to as "noise"). Understanding the reliability of an assessment is essential to understanding and interpreting the results of that assessment. For example, consider a math test comprised of a single problem. Students' responses are probably not a good reflection of what they actually know or do not know. The test did not provide enough coverage of the content to provide reliable information. This section describes some common factors that influence scores but are not related to what students actually know or do not know in a subject. It is imperative that DDMs show good reliability so that they can be used to inform an educator's student impact rating.

Many factors may introduce error and hence reduce reliability in test scores. For example, students may be "lucky" and be presented with a testing scenario they experienced in class or "unlucky" and presented with two vocabulary words from texts they had not yet read. In fact, no test is perfectly reliable. All other things being equal, multiple-choice tests tend to be more reliable than tests featuring open response or performance items because raters trained to exercise judgment are not required for scoring multiple-choice items. Similarly, longer tests with more items—assuming the items are of similar quality and are focused on the same content domain—tend to be more reliable than shorter tests. Standardization of test administration procedures can also increase reliability.

Reliability is measured on a 0 to 1 scale, with "1" representing "perfect reliability" and no measurement error and "0" representing "no reliability" and all measurement error. High reliability, then, is *one* characteristic of a high-quality instrument. Measures of reliability are provided in the documentation that accompanies high-quality commercial assessments.

Because reliability is very hard to "see" or detect in educational testing, there are four basic methods for estimating it. The factors affecting stability and consistency play out in different ways in different test situations:

- *Internal Consistency Reliability* refers to consistency among the items in an assessment. If the items represent the same tested content (e.g., content representing 4th grade fractions curricula), they are internally consistent. Consider internal consistency when you have a traditional assessment with many items (such as an end-of-course exam). Because internal consistency is concerned with the degree to which items represent the same content, it is important to consider the internal consistency reliability for each test section separately when tests represent two or more content areas. Internal consistency reliability can be estimated using just one administration of a single form of the assessment.
- *Test-Retest Reliability* refers to consistency (i.e., stability) in test scores over time. To estimate this type of reliability, the same exam needs to be administered to the same group of students after a period of time (say, 2–3 weeks) without teaching students the tested content between administrations. This is a form of reliability for assessments designed to measure student growth using a pretest-posttest design.
- *Alternate (Parallel) Form Reliability* refers to consistency in test scores across different forms of the same test. This type of reliability is important to consider when educators create different forms of the same test (such as parallel forms for a pre- and a posttest) or when educators compare tests that have been altered in some way (e.g., items or directions have been changed). The process for estimating alternate form reliability is similar to test-retest reliability. Administer the two forms to the same group of students, but this time keep the time interval between testing periods short (e.g., within a week or so to measure the relationship between the parallel tests without respect to the time interval). The correlation between the two sets of scores is the alternate form reliability estimate. This form of reliability will only be important for DDMs if districts are using alternate forms of the same assessment for the pre- and posttest administration or if districts are changing a test form from one year to the next and want to ensure that the revised form is consistent with the earlier form.

Illustrating the Importance of Reliability and Validity

Three tailor's assistants measure the waist sizes of clients for fitting pants and skirts. The assistants have slightly different ways of measuring clients' waists. This results in three slightly different measurements going to the tailor and three slightly different fits for the clients. One assistant may pull the tape tighter than the others; one may put his or her fingers inside the tape measure while the others are careful to keep their fingers outside of it. Even for the same assistant, unintentional differences in procedures from one fitting to the next may result in slightly different measurements and fits (e.g., measuring at the "true" waist versus measuring above or below the true waist). These differences in measurement are examples of reliability, specifically:

- Inter-rater: differences in measurement within each assistant
- Internal Consistency: differences in measurement for a single individual assistant

The tailor's business also offer dresses and suit jackets. For fitting jackets, the assistants take shoulder and arm measurements. Client complaints illustrate different types of validity.

- Content Validity: The jackets did not fit well because no chest measurements were taken.
- Relationships with Other Variables and Outcomes: The jackets did not fit as well as the jackets made by the tailor down the street.
- Consequential Validity: Some clients alleged they were harmed by the poor fits. A bride and groom asked for their money back after their groomsmen complained the jackets didn't fit and they were forced to wear their own clothes to the wedding instead.

- *Inter-Rater Reliability* refers to the degree of consistency among raters who are rating the same performance. This type of reliability is important to estimate when DDMs contain items that have constructed response, performance, or portfolio responses that must be scored using a rubric or other scoring mechanism that relies on rater judgment. The process for estimating inter-rater reliability is to have multiple raters rate a student on some measure of performance and to evaluate the agreement between the raters. The goal in assessing this type of reliability is to ensure that regardless of the rater, students who perform the same will receive the same or a very similar score. Inter-rater reliability is enhanced when there are clear scoring materials and training for scorers. After these procedures are in place, inter-rater reliability must be monitored to ensure that scoring is conducted as expected. The simplest method for estimating inter-rater reliability is to use a concordance table that records the scores for two raters rating students on the same item. The agreements are recorded along the diagonal of the table. The simple percentage agreement is computed by dividing the number of times the two raters agree by the total number of items scored. As shown in Table 4, the two raters agree 47% of the time (14 ÷ 30), which is not a terribly high level of agreement. The reliability coefficient of .47 indicates that (a) more can be done to improve scoring consistency for this item, (b) the scorers need additional training, or (c) both.

Table 4. Concordance: Two Raters

		Rater 1				
		1	**2**	**3**	**4**	**Total**
Rater 2	**1**	4	2	1	0	7
	2	3	5	3	1	12
	3	0	2	2	0	4
	4	1	1	2	3	7
	Total	8	10	8	4	30

These four types of reliability are summarized in Table 5.

Table 5. Four Types of Reliability

Type of Reliability	Description	Type of Reliability Coefficient	When to Establish This Type of Reliability
Internal Consistency	The degree to which the items are measuring a similar set of content in the same way	Coefficient Alpha or similar	When the DDM is comprised of many items representing the same content area
Test-Retest	The degree of stability of scores over time, estimated in the absence of instruction in the content area	Correlation between the scores at Time A and Time B	When a DDM is used to establish pre- and posttest scores
Parallel Forms	The degree of similarity between two different but parallel forms of an instrument	Correlation between the test scores from Form A and Form B	When two parallel forms of a DDM are used or when the DDM is changed from one year to the next

Type of Reliability	Description	Type of Reliability Coefficient	When to Establish This Type of Reliability
Inter-Rater	The degree of consistency between two or more raters (scorers)	Simplest: Percentage of joint agreement using a concordance table; More Sophisticated: Cohen's Kappa, Correlational Methods	When a DDM contains items that require rater judgment

For a potential DDM, ESE recommends that districts begin to examine reliability evidence to determine the degree to which the identified DDM is providing stable and consistent measures. If districts find that a potential DDM has low reliability (e.g., less than .8 for internal consistency measures and less than .7 for test-retest, parallel forms, and inter-rater reliability), districts are encouraged to revise the DDM and/or the administration/scoring procedures to improve the reliability of the assessment.

Validity

Reliability is a required feature of a high-quality instrument. However, reliability alone is not enough; an instrument with high reliability may not be valid for a particular use. Validity is concerned with whether the instrument is appropriate for its intended use. Because DDMs will be used to estimate student growth in a content area, an instrument valid for that use must show evidence of three important validity characteristics, which are described below. Validity is concerned with the development and use of the DDM; as a result, districts can start to collect validity evidence on DDMs as they are being selected, developed, and piloted.

- *Content Validity*. Content validity provides evidence that the instrument content aligns with and samples appropriately from the intended content. The intended content for many educational instruments is specified in the district's curriculum scope and sequence or curriculum map. The DDM will sample content from the curriculum map. The content on the DDM will be described in the Table of Test Specifications, as shown on p. 13.

  Determining content validity is an important up-front consideration in the selection of an assessment for use as a DDM. A district team of content and assessment personnel should review the instrument against the Table of Test Specifications and compare the Table of Test Specifications to the district's curriculum map to determine the degree to which the instrument represents both documents. The team will further examine the alignment to rigor to determine if the cognitive (or other) demands represented in the curriculum map are apparent in the instrument itself and in the Table of Test Specifications. Note that other characteristics of the instrument development process are also important in establishing content validity and development evidence, including the degree to which the instrument is administered in an appropriate and standardized way and that the time allotted to take the test and other administration conditions are also appropriate. Any issues identified in this process can be addressed by revising the tested content (for district-developed or district-modified DDMs) or by selecting a DDM with better content representation (for commercial assessments).

- *Relationships to Other Instruments, Outcomes, and Variables*. This type of validity evidence is accumulated if scores are found to be related to similar measures or other intended outcome measures and unrelated or less related to dissimilar measures and outcomes. For example, one would expect for scores on two general 5th grade reading tests to be related (high scores on one test being associated with high scores on the other; low scores on one test being associated with low scores on another). If students take both tests and the scores are indeed related, this provides validity evidence. The inverse can also be true. Validity evidence is accumulated if the DDM is less related or even unrelated to dissimilar instruments, measures, and outcomes. For example, one would not expect scores on a 5th grade reading test to be strongly related to scores on a physical fitness test, and a weak relationship between the two also provides validity evidence. Similarly, validity evidence is accumulated if scores are related to expected outcomes. For example, validity evidence would be accumulated if scores on math achievement at the end of 7th grade were related to scores on an end-of-year algebra test in 8th grade. Predictive validity evidence would allow a district to use a DDM to predict the results on a future outcome, such as using a DDM to predict students' results on a state or national exam.

Validity evidence in this category is generated by comparing results from two or more measures or outcomes by computing a correlation between pairs of instruments..

- *Consequential Validity*. Consequential validity evidence is accumulated if the use of the scores is generally experienced as fair and beneficial for the students and other persons affected by the test results. To establish consequential validity for DDMs, the instruments should be shown to contribute to student learning and to provide benefits to teachers. For example, results can be used to ensure the following:
  - Improvement of instruction to students
  - Realignment of the curriculum to provide all students with more opportunity to learn the key material
  - Provision of high-quality professional development opportunities for teachers

Table 6. Types of Validity Evidence

Type of Validity	Description	Question to Ask
Content Validity	• The degree to which the content of the DDM aligns with the district curriculum at the expected level of rigor	• Does the DDM represent the content and rigor of the instructional/curricular content?
Relationships to Other Measures, Outcomes, and Variables	• The degree to which the scores are in agreement with or predict other tests and/or criterion	• Is the DDM measuring what it is purporting to measure?
Consequential Validity	• A comparison of the intended use(s) of the assessment to the intended and unintended outcomes of that use(s)	• Does the DDM confer the intended benefits and reduce unintended harms for students and teachers?

Validity is seen as a "unifying concept" that allows districts to describe instruments as more or less valid for each intended purpose. For the purpose of DDMs, ESE recommends that districts identify validity evidence for each category described above, particularly emphasizing evidence related to content and consequential validity. These categories of validity evidence are summarized in Table 6.

Fairness

To be acceptable to teachers, students, the public, and other interested stakeholders, assessments must provide examinees an equal chance to show what they know and can do. Numerous practices indicate that an assessment is fair, including whether all students have an adequate chance to demonstrate their knowledge during the assessment process, whether students had an ample opportunity to learn the content, and if the instrument is not biased (see definition of bias below). It is important for DDMs to be perceived as fair by stakeholders, including educators, district administrators, and others.

Bias

The primary source of bias is item bias. Individual instrument items may perform in a biased way against specific groups by referencing persons, groups, experiences, or cultures that the examinees may be more or less familiar with. To avoid bias in item writing, test developers are advised to review the instrument content to ensure that the following guidelines are adhered to:

1. Selected language on the instrument holds the same semantic meaning for all examinees.
2. Selected language on the instrument communicates the intended affective (emotional) effects for all examinees
3. Stereotypical language is avoided, especially language that characterizes groups as more or less powerful, advantaged, smart, attractive, etc., than other groups.
4. To the extent that cultural or demographic groups are represented on the instrument, the representation attempts to acknowledge all groups.
5. Main characters in texts show good representation across cultural and demographic groups.

Specifically, instruments should strive for gender, cultural, and demographic balance and should be inclusive of the groups of students taking them. While it is important to note that instrument items will draw from examinees' backgrounds in often unintended ways, it is the work of the instrument review team to ensure that the content used in the instrument prompts consists of a reasonable range of experiences, group representations, and backgrounds. With respect to experiences, the instrument should include items that tap into common experiences for the group and not include experiences that favor one group over another (e.g., using more than one sports example when the group of examinees are not all athletes).

National Evaluation Systems (NES; 1991, pp. 4 and 15) provides examples of biased and nonbiased language in assessments to assist review teams in reducing instrument bias. Table 7 provides a few examples:

Table 7. Examples of Bias in Item Writing

	Poor	Better
Gender Bias	Identify the major stages in the evolution of man.	Identify major stages in human evolution.
Power and Status	Congress finally granted African Americans broad enforcement and protection of their right to vote in 1964.	After a long struggle, African Americans won legal enforcement and protection of their right to vote in 1964.
Power and Status	Many universities are now permitting retirees to enroll in degree programs.	Older persons are now enrolling in university courses and degree programs in ever increasing numbers.

A related problem is known as the *stereotype threat*[7] (Steele & Aronson, 1995). A stereotype threat is one in which examinee performances on an assessment is changed when examinees belonging to a particular subgroup are reminded that that subgroup performs better or worse on that particular type of task. For example, if a narrative prompt on an assessment describes negative impacts of poverty on student performance, the prompt can activate that stereotype in examinees who are receiving free or reduced-priced lunch, perhaps negatively affecting their performance on the exam.

After examinees take the exam, there are quantitative methods for identifying possible bias on the instrument through an examination of the students' responses by subgroup. These methods are described in Appendix B.

Documentation

To ensure that assessments are used in an appropriate and standardized way, they are typically accompanied by documentation. This documentation ensures transparency in the development and administration of assessments and can include the following documents:

- *Technical manual*. An assessment's technical manual is a comprehensive technical document. It should identify the purpose of the assessment as well as when, how, and to whom it can be appropriately administered. The document should explain how the instrument content was identified and developed, specific requirements regarding the administration of the instrument, the process for scoring the instrument, the types of scores reported by the instrument, and information regarding the proper interpretation of scores. It should include information a potential user may need for determining the assessment's psychometric quality, such as reliability, validity, and bias analyses. The technical manual may also include other policies describing the appropriate use of the instrument, such as the training requirements for instrument administration and the interval of time before which the instrument must be reevaluated. Technical manuals are typically developed for commercial assessments; districts wishing to use commercial assessments for DDMs should consult the technical manual to review the instrument quality information reported there.

[7] For more information about stereotype threat, see http://www.reducingstereotypethreat.org/definition.html.

- *Administration manual*. This document details the instrument administration procedures. When followed closely, it standardizes the administration procedures, enhances instrument security, supports the equitable treatment of examinees, and minimizes errors in instrument administration and scoring. The instrument administration manual typically includes a list of the examinee resources (e.g., calculators or dictionaries) that are required and prohibited during administration, a description of the appropriate conditions for administering the instrument, a script that the administrator reads to students, details regarding what can and cannot be said or done by students and by those administering or proctoring the assessment, instructions for timed tests, insights into how to deal with emergencies that may arise during a test, and a list of the examinee accommodations that are permitted (e.g., offering extra time or administering the exam in a quiet setting outside the classroom). It should explain clearly the procedures for scoring the instrument and procedures for training scorers to score the instrument items reliably. Finally, the instrument administration manual should include instructions to ensure instrument security, including procedures for accounting for instrument materials (e.g., how to check out and return instrument materials) and other administrative details (e.g., how to process or score answer sheets).

If a district decides to use or buy an existing assessment, the content described above should be provided by the test developer; and if not, it can be requested from the test developer, as the information contained in these documents will help determine whether the assessment is of high quality and whether it is appropriate for use as a DDM. If your district decides to build an assessment from the ground up, this is the type of content the district can begin to develop to provide documentation for use of the assessment as a DDM.

Note that formal versions of these documents are expected if the assessment is commercially developed. If the assessment is locally developed, then this is documentation the district can begin to accumulate. For more information on how to select an existing DDM, see Section 3. For details on how to build a new assessment, see Section 4.

Section 3. Selecting an Existing Assessment as a DDM

Districts have three basic options regarding DDM development. They can do any of the following:

- Select an existing assessment
- Modify an existing assessment
- Build a new assessment

Assuming that a high-quality assessment exists, using an existing assessment may be a better choice over building a new assessment given the nature and level of resources typically required to develop a new assessment. In this section, ESE describes the steps involved in identifying and selecting an existing assessment for use as a DDM of educator impact on student learning. A firm grasp of the concepts presented in Section 2 is required for understanding the content of this section as well as Section 4, which describes the process for developing a new assessment.

Identifying Coverage Areas for DDMs

Before beginning the process of identifying an existing assessment (or building a new one) to serve as a DDM, district teams can start to identify the number and type of DDMs that will be required. Recall that DDMs must be comparable across the district for all educators in a grade or subject and that districts should identify at least two measures to be used to inform each educator's student impact rating. ESE recommends that each district begin the work by engaging in a DDM Gap Analysis, which involves linking the potential DDMs to courses taught in the district and linking educators to be evaluated to the courses they teach. In the case of noninstructional staff, districts can begin to identify the support or services noninstructional staff provide as well as the number of students receiving those services.

A DDM-Educator Alignment Tool[8] has been developed to support districts as they perform this task. The tool requires districts to identify the courses taught in the district as well as to identify each educator and the content he or she teaches or the support services he or she provides.

Identifying Potential Existing Assessments

After becoming familiar with the DDM requirements, districts can begin the process of identifying existing assessments that have the potential to be used as DDMs for each subject/course area. The DDM-Educator Alignment Tool supports district teams in the performance of this task.

The first step is to match statewide measures of student learning, growth, and achievement to district educators. ESE provides districts with the following:

- Student Growth Percentiles (SGPs) for English language arts and mathematics in Grades 4–8 and 10
- MCAS-Alt scores for students with disabilities who are unable to take the MCAS exams
- ACCESS (formerly MEPA) scores for students who are eligible English Language Learners

[8] To access the DDM-Educator Alignment Tool, see http://www.doe.mass.edu/edeval/ddm/webinar/Educator-Alignment-Tool.xlsx.

The second place to look for available DDMs is to review the assessments already in use in the district. ESE has developed an Assessment Quality Tracking Tool[9] to help districts review the properties of existing assessments to determine those that may work as potential DDMs as they already exist and those that may work with modification.

After the available and existing DDMs are identified, districts can begin the process of searching for additional DDMs or developing new ones. Identified resources include the following:[10]

- Part VII of the Massachusetts Model System for Educator Evaluation titled Rating Educator Impact on Student Learning Using District-Determined Measures of Student Learning, Growth and Achievement
- *Tests in Print*, published by Buros Center for Testing, which provides a comprehensive list of commercially available English-language tests[11]

Many other potential sources of information about existing assessments include materials released by other districts and other states.

The remaining information in this section is provided to guide districts when selecting an existing district or commercial assessment, while information in Section 4 is provided to assist districts with building a new assessment.

Evaluating Potential Existing Assessments

Districts should evaluate each potential DDM to determine quality and verify the appropriateness of the assessment for use as a DDM. The basic process involves gathering information about the assessment's purpose and quality and conducting an internal review of the assessment. During the internal review, districts should consider numerous criteria using the core assessment concepts that were described in Section 2. Descriptions of the criteria to be evaluated are outlined below to guide districts through the remainder of the process.

At the outset, ESE wishes to address the potential desire to repurpose classroom assessments for use as DDMs. Classroom assessments such as formative and summative assessments will not, in their raw form, be suitable for use as DDMs because their content is typically tied to that specific classroom use for a specific population of students. Districts may wish to consider scaling up rigorous classroom or unit-based assessments for use as DDMs. This process would involve some level of assessment development to revise the content to better reflect the district curriculum map and to revise some of the instrument items to better reflect the intended level of rigor. Further, other assessment development work may need to be done, such as creating administration protocols and scoring guides as well as identifying student accommodations. For more information on developing DDMs, please refer to Section 4 of this guide.

Note that districts may follow the steps below to decide whether an assessment is acceptable to pilot as a DDM during the 2013–2014 school year. However, districts may instead find that piloting an assessment will actually yield much of the information needed to decide whether it is a suitable DDM described below. In other words, piloting may happen before or after districts complete the evaluation process described in this section.

[9] To access the Assessment Quality Tracking Tool, see Appendix A.

[10] Additional information about commercially available assessments is included in Appendix B.

[11] For more information about *Tests in Print,* see http://buros.org/test-reviews-information.

Step 1: Gather Documentation Regarding the Quality of the DDM (Reviews, Critiques).
For each promising assessment considered for use as a DDM, districts should gather documentation regarding the quality of the instrument. When possible, information should be gathered from sources external to the publisher or developer as well as from the authors/test developers. For assessments developed by the district, this process may include collecting feedback from teachers who have used the assessment. For commercially developed assessments, this process may include reviewing documentation from the following sources, where available:

- Published information from sources external to the test developers such as:
    - Formal reviews published by sources external to the author or publisher
    - Informal reviews of the assessment published by research or evaluation groups
- Published or unpublished information from the author or publisher of the assessment such as:
    - Technical manual
    - Administration manual
    - Policies regarding the assessment
    - Any other available information (e.g., recent reliability data, newly created norms)
- Unpublished information from sources external to the assessment authors (e.g., teachers in the district or other districts who are using or have used the assessment) such as:
    - Testimonials, ratings, or other user feedback
    - Data that demonstrates the quality of the assessment
    - Any other available information that can be shared

Vet the material gathered about the assessment's quality with an eye toward how knowledgeable and current the source is. Knowledgeable sources will include information about the assessment concepts outlined in Section 2, including construction, reliability, validity, bias checks, and administration and reporting procedures. Knowledgeable sources will further address the quality of the assessment as it aligns to the purpose or use of the assessment. (Here, reviewers will be interested in information about assessment quality for the purpose of tracking student progress.)

Step 2: Use the Documentation to Evaluate the Assessment.
Review the documentation gathered to determine whether the assessment is of sufficient quality for use as a DDM. Because the information may exist in one or more locations in the gathered documentation, the criteria below are organized by topic. Following the tasks below will assist the district in conducting a systematic evaluation. Appendix A, the Assessment Quality Checklist and Tracking Tool, provides districts with a place to collect evidence and information on potential DDM characteristics and quality.

The following list provides the basic components of and information about an assessment that districts should collect and review. If components or information is missing but the assessment is a promising candidate for use as a DDM, districts can begin to develop missing components and/or gather evidence of the assessment's quality (e.g., pilot data). Note that the evaluation process can be halted at any time if the review team determines that the assessment is not appropriate for use as a DDM.

The Assessment Quality Checklist and Tracking Tool will aid districts in documenting the following steps:

1. Identify general information about the assessment.
    a. *Grade/subject or course:* Identify the grade/subject or course aligned to the DDM.
    b. Potential DDM name or title
    c. *Potential DDM source:* Identify the source of the assessment (e.g., district-developed, commercially developed, developed by another school district).
    d. *Type of assessment:* Refer to the types of assessment described previously in Section 2. Indicate whether the potential DDM is an on-demand assessment, performance/project, portfolio, hybrid, or other type of assessment.
    e. *Item types:* Refer to the item types described in Section 2 of this guide.
2. Consider the utility and feasibility of using the assessment
    a. *Utility:* Districts are strongly advised to select DDMs that provide useful results for students and educators. Districts are encouraged to include teachers and administrators in the DDM identification process to determine the utility of DDMs.
    b. *Feasibility:* Take stock of the cost, length, accommodations, technology needs, and report types of the DDM.
3. Identify and evaluate the components of the assessment.
    a. *Table of Test Specifications:* The Table of Test Specifications describes the alignment and rigor of the instrument’s content by matching all items on the assessment with the tested standards (and sometimes learning objectives) and the level of rigor (such as Bloom’s Revised Taxonomy).
    b. *Administration protocol:* This protocol includes proctoring directions, security provisions, and how to provide student accommodations.
    c. *Instrument:* The instrument refers to the test itself.
    d. *Scoring method:* The scoring method refers to the availability of a scoring key for selected response items and scoring papers or rubrics for the scoring of constructed response items. It could take the form of a scoring guide.
    e. *Technical documentation:* The assessment may be accompanied by additional documentation. This documentation may include a technical manual, which describes the reliability and validity evidence associated with the assessment, along with instrument development procedures. Well-known commercial assessments will frequently be accompanied by technical documentation.
4. Evaluate the level of alignment to the curriculum and to the intended level of rigor.
    a. *Alignment to curriculum:* Indicate the procedure(s) used to establish the alignment between the district curriculum and the DDM.

   b. *Alignment to intended level of rigor:* Indicate the procedure(s) used to establish the intended degree of rigor on the assessment. Indicate the use of taxonomy, such as Bloom's Revised for establishing rigor.

5. Gather evidence of and evaluate the technical qualities of the assessment.

   a. *Reliability:* Evidence for reliability is collected as described earlier in Section 2. The type of reliability evidence collected should conform to the type of assessment. For example, an internal consistency reliability coefficient is typically reported for on-demand tests.

   b. *Validity:* Evidence for validity is collected as described in Section 2 of this guide. Districts are advised to collect three types of validity evidence: content, relationships to other measures, variables, outcomes, and consequential validity evidence. Districts are advised to pay particular attention to content and consequential validity evidence.

   c. *Nonbias:* Evidence of nonbias indicates that students who belong to particular gender, demographic, and cultural groups are not advantaged or disadvantaged by the instrument items included on the assessment. Nonbias evidence is collected by reviewing items singly and collectively for possible bias, including under- and overrepresentation of student demographic groups. On commercial tests, quantitative measures of test bias may also be reported.

   d. *Item quality:* These properties provide evidence that each instrument item is performing well (i.e., has an appropriate level of item difficulty and discrimination) and that the instrument items collectively show a range of difficulty from easy to hard so that the instrument shows no floor or ceiling effects as either a pretest or a posttest. Finally, evidence that the instrument contains "instructionally sensitive" items is also preferred. Additional information about item difficulty, item discrimination, and floor and ceiling effects, see Appendix B.

Finalizing a Decision Regarding DDMs

After conducting an evaluation of the assessment and recording results using the Assessment Quality Checklist and Tracking Tool, district teams will have enough information about an assessment to begin to make a determination about whether it can be used as a DDM. Often, review teams will find that they have some but not all of the information needed to make a decision. If more information is needed, the district can do one of two things:

- Continue to collect information while piloting the assessment as a DDM
- Collect more information about the potential DDM before making a final decision

Districts may encounter a scenario in which an assessment would have sufficient quality for use as a DDM if it were revised to improve its technical qualities. If the assessment is locally developed or is in the public domain, then the district may decide to pursue the assessment for use as a DDM with modifications. During the process of making modifications, districts can update pertinent information in the Assessment Quality Tracking Tool. (The updated information will be specific to the changes in the modified assessment.) After the assessment is revised, districts can pilot it and update the remaining information in the Tracking Tool to determine if the revisions on the assessment had the intended effect of improving the quality of the assessment.

Monitoring the Assessment’s Use

The process of maintaining a district’s set of DDMs will be ongoing. After a district selects and implements a DDM, the district should monitor the quality of the assessment to determine if it is living up to its promise of being a high-quality assessment. Districts may wish to continue to monitor the following assessment characteristics:

- Continued alignment to the district’s curriculum and intended degree of rigor
- Instrument security (i.e., procedures intended to ensure that assessment results are not tainted by improper instrument administration procedures or by an overfamiliarity with the exam or the exam contents)
- Reliability
- Validity associated with its use as a DDM, including useful results and good score reporting
- Feasible administration and scoring procedures

Section 4. Developing a New Assessment

In some instances, districts may choose to build their own DDM. Building a new DDM allows districts to create an assessment closely aligned to their learning goals and practices. More specifically, local developers have greater latitude in selecting:

- The tested content and the level of cognitive demand, often from a much broader range of content areas than is typically available in commercial testing programs
- The item types (Local tests may go beyond the typical selected response/constructed response item types available in commercial programs.)
- The administration and security procedures

In addition, the process of developing the assessment is an opportunity for rich conversations among educators about student learning and effective educator practice.

The following material outlines the steps for building a DDM. (Note that these steps describe a "generic" DDM, not one for a specific subject or grade.) Some of these steps will appear somewhat differently based on the type of assessment being developed (e.g., on-demand versus portfolio).

Determining What Needs to Be Measured

Districts can begin the process of developing a DDM by identifying the content to be represented in the instrument, which can be accomplished by examining the relevant curriculum map and determining the scope of content coverage. (Coverage may include the entire school year, the semester, or just one unit of instruction.) This information will be considered when determining which type of assessment to develop to fit the intended content.

Selecting an Assessment Type and an Administration Method

After the type and degree of content coverage has been identified, it is time to consider the type of assessment and an administration method. Regarding assessment type, choose the type of assessment (e.g., on-demand, performance, portfolio, or hybrid) that is well aligned to the intended content and performance domain. Districts may wish to consider multiple assessment types before determining which one to use. There may be several different ways to assess the identified content. For example, performance assessments that require the student to perform a task or activity may be a good choice for assessing skills and concepts, but a hybrid assessment that includes a performance component may also work well and may provide greater coverage of the identified content (through the inclusion of more instrument items). A district should evaluate the advantages and challenges of different types of assessments based on its particular situation using the information presented in Section 2.

Next, taking into consideration the assessment type selected, identify the types of items to use in the assessment. When designing an instrument to assess broad areas of content (such as a grade-level math test), districts may wish to consider using a mix of item types to improve content coverage and to incorporate a range of cognitive demands and task complexities (each item representing a different facet of the content, eliciting a range of concepts and skills). The item types on such a test may range from selected to constructed response and may include some performance items. In some situations, a particular item type may be used more than other item types, depending on the intended area of content representation. For example, in a career/vocational technical assessment of carpentry skills, the test may include more performance items than selected or constructed response items. In all cases, however, districts are advised to build assessments to reflect a range of cognitive demands and a range of item difficulties (from easier to harder items) so that (a) the content coverage of the test provides good representation for the district curriculum, and (b) the test is able to provide scores for examinees with a range of ability levels in that content area. What districts should avoid is building tests that (a) reflect poorly on the content and rigor represented in the curriculum mapping and (b) show "floor and ceiling" effects.

Floor and Ceiling Effects

Floor and ceiling effects occur when a test does not provide enough items at the low (floor effects) or high (ceiling effects) levels of the test scale so that, in the case of floor effects, there are a lot of low-scoring students receiving almost no score points on the test (e.g., "bottoming out" of the test scale), or in the case of ceiling effects, there are many high-scoring students receiving all or almost all of the score points on the test (e.g., "topping out" of the test scale).

DDMs should measure student growth. As a result, these are assessments that will be administered more than once, and they must be built for that purpose. While information about methods for measuring student growth will be addressed in detail in *Technical Guide B,* here are three points to keep in mind when developing DDMs:

1. All DDMs will be administered to students in different cohorts. (DDMs are intended to be administered to students every year.) As a result, the instruments need to be of sufficient quality for yearly administration, and the content of the assessment should be secured to ensure that educators are focused on teaching to the curriculum map and not specifically focused on teaching to the content on the DDM. If the district plans to make changes to the instrument, the impacts of those changes can be studied using the parallel reliability approach described on p. 15. Districts may explore measuring growth using a posttest-only design.

2. Some DDMs will use a pretest-posttest design. In this type of design, a pretest is given to students prior to receiving instruction in the content area, and a posttest is given after receiving instruction. Some districts will expand on the pre-post design by administering the assessment at a mid-way point through the year. Best practice for test development using the pretest-posttest design method is to develop an assessment that represents a wide range of item difficulties and rigor so that examinees of all abilities can show some capability on the pre-test (and on the posttest). This best practice is often accomplished by including some content represented in the previous year's curriculum map as well as content represented in the next year's curriculum map (again, to avoid ceiling effects). The items selected on a pretest-posttest design can further be piloted to identify items that are "instructionally sensitive." Instructionally sensitive items are those that pick up changes in student capabilities resulting from instruction. Test security remains an issue for pretest-posttest designs, particularly for on-demand, hybrid, and even performance assessments. Finally, some developers of tests using this design utilize parallel assessments so that examinees do not become overly familiar with the tested content and so that the amount of content represented on the assessment increases.

3. Some districts may use a “repeated measures” design. Such a design involves administering the same assessment to students, individually or in groups, many times over the course of a year to “track” student gains over time by using the results of the assessment to identify gains in performance. Assessments designed for this purpose include many portfolio and observational assessments.

Example: Repeated Measures Design Using a Performance Assessment

A reading inventory can be administered to students in Grade 1 where proctors record students’ fluency and comprehension using a series of “leveled” readers (reading passages that are calibrated to a range of reading difficulty that spans from early reading to advanced reading). Students’ progress is recorded and monitored throughout the year, and the average progress and attainments of students are assessed at various points of the year to determine student gains at the classroom level.

Example: Repeated Measures Design Using a Portfolio Assessment

A portfolio assessment program in writing asks students to purposefully select finished pieces that represent various aspects of the writing curriculum (a persuasive writing piece, an informational brochure, etc.). The pieces can be scored in two ways. First, a rubric or scoring papers can be developed to score each individual assignment against the criteria for that particular task. Second, a rubric or rubrics can be developed to evaluate students’ progress in writing conventions and progress in topic development over time (for example, to capture longitudinal changes in writing).

The actual mode of assessment administration must also be considered. For example, an on-demand assessment may be administered electronically as a computer-based test (CBT) or as a paper-based test (PBT). If using paper-and-pencil administration, determine whether it will be possible to use a machine score answer sheet to save time and reduce hand-entry errors of the test data. Determine scoring techniques and turnaround time for results. Portfolios and performance assessments require other considerations such as if and how the activity or performance will be recorded.

Building an Initial Draft of the Assessment

After identifying the content and the type of assessment needed, build a draft of the assessment that can be piloted.

Create a Table of Test Specifications

After the district has identified the content to be captured on the assessment, it should make a Table of Test Specifications that identifies the sampling plan and organization of the items (see description of Table of Test Specifications in Section 2). Individuals who are experts in the content of that performance domain and in assessment practices should be involved in making decisions regarding the number and type of items from each content area and in determining the degree of rigor reflected in the items.

Finalizing the total number of items is a key part of the table. The decision about the number of items is a function of many factors, including the size and complexity of the domain to be captured, the acceptable length of time for students to take the test, and how long students will need to respond to the type of items that are planned.

Tables of Test Specifications also require identifying the type of items that will be used to capture student growth. If multiple types of items are used to construct the assessment, there needs to be a thoughtful balance between them. In addition, the rigor of the items must be aligned with the curricular goals of the district. Bloom's Revised Taxonomy is typically used for identifying item rigor, although other taxonomies (e.g., Webb's Depth of Knowledge criteria) may be used. Finally, with respect to rigor, districts may wish to include items that range in task complexity from simple tasks (e.g., answering a selection item on a test) to more complex tasks (e.g., tasks that require examinees to demonstrate multiple abilities, such as project development, understanding and creativity, research and information gathering, organization, and project completion).

Select or Write Items

Some districts may have access to item banks from which they can select items for their assessment. Districts using item banks are advised to select items according to the design and purpose of the assessment, using the Table of Test Specifications as a guide, selecting items that appear to best fit the instrument design with respect to content, rigor, and—perhaps—complexity.

Many districts will find that they want to develop items for their DDMs. Item development is typically an iterative process with several rounds of writing and editing. While item development somewhat exceeds the scope of this document, ESE offers the following advice to start on the process:

- Convene a team of educators to survey the assessment content embedded in the instructional materials that align with the tested content. Assessment content may include unit tests and quizzes, formative assessment questions and tests, embedded performance tasks, and end-of-term or end-of-semester grade-level tests. Ask the team to provide examples of items that:
    - Have been shown to be "instructionally sensitive" or that provide a good demonstration of student learning from the beginning to the end of a unit
    - Demonstrate levels of content knowledge at the start, the middle, and the end of learning units
    - Encourage the use of good instruction in the classroom by providing insights into student learning, misconceptions, or other instructionally relevant information
    - Embody desired levels of rigor and/or complexity
- Use surveyed items as ideas or shells for the development of items with similar desired properties.
- If developing items from scratch, employ a basic process of item development:
    - Begin asking experts in the content area to generate items that align with the curriculum, are at the appropriate level of complexity, and exhibit standards of universal design (i.e., items that are accessible by all students).
    - Refine the initial item drafts by (a) revising text to make items more clear and correct, (b) providing good instructions, and (c) improving response options for elected response items.
    - When draft items are complete, ask the item writing team to review items for accuracy and quality, edit the items, and then conduct additional rounds of review.

Items must be written so that they stand alone and do not provide clues to other items on the instrument. If response options are provided, the correct answer(s) should be clear to knowledgeable examinees. The distractors should be plausible (to examinees who do not know the material on the item) but incorrect.

ESE has several specific guides that detail the item writing process and requirements, including the following:

- http://www.doe.mass.edu/mcas/student/2011/
- http://www.doe.mass.edu/mcas/student/elacomp_scoreguide.html

Build a Scoring Key or Rubric

A scoring key denotes the correct response for each selected response item. Often, the key identifies the correct answer through a letter (e.g., A, B, C, etc.). Typically, each selected response item has one correct answer identified, and no other correct answers should appear in the item options. Occasionally, selection items allow two or more correct answers, and in these instances, more than one correct answer needs to be identified in the scoring key.

Constructed response items cannot be scored through a key and require scoring materials that typically include either a scoring rubric and/or scoring papers. The scoring rubric and scoring papers assist scorers with identifying the number of points to assign each student answer, based on the criteria of what constitutes a good or poor response for each constructed response item. The following are the basic steps in developing scoring materials for constructed response items:

- Prior to administering the item to students, attempt to delineate and describe levels of performance from low to medium to high. Try to define clearly what performance would look like in each level. Determine if you need more levels of performance to adequately capture the range of performance. After this step is done, accurately describe what performance looks like in each range and include these descriptions in a checklist or a scoring rubric that will later be used to score the item.
- After the item is administered, divide the examinee item responses into the levels previously identified. Determine if the number of levels is adequate to delineate the levels of student performance and whether the descriptions adequately captured what performance looked like at each level. Revise the number of levels and/or the performance level descriptions as needed.
- Divide the student responses into the levels. Identify responses that "exemplify" these levels of performance. These exemplars can serve as scoring papers that illustrate each score point.

The rubrics and scoring papers developed in this process define the levels of performance for raters to ground their understanding of what responses look like at each score point. Training raters on how to use the rubric or score report is a critical part of the process.[12] Refer to the following ESE documents for additional information:

- http://www.doe.mass.edu/mcas/tdd/g3sr.html
- http://www.doe.mass.edu/mcas/tdd/longvopen.pdf

[12] To view a sample rubric created by ESE, see http://www.doe.mass.edu/mcas/student/2011/question.aspx?GradeID=5&SubjectCode=ela&QuestionID=15359.

Review Items for Bias and Sensitivity

Finally, the item writing or item selection team can review all items for bias. The purpose of this review is to ensure that students are not unfairly advantaged or disadvantaged when responding to items due to any number of factors in the item contents, including insensitive language, embedded negative stereotypes, or an overemphasis of cultural or experiential material. To ensure that no student is disadvantaged, item content (stimulus material, stem, response options) is reviewed to ensure that it does not include elements that may be insensitive or disrespectful to the student's ethnic, religious, or cultural background. Keep in mind that "culture" is a broad term that includes disability, language proficiency, socioeconomic status, and regional differences. In selecting items for use on the MCAS exams, the bias committee "ensures that no questions appearing on the tests may potentially disadvantage particular groups of students taking the test and no questions will likely favor one group of students taking the test over another group of students for non-educationally relevant reasons" (ESE, 2008, pp. 9–10). Examples of bias in item development are provided on p. 20 of this guide.

Construct Draft Versions of the Instrument and Administration Manual

Populate the instrument with items deemed to be of sufficiently high quality (based on the content, scoring, and bias reviews conducted earlier), and fill in the desired specifications in the Table of Test Specifications: number of items, variety of item types and difficulty, and the number of points appropriately spread to cover standards. Organize the instrument items so that they are accessible to the examinees. Divide the instrument into sections, if necessary, by grouping similar item types together. (One grouping that must remain together is when you have a set of instrument items associated with a single passage or prompt, for example.) Provide sufficient examinee directions for each instrument section. Review the entire instrument to determine the amount of time to allot to examinees, and if the instrument extends beyond a single administration period, decide how the instrument should be divided by administration periods. Finally, review the instrument to ensure the following:

- The material in the Table of Test Specifications is well represented with respect to content and rigor.
- There is a desired mix of intended item types.
- The instrument sections are accessible to examinees.
- Items using a single prompt appear on the same page as the prompt, or the prompt is viewable when responding to each item in the set.
- All instrument items appear on their own page (i.e., no item is split over two pages).
- Items are properly numbered. The numbers correspond with the scoring key and scoring guides.
- Item formatting is consistent through the instrument (similar fonts and font sizes, similar spacing, etc.).
- The directions for the entire instrument are clear, as are the directions for each instrument section.

It is important that instruments are administered the same way every time (i.e., that the assessment is standardized). To ensure such standardization, develop an administrator's manual that details the process for administering the test. Such manuals typically include information regarding the required characteristics of the testing environment (e.g., quiet, adequate space for students to write or perform), required materials for examinees (e.g., pencils, instruments such as calculators, etc.), if and how to time the instrument, the script to be read by the administrator, and information regarding allowable student accommodations (e.g., calculators, word-to-word dictionaries) and how they are to be managed. In addition, the administration manual can identify student accommodations allowable on the exam and the conditions under which examinees would qualify for using the accommodations. Typically, examinees may use accommodations that are habitually used for instructional purposes, provided the use of these accommodations does not interfere with the information learned in the test. (For example, examinees might not be allowed an accommodation of having a test "read" to them if the test is assessing reading fluency.)

Conducting a Pilot of the Draft Assessment

The next step in assessment development is to conduct a pilot or field test of the draft assessment.

Administer the Draft Assessment to a Representative Sample of Students
Try out the draft assessment by conducting a pilot or field (i.e., practice) test. Pilot testing the assessment and its administration manual in advance will serve many purposes:

- Provide data that can be used to evaluate the quality and the efficacy of the instrument items as well as the instrument overall
- Provide material and guidance for use on the development of scoring papers, scoring rubric(s), and scoring training materials
- Provide information regarding the appropriate amount of time to give students to take the assessment
- Identify challenges with the administration manual

There are several ways to pilot an assessment. One practical method is to administer an assessment developed for Grade 5 students to Grade 6 students at the beginning of the school year. This method could serve two purposes. First, it would serve as a test of the validity and fairness of the assessment items. Second, it would serve as a formative assessment for the Grade 6 students and teacher. Other approaches might include administering the instrument to target students (students in the identified grade, at the correct time of the year) and to consider the first administration the pilot assessment or piloting with students in another district. Collaborating across districts allows an opportunity to control overusage or overexposure of items, but it does increase the difficulty of conducting the pilot. Feasibility should always be a consideration in the development and use of DDMs.

ESE recognizes that it may not always be possible to conduct a formal pilot test for a locally developed assessment. However, some sort of "tryout" to check for fairness and accessibility is advised. For example, if a pilot is not feasible, districts can have the assessment reviewed by a second content expert who can review it from a fresh perspective. Weaknesses found prior to the instrument administration can thereby be redressed.

Evaluate Items for Quality

After the items from the pilot test have been scored, district teams can look closely at each item to evaluate its performance. Again, this is a multistep process that can only be highlighted here, but the basic process is to conduct both qualitative and statistical analyses on items to identify those that did and did not perform well. The following are among the specific indicators that an item is not performing well:

On selection items:

- Distractors that are not selected may indicate they are not plausible to examinees who do not know the tested content.
- Distractors selected by large numbers of high-scoring examinees may indicate that an item has two possible correct answers.
- If all or nearly all students get an item correct or incorrect, this suggests a problem with item difficulty that results in the item not discriminating between high and low performers. (However, on a pre-test, you may expect to see more items examinees found difficult, and on an end-of-year or posttest, you may find more items that many examinees got correct.)
- Large unexplainable differences in the performance of various subgroups on an item suggest that item may not be measuring relevant content.

On constructed response, performance, and portfolio item types:

- Poor-quality student responses may be an indication that the testing prompts are not adequate and need to be reworked or that the expectations for the task detailed in the rubrics for the item are not adequately communicated to examinees.
- Responses that are ancillary or only partially related to the prompt, again, may be the result of poor-quality prompts and administration procedures or rubrics that are poorly communicated to examinees.

Poorly performing items should be evaluated to determine (a) if the item is flawed and (b) the likely source of the problem. Then, if the source is known or suspected, the items may be edited and retested. Items that perform well can be retained and used.

Finally, for all types of assessments, the review team should examine item and test scale (test score) qualities by asking the following questions:

- Does the range of item difficulties span from easy to hard?
- Does the range of test scores span across the examinees so that the test shows no floor or ceiling effects?
- Are the items positively discriminating between students who know and do not know the material?

Evaluate the Scoring Method

If a scoring rubric was required, evaluate it by collecting student responses and dividing them into low, medium, and high “buckets.” Examine the sorted responses in the buckets against the rubric score-point descriptors. Identify specific student responses that represent each score point on the rubric. This is the final opportunity to adjust elements of the rubric and to ensure that it captures low, medium, and high responses accurately. After the rubric is edited and student responses are matched to the score points, the rubric must not be revised. This is important, as it establishes the standards of expectations that result in a student's score.

When pilot test data are available, it will be possible to use those data to help set the performance standard for the assessment. Setting the performance standard typically means identifying the score that represents a passing or acceptable score for students.

Evaluate Reliability

Determine which form(s) of reliability (e.g., internal consistency, test-retest, alternate forms, or inter-rater reliability) is most appropriate to establish, based on the characteristics of the assessment and the DDM context. Calculate the reliability coefficient for this type of reliability. Evaluate the size of the coefficient. If it is sufficiently large, then reliability has been established for the tested sample.

Evaluate Validity

Determine which form(s) of validity evidence (e.g., evidence based on content, evidence based on patterns of relationships, or evidence based on outcomes) is most meaningful based on the characteristics of the assessment and the DDM context. Note that evidence based on test development and content has already been established by the district by following the steps listed previously. Conduct any additional post-hoc analyses by establishing relationships (correlations) between the DDM and other measures or outcomes. If the correlations are sufficiently large, then evidence for validity related to the relationships with other variables has been established for the tested sample. Monitor the use of the assessment and the results to ensure that the benefits for students and educators continue to far outweigh any drawbacks associated with the assessment. Evidence collected from each of the processes (establishing validity through content/development, relationships with other variables, and consequential validity) can be collected to establish a level of validity for the instrument (judged from low to high).

Revise the Assessment and Construct the Final Version

Working through each of the steps listed in this section should have provided the district with a great deal of information about the assessment, including the items on the instrument, the scoring mechanisms, and the administration manual and its procedures. In this step, a district should revise the components of the assessment that were found to be lacking in the pilot administration. For example, any items that appear to have two correct answers should be removed entirely or edited to address the problem. The administration manual should be edited to capture natural speech more closely to prevent test administrators from deviating from it. The scoring rubrics, papers, and guides should be finalized.

Finally, revisit the purpose of the assessment (establish a baseline and select a diagnostic to measure growth) to ensure that it has been met.

Administering the Final Assessment

After the assessment has been finalized, begin using it operationally.

Establish Local Policies

Before administering the assessment, it may be necessary to establish district- or school-level policies to ensure best use and interpretation of assessment scores through the establishment of good score reports, standardization in administration, security, training of administrators, and other policies required to ensure that the assessment and its administration manual are kept up to date.

Communicate Results

After administering and scoring the assessments, information about student performance may be communicated to students and their parents as well as to teachers and other educators. These communications can take many forms; what is important is that they are appropriate for the audience. In addition, communications should use language regarding performance that is familiar within the ESE environment. (For example, ESE has established definitions for general performance, including Warning, Needs Improvement, Proficient, and Advanced.)[13]

Monitor Assessment's Use

After your district begins using an assessment operationally, it is important to reevaluate the instrument and the items periodically to ensure they are still performing adequately. Changes in things like the passing rate over time may indicate that instrument items have been compromised through overexposure. Psychometric quality should also be evaluated from time to time, as should alignment of the assessment with the curriculum.

[13] For more information about ESE's MCAS Achievement Level Definitions, see http://www.doe.mass.edu/mcas/tdd/pld/default.html.