



# ***WHAT IT TAKES TO DEVELOP AND IMPLEMENT STANDARDS-BASED ASSESSMENTS***

*Li-Ann Kuan, PhD  
American Institutes for Research*

*November 2011*



**USAID**  
FROM THE AMERICAN PEOPLE



Educational Quality Improvement Program  
Classrooms • Schools • Communities





# ***WHAT IT TAKES TO DEVELOP AND IMPLEMENT STANDARDS-BASED ASSESSMENTS***

*Li-Ann Kuan, PhD  
American Institutes for Research*

*November 2011*



**USAID**  
FROM THE AMERICAN PEOPLE



Educational Quality Improvement Program  
Classrooms • Schools • Communities



# Contents

Acknowledgments .....	v
Acronyms .....	vii
<b>Introduction: The purpose of standards and standards-based assessments in the classroom .....</b>	<b>1</b>
<b>Definition of standards.....</b>	<b>3</b>
<b>Things to consider when developing and implementing assessments .....</b>	<b>7</b>
There is a clear purpose for testing. ....	7
The assessments are aligned with the standards. ....	10
The test development process is technically defensible, and assessments produce valid and reliable test data. ....	14
The assessments are unbiased and administered and used fairly. ....	23
Performance standards are used to determine students' learning proficiency. ....	25
Test scores are organized in a manner that is useful. ....	31
The assessment results are used to guide policy analysis, programmatic decision making, instructional planning, and resource allocations. ....	33
What are some of the indicators of successes, challenges, and limitations of an assessment program? .....	37
<b>In conclusion .....</b>	<b>40</b>
<b>Suggested reading.....</b>	<b>41</b>
<b>References .....</b>	<b>42</b>



# Acknowledgments

---

USAID commissioned this document, *What It Takes To Develop and Implement Standards-Based Assessments*, through the Educational Quality Improvement Program 1 (EQUIP1), with the American Institutes for Research.

*What It Takes To Develop and Implement Standards-Based Assessments* was developed under the guidance of Yolande Miller-Grandvaux, current AOTR of EQUIP1, Pamela Allen, Director of EQUIP1 at AIR and Cassandra Jessee, AIR Deputy Director of EQUIP1. Feedback, excellent criticism and suggestions on improving the material have been provided by Markus Broer and Abdullah Ferdous. Editorial support was provided by Sue Bratten and design support was provided by the AIR Design Team.

EQUIP1: Building Educational Quality through Classrooms, Schools, and Communities is a multi-faceted program designed to raise the quality of classroom teaching and the level of student learning by effecting school-level changes. EQUIP1 serves all levels of education, from early childhood development for school readiness, to primary and secondary education, adult basic education, pre-vocational training, and the provision of life-skills. Activities range from teacher support in course content and instructional practices, to principal support for teacher performance, and community involvement for school management and infrastructure, including in crisis and post-crisis environments.



# Acronyms

---

AIDS.....	acquired immunodeficiency syndrome
CRQ.....	constructed-response question
DIF.....	differential item functioning
EFA-FTI.....	Education for All—Fast Track Initiative
EQUIP.....	Education Quality Improvement Program
HIV.....	human immunodeficiency virus
MCQ.....	multiple-choice question
MIDEH.....	Improving Student Education Achievement in Honduras (Mejorando el Impacto al Desempeño Estudiantil de Honduras)
NSAT.....	National Student Achievement Test
PIRLS.....	Progress in International Reading Literacy Study
PISA.....	Programme for International Student Assessment
SACMEQ.....	Southern and Eastern Africa Consortium for Monitoring Educational Quality, The
TIMSS.....	Trends in Mathematics and Science Study
USAID.....	United States Agency for International Development



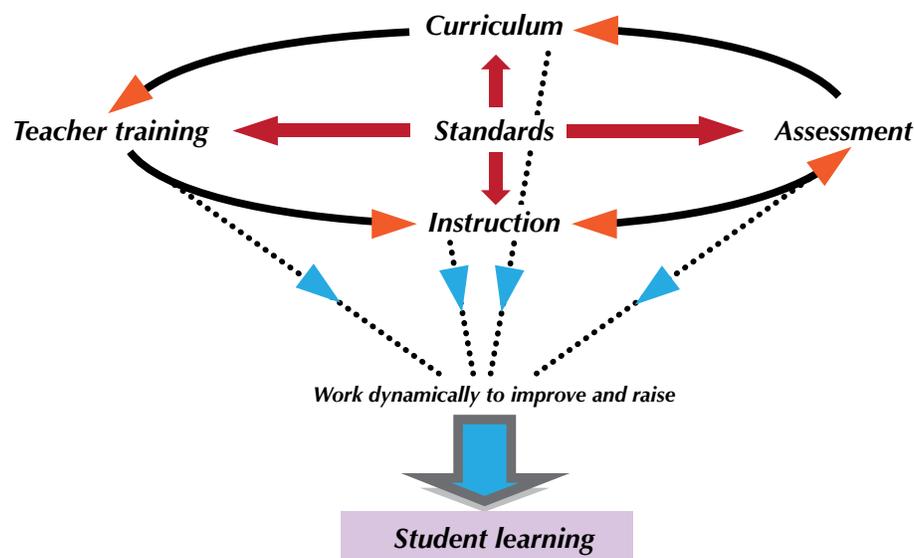
# Introduction: The purpose of standards and standards-based assessments in the classroom

Through the results obtained from international assessments, such as the Trends in Mathematics and Science Study (TIMSS), the Progress in International Reading Literacy Study (PIRLS), and the Programme for International Student Assessment (PISA), it is becoming increasingly apparent that standards and assessments play an important role in achieving high levels of student learning. Specifically, countries that have consistently performed well on these international tests have implemented an education system that comprises standards and assessments. This supports the findings from an earlier study conducted through the New Standards Project suggesting that the education systems of high-performing countries on these international assessments share two key features: (a) clear, consistent, demanding, and publicly articulated

academic goals and (b) regular, mandated programs for assessing student learning (Resnick & Nolan, 1995).

In simple terms, standards define the expectations for what students must know and be able to do at the end of a specific course of study (typically, at each grade level or grade span), and assessments (e.g., classroom-based assessments, large-scale examinations, national assessments of student learning) provide the means for measuring whether students are successfully learning the content delineated in the standards. However, standards and assessments by themselves are not sufficient to improve student learning. In a functional, standard-driven education system, the desired improvements in education depend greatly on the degree of alignment of the curriculum and

**Figure 1. Driving Improvements in Student Learning Through Standards**



---

instructional materials, teacher training, and assessment with the standards (Linn & Herman, 1997; Pellegrino, Chudowsky, & Glaser, 2001; Weiss, Knapp, Hollweg, & Burrill, 2001) and the systematic changes to each of these components (Linn & Herman; Goertz, Floden, & O'Day, 1995) (see Figure 1). Aligning the curriculum, instructional materials, and teacher training with the standards ensures that students have the maximum number of opportunities to learn the knowledge and skills outlined in the standards. With increased opportunities to learn, improved student outcomes will be reflected in the results derived from assessments that are aligned with the standards. In this way, assessments play a crucial role in monitoring improvements in student learning in any education reform effort. In fact, without assessments, it would be impossible to systematically measure the impact of such efforts.

# Definition of standards

---

Standards define student *learning goals*—that is, *what* students should understand and be able to do, and *how well* they should be able to demonstrate their understanding. Standards may be broken down into three components: content, process, and performance. Each component plays a specific role in ensuring a quality education for students. *Content* and *process standards* provide definition and clarity regarding what students are expected to know and be able to do during the course of the school year or grade span. Content and process standards also provide guidance to teachers about what they should teach in the classroom. Specifically, *content standards* identify the knowledge and concepts that students are expected to learn, and process standards define the cognitive skills and processes with which students are expected to demonstrate their knowledge and concepts (see Table 1).

Content *and* process standards are referred to differently in different countries. For example, in Namibia these standards are called *competencies* and in Pakistan *student learning outcomes*, while in the United States they are known as *content standards*. Other terms that have been used to describe content and process standards include *indicators*, *objectives*, and *expectations*. Collectively, content and process standards are also referred to as the *intended curriculum* or *learning frameworks*.

*Performance standards*, on the other hand, are established in conjunction with student assessments and define the levels of test performance that examinees are expected to attain relative to the content and process standards (Hambleton & Pitoniak, 2006). Performance standards describe *how well* students must perform on a test in order to be considered proficient learners. For example, in some contexts, students may be expected to answer at least 80% of the test questions—for example, 44 of 55—correctly to be considered advanced or excellent learners. Typically, student performance may be categorized in three or four *levels* of performance (see Tables 2a and 2b). In addition to providing a means for characterizing individual student learning, performance standards provide a basis for comparing all student performance against the same criteria when applied across an entire school system (National Academy of Education, 2009). Performance standards will be discussed in greater detail under First Principle #5 *Performance standards are used to determine students' learning proficiency*.

**Table 1. Examples of Content and Process Standards**

<b>Domain</b>	<b>Content standards</b> Define what students should understand and be able to do	<b>Process standards<sup>a</sup></b> Delineate the cognitive processes (such as reasoning, problem-solving, synthesis, and analysis) that students should acquire through their education
Operations and algebraic thinking	<p>Represent and solve problems involving multiplication and division.</p> <ol style="list-style-type: none"> <li>1. Interpret products of whole numbers, e.g., interpret <math>5 \times 7</math> as the total number of objects in 5 groups of 7 objects each. <i>For example, describe a context in which a total number of objects can be expressed as <math>5 \times 7</math>.</i></li> <li>2. Interpret whole-number quotients of whole numbers, e.g., interpret <math>56 \div 8</math> as the number of objects in each share when 56 objects are partitioned equally into 8 shares, or as a number of shares when 56 objects are partitioned into equal shares of 8 objects each. <i>For example, describe a context in which a number of shares or a number of groups can be expressed as <math>56 \div 8</math>.</i></li> <li>3. Use multiplication and division within 100 to solve word problems in situations involving equal groups, arrays, and measurement quantities, e.g., by using drawings and equations with a symbol for the unknown number to represent the problem.</li> <li>4. Determine the unknown whole number in a multiplication or division equation relating three whole numbers. <i>For example, determine the unknown number that makes the equation true in each of the equations <math>8 \times ? = 48</math>, <math>5 = ? \div 3</math>, <math>6 \times 6 = ?</math>.</i></li> </ol>	<ul style="list-style-type: none"> <li>• Make sense of problems and persevere in solving them.</li> <li>• Reason abstractly and quantitatively.</li> <li>• Construct viable arguments and critique the reasoning of others.</li> <li>• Model with mathematics.</li> <li>• Use appropriate tools strategically.</li> <li>• Attend to precision.</li> <li>• Look for and make use of structure.</li> <li>• Look for and express regularity in repeated reasoning.</li> </ul>

Note. From Council of Chief State School Officers (CCSSO) & the National Governors Association Center for Best Practices (NGA Center). (2010). Common Core Standards for Grade 3 mathematics. Retrieved November 10, 2011, from <http://www.corestandards.org/the-standards/mathematics/grade-3/operations-and-algebraic-thinking/>

<sup>a</sup> Referred to as “Mathematical Practice” in the Common Core Standards.

**Table 2a. Namibia’s Performance Standards–Description of Each Level of Performance**

<b>Below Basic Achievement</b>	The learner demonstrates <b>insufficient</b> knowledge and skills across all themes in the syllabus.
<b>Basic Achievement</b>	The learner demonstrates <b>sufficient</b> knowledge and <b>limited</b> skills across all themes in the syllabus.
<b>Above Basic Achievement</b>	The learner demonstrates <b>proficient</b> knowledge and skills across all themes in the syllabus.
<b>Excellent Achievement</b>	The learner demonstrates <b>excellent</b> knowledge and <b>advanced</b> skills across all themes in the syllabus.

From Ministry of Education & Directorate of National Examinations and Assessments, Namibia. (2010a). National Performance Standards for Grade 5 English. Windhoek, Namibia: Namibia Department of Education.

By delineating the knowledge and skills and levels of performance that students are expected to develop and demonstrate in a core subject, standards can provide guidance for educational practices within classrooms, schools, districts, and ministries of education, as well as help shape policies affecting curriculum, teacher development, assessment, and accountability (Weiss et al., 2001). When all educational practices and policies are aligned with a set of standards, they are more likely to operate seamlessly across the education system to achieve maximum student learning (Briars & Resnick, 2000; Ginsberg, Leinwand, Anstrom, & Pollack,

2005; Goertz et al., 1995; Linn & Herman, 1997). Therefore, in order for standards to effect change at the classroom level, they have to be “specific [and clear] enough to enable everyone (students, parents, educators, policy makers, and the public) to understand what students need to learn . . . [and] precise enough to permit a fair and accurate appraisal of whether the standards have been met (Linn & Herman, 1).”

**Table 2b. Expected Level of Performance for Each English Competency for Grade 5 Students in Namibia**

Theme	Topic	Competency <sup>a</sup>	Below Basic For example, a student who is considered BELOW BASIC	Basic For example, a student who is considered BASIC	Above Basic For example, a student who is considered ABOVE BASIC	Excellent For example, a student who is considered EXCELLENT
Reading and responding	01. Read intensively a range of texts across the curriculum, for example, reading texts on HIV and AIDS, Population, Education, Environmental Education, Human Rights and Democracy - for pleasure - for information - to complete a task - to give personal opinions	<b>1.01.01:</b> Predict outcomes	Cannot predict outcomes	Can make limited predictions based on simple texts	Can make some predictions using evidence from texts	Can locate and use words/phrases to support predictions and inferences
		<b>1.01.02:</b> Distinguish chronological order or sequence of events	Cannot distinguish chronological order or sequence of events	Can distinguish some sequential events but is unable to identify chronological order	Can distinguish most sequential events and chronological order	Can distinguish almost all sequential events and chronological order
		<b>1.01.03:</b> Identify main idea	Demonstrates little to no understanding of main idea	Understands the main ideas of simple text	Understands and can identify the main ideas of moderately difficult text	Understands and can identify the main ideas of complex text

<sup>a</sup> In Namibia, competencies are synonymous with content and process standards.

*Note.* Table 2a illustrates the overall achievement expected of grade 5 students in Namibia at each performance level (Below Basic Achievement, Basic Achievement, Above Basic Achievement, and Excellent Achievement) on the National Student Achievement Test (NSAT). These expectations are used to describe student performance on both the NSAT English and Mathematics. To produce Table 2b, Namibian educators applied the performance levels to the Grade 5 English and mathematics competencies. Essentially, educators defined for each competency the knowledge and skills students would have to be able to demonstrate at each level of performance.

From Ministry of Education & Directorate of National Examinations and Assessments, Namibia. (2010a). *National Performance Standards for Grade 5 English*. Windhoek, Namibia: Namibia Department of Education.

# Things to consider when developing and implementing assessments

The purpose of this paper is to highlight seven key principles that are critical to implementing and developing assessment programs, and must not be overlooked (see Table 3). All the principles presented in this paper are important for ensuring that the maximum amount of benefit is gained from an assessment program.

## There is a clear purpose for testing.

The most important stage in the development of a test is the design stage (Schmeiser & Welch, 2006), and the most important consideration

during the design stage is to determine test purpose or the intended uses of the test results (Cizek, 2009; Millman & Greene, 1989; Pellegrino et al., 2001). The test design stage is critical because this is the time when key decisions concerning test structure, item types, administration time, and use, as well as examinee population, are made (Schmeiser & Welch). Clear statements describing the purposes of a test allow test developers to create an overall framework for what the test will measure and to determine how best to go about developing the test. For example, during the test design stage, test purpose works hand in hand with the knowledge and skills delineated in the

**Table 3. Principles for Developing and Implementing Standards-Based Assessments**

Principles
1. There is a clear purpose for testing.
2. The assessments are aligned with standards.
3. The test development process is technically defensible, and assessments produce valid <sup>a</sup> and reliable <sup>b</sup> test data.
4. The assessments are unbiased and are administered and used fairly.
5. Performance standards are used to determine students' learning proficiency.
6. Test scores are organized in a manner that is useful.
7. The assessment results are used to guide policy analysis, programmatic decision making, instructional planning, and resource allocations.

<sup>a</sup> *Validity* refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999;p. 9).

<sup>b</sup> *Reliability* refers to “the consistency of measurements when the testing procedure is repeated on a population of individuals or groups” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999;p. 25).

standards, to determine *what* would be measured in the test and *how* these things will be measured.

Test purpose informs how inferences about the test results should be made and what inferences can be made about the examinee population. So, if the purpose of a test is to measure 10th-grade algebra knowledge and skills, the test scores cannot be used to make inferences about students' understanding of 9th-grade algebra or 10th-grade overall math skills. Table 4 presents more examples of the way test purpose influences decisions about test administration time, test delivery, test location, and test security.

There is not a universally accepted way for categorizing tests, since they may serve many different functions in various contexts or may be developed to meet multiple purposes (Schmeiser & Welch, 2006). Typically, tests are categorized according to the purposes they serve and may be classified as *diagnostic* versus *achievement tests*, *formative* versus *summative tests*, or *classroom* versus *large-scale tests*, or for *accountability* versus *instructional* purposes. These labels overlap considerably and are not always mutually exclusive; therefore it is possible for the same test to be listed under more than one category. There are two important thoughts to bear in mind when developing a test. First, it is crucial to be clear on what the intended uses of the test results are and to use the results accordingly. In other words, a test should not be used to meet a different purpose from the one that it was originally created for. Second, the more purposes a test is created to

serve, the less accurate the inferences from the test results will be (Pellegrino et al., 2001).

In light of the difficulties associated with categorizing assessments, for the purposes of this paper, assessments will only be discussed in terms of *formative* and *summative* assessments. In a nutshell, formative and summative assessments have different purposes. The former enables learning by providing early feedback, and the latter documents achievement (Shepard, 2006). Thus, formative assessments are often referred to as assessments *for* learning; while summative assessments have been called assessments of learning.

*Formative assessments.*<sup>1</sup> Results from formative assessments are used by teachers to gain a better understanding of what each student is learning successfully and where there are problems, and to adjust instruction accordingly. Therefore, for formative assessments to have an impact on learning, the results from these assessments need to provide effective feedback to students and teachers, engage students in their own learning, and provide insight into teaching practices (Broadfoot et al., 1999). Formative assessments are administered by teachers throughout the school year. These assessments

---

1 The term *formative assessments* is often used synonymously with *continuous assessments*. However, there is a distinction between formative assessments and continuous assessments. The term continuous assessments implies that assessments are administered continuously throughout the academic year. Hence, diagnostic, formative, and summative assessments may all be regarded as different types of continuous assessments.

**Table 4. Examples of the Way Test Purpose Influences Testing Decisions**

Decision about	If the test purpose to...	Then...	This requirement results in...
Test administration time	Measure students' mathematics knowledge and skills taught in a specific textbook chapter.	The test must assess only content taught in that chapter.	A test with narrow content coverage and fewer items, thus requiring <i>shorter time</i> for completion.
	Measure students' mathematics knowledge and skills at the end of the year.	The test must assess content taught throughout the year.	A test with broad content coverage and more test items, thus requiring <i>longer time</i> for completion.
Test delivery	Measure kindergarten students' skills in listening and speaking.	The test must allow students to demonstrate their oral skills.	A test that is administered by the teacher to the student <i>one-to-one</i> and the student responds <i>orally</i> to questions.
	Measure Grade 4 students' skills in reading comprehension.	The test must allow students to read a stimulus and respond to questions about the stimulus.	A test that is administered in a <i>group setting</i> on a <i>paper-and-pencil test</i> .
Test security	Measure knowledge and skills to determine students' graduation status.	To prevent cheating, the test items must not be circulated to students before the test; to ensure that the test items may be included in future tests, the items must not be circulated after the test.	<i>Higher test security</i> , whereby test items are kept in a secure location at all times.
	Help students understand what they have learned at the end of the chapter.	The test must not be circulated to the students before the test is administered, but is available to students afterward so that they may review their responses.	<i>Lower test security</i> , whereby students (and parents) may look at the test items and their responses after the test has been graded.

may be created informally and individually by teachers, and can be unique for each class or be created by groups of teachers at the district, regional, or national levels and common across classrooms. The decision of whether to employ common formative assessments rests on how the assessment results are intended to be used. One advantage of a common formative assessment is the clear link over the standards that it provides with the summative tests—that is, results in the formative tests will be indicative of how students will perform in the summative end-of-year test if no further improvements are made.

*Summative assessments.* Summative assessments are achievement tests intended to measure student learning after learning occurs. Specifically, summative assessments should measure students' achievement against a clearly defined set of standards. The results from these assessments are typically used in a variety of ways: to communicate to the public the academic proficiency of students on the national, district, or school level; to provide information for evidence-based decision making about student readiness for promotion and graduation, the effectiveness of the curriculum, the number of staff to hire, the goals of professional development, and budgetary needs at all levels of education (school, district, and ministries of education); and to assist students and parents with personal decisions and the setting of personal goals.

## **The assessments are aligned with the standards.**

Since assessments provide confirmation that students have learned what was taught in class, assessments may be regarded as a bridge between teaching and learning (William, 2010). However, to accurately determine whether students have learned the knowledge and skills outlined in the standards, the test needs to be aligned with the standards—that is, the questions on the tests need to be written so that they assess the targeted knowledge and skills (see Table 5). Thus, in addition to test purpose, a second crucial element needed for test development is a set of standards that define the domain (i.e., content knowledge and cognitive skills) to be measured on the test. Without a set of standards, test developers may create a test to suit a specific purpose but the test domain may be ill defined, haphazard, and/or ambiguous. If a test is intended to measure student achievement after learning has occurred in third-grade mathematics, then the test needs to be defined with respect to the content that has been taught. It would be impossible to evaluate an individual's overall understanding of third-grade mathematics if the test that is administered only contains questions pertaining to third-grade geometry. Tests that are written to assess student learning relative to a specific set of standards are commonly referred to as *standards-based assessments* or *standards-referenced assessments*.

**Table 5. Alignment Between the Standards and Test Questions**

Standard	Test question
<i>Content standard</i> Determine the unknown whole number in a multiplication or division equation relating three whole numbers.	In the equation below, the □ represents an unknown number. What number would □ have to be for the equation to be true? $3 \times \square \times 5 = 15$
<i>Process standard</i> Make sense of problems and persevere in solving them.	(a) 7 (b) 5 (c) 1 (d) 0

When standards and assessments are closely aligned, a high level of student performance on the assessments implies that students are learning the expected knowledge and skills. In contrast, test results indicating poor student performance suggest that students are not learning what they are supposed to, and *may* be indicative of other issues, such as poor alignment between the curriculum and instructional materials or teachers who lack the expertise and training to teach the knowledge and skills outlined in the standards. Thus, poor student performance should prompt a closer examination of the educational components, such as curriculum, classroom instruction, materials development, teacher professional development, supervision, management at schools, districts, and MOEs, and accountability.

The development of standards is typically a separate process that occurs prior to assessment development. Because the discussion concerning standards development is beyond the scope of

this paper, Table 6 has been included to provide a general idea of the activities involved in the development of content standards.

Typically, there are more standards to be measured than there is testing time; so, early in the development process, decisions have to be made about which standards to include on a test or how the standards should be rotated over the years to achieve optimal coverage. The decision of which standards to include in a test is made by assessment experts and content specialists together: The assessment experts provide the psychometric guidance to help determine which standards are measureable through a paper-and-pencil test (e.g., in a language paper-and-pencil test, standards that emphasize skills in speaking have to be omitted), while the content specialists provide the content expertise to identify and prioritize the standards that are important to measure on a test.

**Table 6. Development and Review Process for Learning Standards**

Step	Timeframe	Activity
1	2 months	Recruit a core writing team and a team of expert collaborators composed of content experts and respected leaders at the school and university levels from around the country.
2	1 month	Develop a purpose statement, philosophy, and outline for the standards.
3	1 month	Conduct a review of the purpose statement, philosophy, and outline for the standards by a group of diverse stakeholders, such as teachers, university teacher educators, university subject-matter experts, assessment experts, and instructional materials developers.
4	1 month	Use stakeholder feedback to revise the purpose statement, philosophy, and outline for the standards.
5	1 month	Make the revised purpose statement, philosophy, and outline for the standards available to the public for feedback and comment.
6	6 months	Develop an initial draft of the standards that responds to the revised purpose and philosophy statements and draws on high-quality standards developed in other countries; available research; and the experience of teachers, content experts, and leading thinkers on the subject.
7	2 months	Broadly disseminate the draft standards for review and comment, including specific requests for feedback from key stakeholders and stakeholder organizations.
8	1 month	Summarize and prioritize the feedback.
9	3 months	Revise and refine the standards in response to the feedback.
10	1 month	Recruit a validation team of content experts (who have not been part of the process ) from schools and local universities.
11	2 months	Disseminate the final draft for review by the validation team and by those who participated in the review of the purpose statement, philosophy, and outline for the standards, to build acceptance of and support for the final product.
12	2 months	Make any necessary final revisions to the standards.
13	2 months	Publish the standards and disseminate them to schools.
14	6 months	Conduct a dissemination and awareness campaign among educators, parents, and other stakeholders, to build familiarity with and support for the new standards.

Note. Taken from Kuan, L., Leinwand, S., Molotsky, A., Reeves, H., and Williams, C.H. (Draft, not for citation). *Learning Standards: What Matters Most for Quality Education*. Washington, DC: World Bank

**Table 7. Test Specifications for a Mathematics Formative Test in Grade 6**

Standards	# of test items	Cognitive skills		
		Knowledge	Comprehension	Application
<b>Number and number relations</b>				
Factor whole numbers into primes	4	1	1	2
Determine common factors and common multiples for pairs of whole numbers.	2		1	1
Find the greatest common factor (GCF) and least common multiple (LCM) for whole numbers in the context of problem-solving.	2			2
Multiply and divide by powers of 10 (e.g., $12.56 \times 100 = 1,256$ ).	2			2
Divide 4-digit numbers by 2-digit numbers with the quotient written as a mixed number or a decimal.	2			2
<b>Patterns, relations, and functions</b>				
Describe patterns in sequences of arithmetic and geometric growth and now-next relationships (i.e., growth patterns where the next term is dependent on the present term) with numbers and figures.	4	1	1	2
<b>Data analysis, probability, and discrete math</b>				
Collect, organize, label, display, and interpret data in frequency tables, stem-and-leaf plots, and scatter plots and discuss patterns in the data verbally and in writing.	4	1	1	2
Describe and analyze trends and patterns observed in graphic displays.	2		1	1
Calculate and discuss mean, median, mode, and range of a set of discrete data to solve real-life problems.	4	2	1	1
Create and use Venn diagrams with two overlapping categories to solve counting logic problems.	3		1	2
Use lists, tree diagrams, and tables to determine the possible combinations from two disjoint sets when choosing one item from each set.	2			2
Apply the meaning of equally likely and equally probable to real-life situations.	4		2	2
<b>Total # of test items</b>	<b>35</b>	<b>5</b>	<b>9</b>	<b>21</b>

*Note.* The test specifications above indicate the number of test items that have been included to measure specific mathematics standards. For example, in this mathematics test there are a total of four test items measuring the standard “factor whole numbers into primes” (see standard highlighted in pink). Of the four test items, one item measures the “knowledge” cognitive skill, one item measures the “comprehension” cognitive skills, and two items measure the “application” cognitive skill. Overall, the test specifications show that the test is to contain 35 test items, with 5 items measuring knowledge, 9 measuring comprehension, and 21 measuring application.

Together, test purpose and the standards allow for the development of the test specifications, sometimes also referred to as *test blueprints* (see Table 7). Test specifications provide guidance for the way current and future versions of the test should be constructed by describing the *content* (e.g., language, math, science, social studies), *form* (e.g., item-types: true–false, multiple-choice, and/or constructed-response questions), and *functional requirements* (e.g., statistical prerequisites for each question) of the test (Schmeiser & Welch, 2006). Test specifications may also outline the protocols for subsequent stages of item development (see Table 7), test review process, pilot testing procedures, test form assembly, and the evaluation of the end product. Overall, test specifications are important for maintaining the consistency of test forms created by different groups of individuals over time, and are an important part of development for formative and summative assessments.

If a country administers a regional or international assessment (e.g., The Southern and Eastern Africa Consortium for Monitoring Educational Quality [SACMEQ], PIRLS, TIMSS), an important and worthwhile activity to conduct is an *assessment to standards alignment activity*, to examine the degree to which the knowledge and skills tested on these regional or international assessments correspond with the country’s learning standards. If the degree of alignment is high, this means that the knowledge and skills expected of students throughout the country is consistent with what is being tested on the regional or international test. The opposite is true if the degree of alignment

is low, but results obtained from these tests must be interpreted with caution. It is important to mention here that a low degree of alignment does not necessarily mean that the standards or assessments are of poor quality; it merely points to the fact that the knowledge and skills taught in class are not the same as those being tested. Educators may want to reexamine the quality of the standards if they were not assembled following a systematic development process (such as the one described in Table 7) or with comprehensive input from experienced local educators (e.g., teachers, content specialists, university scholars) representing all regions of the country.

### **The test development process is technically defensible, and assessments produce valid and reliable test data.**

With the purpose for testing clearly established, and the knowledge and skills to be evaluated defined, and the test specifications in place, the development of test items is ready to begin. It is important to bear in mind that “sound test development depends on well-defined, *technically well executed item development and review processes. Sound item development is critical for providing the quality and consistency necessary to produce reliable test scores upon which validated test score-inferences can be made*” (Schmeiser and Welch, 2006, p. 324). *Text Box 1 provides a brief description of validity and reliability.*

### ***Text Box 1. What are Validity and Reliability, and Why Are These Concepts So Important to Testing?***

*Validity* refers to “the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 9). In other words, “to validate a proposed interpretation or use of test scores is to evaluate the rationale for this interpretation or use” (Kane, 2006, p. 23). Viewed from this perspective, it can be said that the entire process for developing test items is built around the collection of solid validity evidence to support the test score interpretations.

Another central concept to testing is reliability. *Reliability* refers to “the consistency of measurements when the testing procedure is repeated on a population of individuals or groups” (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999, p. 25). Like validity, reliability must be thought of relative to the intended purpose(s) of the test and its uses. However, reliability is concerned with the consistency or stability of test scores across repeated test administrations (Haertel, 2006).

Two examples that assessment experts often employ to illustrate the critical nature of validity and reliability to testing is room temperature and weight measurements.

*Validity.* If an individual wants to take the temperature of a large ballroom, he or she would not place the thermometer by the window under the direct exposure of the sun; nor would he or she place the thermometer directly under the air-conditioning vent located in the corner of the room. This is because neither of these locations would yield valid temperature measurements for the part of the room in which most people will gather—which is presumably in the center. Likewise, in order for a test to yield valid measurements about a group of individuals, its scores must be interpreted correctly and used appropriately.

*Reliability.* A weighing scale is thought to be *reliable* if the measures it produces of the same person are consistent over time. So, if the scale produces a reading of 180 pounds on Monday and Friday, 250 pounds on Tuesday, and 100 pounds on Wednesday and Thursday, it is safe to conclude that the scale is not reliable and produces weight measurements that fluctuate greatly from one day to the next. Likewise, a test that produces similar fluctuations in measurements of students’ abilities from one day to the next is considered unreliable.

---

*Recruiting item writers.* The first step in any item-development process is the recruitment of item developers on the basis of their qualifications and representativeness. A qualified item writer is one who is well versed in the specifics of the content and the appropriate level of difficulty of the intended assessment. Furthermore, these item writers should represent a range of demographic characteristics (e.g., geographic location, racial, ethnic, and gender backgrounds) similar to those of the student population. Therefore, when developing a large-scale assessment that is to be administered summatively at the national level, the item writers must be recruited from all regions of the country and must represent as many races and ethnicities as possible. Item writers should include a fair representation of men and women. This diverse representation is necessary to capture the unique viewpoints of each constituency. These perspectives are important for developing an assessment that accurately measure the knowledge and skills judged to be critical by each representative group (Schmeiser and Welch, 2006).

*Training item writers.* When recruitment is completed, the item writers must be trained to write items to meet the requirements of the test and item specifications, as well as a variety of other technical criteria of good item writing. The training process must be of high quality and consistently applied across subjects and item development efforts throughout the assessment program. Otherwise, the quality of the items will be poor and the survival rates of the items after field testing will be reduced—possibly increasing

the expenses associated with development and, ultimately, the cost of the overall assessment program and also lowering the quality of the assessments.

Generally, any training program associated with item development must include instructions on how to construct technically sound items. There is a set of procedures and considerations for developing multiple-choice items and constructed-response items put together by various test development experts. However, some of these considerations are context specific and may not apply to certain test development efforts. For example, because of the level of cognitive development, test developers have found that, when writing test items, it is best to phrase items as questions (e.g., *Why did John chase the cat?*) rather than as an incomplete sentence (e.g., *John chased the cat because \_\_\_\_\_*). As part of the training, test developers may want to include samples of exemplary test items, to provide item writers with an idea of what the end product should resemble, and to discuss the characteristics that make the item high quality. During the training, item writers are asked to submit samples of their work for evaluation. These items are reviewed for quality, and feedback about each item is provided to its writer, as needed. A writer may take as long as a week to become comfortable with writing items according to specifications. However, as noted above, since items ultimately determine the quality of the test, taking the time to train individuals to become good item writers is a worthwhile investment of time and resources. Writing assignments are given to item writers as

**Table 8. Criteria for Reviewing Items**

**Alignment with the standards**

Reviewers must make a judgment to ensure that

- The item measures an important aspect of the standard.
- The level of cognitive rigor is grade level appropriate.

**Item accuracy**

Reviewers must make a judgment to ensure that

- The content of the item is accurate.
- The language grade level is appropriate.
- All parts of the item are clear in meaning.
- The graphic, if any, is accurate and relevant to the item.
- The answer is correctly identified.
- The distracters for multiple-choice questions are plausible but clearly incorrect.
- The wording of the question is free of clues that indicate the correct answer.

**Freedom from bias**

Reviewers must make a judgment to ensure that

- The item is free from language, content, or stereotypes that might disadvantage or offend an individual.
- The item is fair to all individuals.

soon as they demonstrate that they are able to write items of adequate quality, which will later be refined through expert review (becoming a highly qualified item writer is a long process). Writing assignments are based on the number of items necessary for the construction of the operational test forms, which is in turn defined by the test specifications.

In addition to writing items, item writers should also develop item rationales for multiple-choice items and scoring rubrics for constructed-response

items.<sup>2</sup> Item rationales provide justifications for the key responses (correct answer) and reasons for why the distracters (incorrect responses) to the multiple-choice questions are incorrect. Item rationales serve as checks for potentially flawed items and should be included in the content review process. In situations in which the

<sup>2</sup> Constructed-response items are test items that do not require students to select an answer from a list of answer choices. In other words, constructed-response questions are questions that ask students to fill in the blanks, construct short response, or write essays.

---

content reviewers disagree with the justifications provided for the key responses, the item should not be used on a test without revision. Scoring rubrics are developed alongside the constructed-response items and provide a set of procedures and criteria for scoring student responses. Sound rubrics are consistent with the test purpose(s), define the characteristics of the response along a continuum, express the performance criteria in an understandable way, and account for a full range of performance that is consistent with the test purpose(s).

*Item review.* Once the new set of items is developed, each item should be reviewed for content accuracy, fairness, editorial style, and sound psychometric characteristics. All item reviews should be conducted by individuals who were not involved in the writing process for a particular item. And like the item writers, item reviewers should be knowledgeable of the content and performance expectations for students being tested at that grade level, as well as demographically representative of the diverse student population to be tested. During the content review process, reviewers are asked to evaluate the items for alignment with the standards, accuracy, and possible bias (see Table 8 for a list of item review criteria).

In addition to reviewing the items for alignment, accuracy, and bias, reviewers should provide suggestions for revising items in situations in which items lack clarity or require changes to improve technical quality. These changes may include rephrasing items that are unclear and

providing new incorrect alternatives responses. The second review process that items typically undergo is a fairness review or a bias review. This process is described in greater detail in the following principle, which discusses test fairness.

When all items have undergone content and fairness reviews, they proceed to editorial review. The editorial review consists of editing or producing the graphics associated with the items and proofreading the items to make certain that they read clearly and make accurate references to the graphic. The editorial process is an important one because it ensures that the same editorial standards are consistently applied to all items written for the assessment program (Schmeiser & Welch, 2006).

*Field testing, form assembly, and test review.* The new test items are ready for field testing after the item review process is completed. The purpose of field testing the items is to acquire information about them in order to evaluate their psychometric qualities, such as difficulty, discrimination, and bias. When a test item does not meet the minimum psychometric standards set by the test developer, the decision may be to delete it from the item pool or revise the item and pilot-test it a second time. The decision to delete an item or revise it depends on the interplay of several factors, such as the extent of the revisions required and the survival rate of items targeting the same standard. Because the item attrition rate could be fairly high, it is recommended that test developers field-test at least twice the number of items needed in an operational test form

---

(Schmeiser & Welch, 2006). However, this ratio may increase or decrease according to the quality of the item pool, the complexity of the items, and so forth.

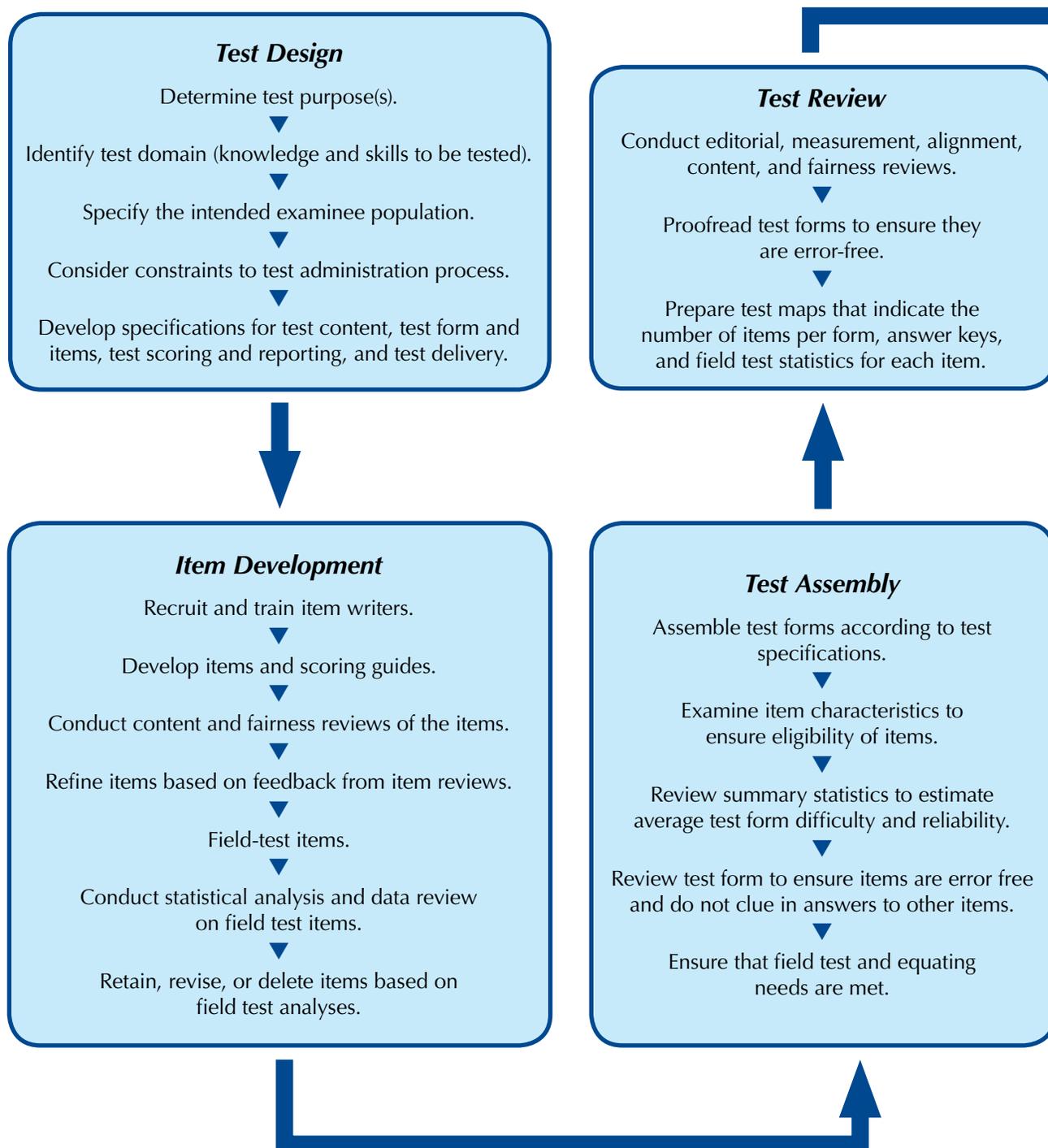
With the identification of a pool of test items that are statistically sound, test developers are ready to assemble the test form(s). In the assemblage of a test form, several things have to be considered. First, test developers have to decide which of all the items that survived field testing are best to include on the test form(s). This decision is made by reviewing the test specifications, each item's difficulty and reliability, and content.

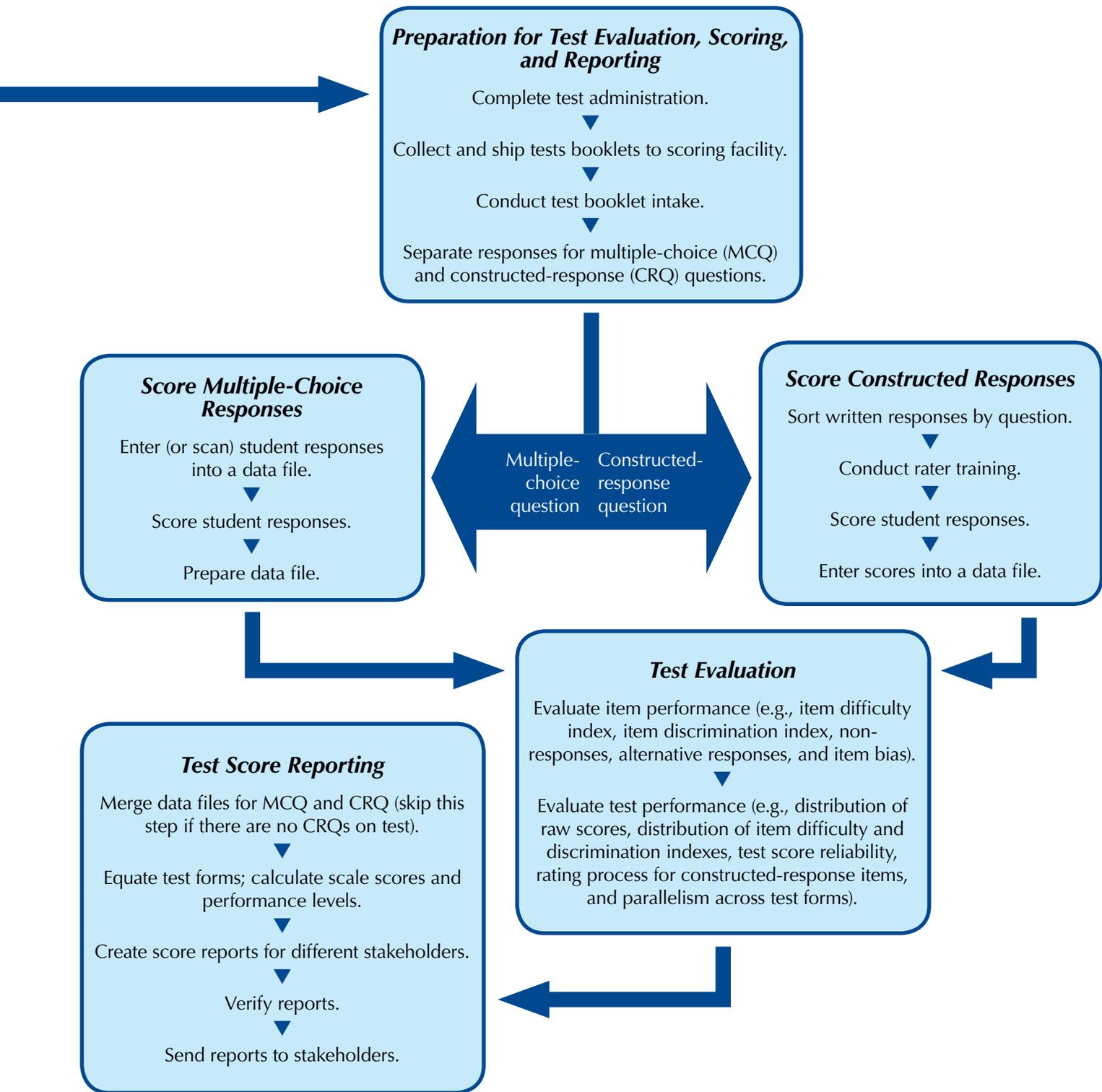
Second, if multiple test forms are employed for a single test administration to achieve comprehensive coverage of the knowledge and skills outlined in the standards. As mentioned previously, there are typically more content standards than can be realistically measured on a test within a reasonable timeframe. By administering multiple test forms, test developers are able to capture learning over a wider spectrum of standards. The other benefits of administering multiple test forms are that it limits opportunities for student cheating and reduces the need to replace an entire bank of items if one of the test forms is lost or stolen. When administering multiple test forms in a single test administration, test developers need to ensure that all the forms are equal in difficulty and content representation; otherwise, a test form may be more difficult or easier than other test forms and the results for students taking that test form may be artificially lower or higher than the results for the students taking the other

test forms. The process of making different test forms comparable in difficulty is referred to as test equating. To statistically equate test forms, test developers need to ensure that the items on each form are equal in difficulty, represent the same content, and share a subset of common items. All common items have to measure a range of standards and represent a range of difficulty levels, and appear in approximately the same position on every test form.

Third, the format of the test should be as simple as possible and should help students perform their best with little or no interference from factors that are irrelevant to the knowledge and skills being tested. If the test formats are too complicated with graphics and layout that are confusing and inaccessible to students, the test results will not be a true reflection of student abilities. In fact, test forms should be designed in such a way that valid inferences can be made about the abilities of the widest range of students possible (Thompson, Johnstone, & Thurlow, 2002). Some suggestions for creating a manageable test include (a) grouping all items written to the same standard or domain on the form, (b) placing any accompanying graphics or passages and items on facing pages so that students do not have to flip back and forth to respond to questions, (c) clearly indicating the general test directions and specific item directions so that students do not miss this information, and (d) designing the test form so that it is optimal for reading (e.g., adjusting the font size, line spacing, margins, and character spacing).

**Figure 2. Work Flow for Developing and Administering Assessments**





---

Finally, prior to sending the assembled test forms for reproduction, packing, and shipping, the forms must undergo a final series of reviews. The test form(s) must be reviewed by the test developer to ensure that the form(s) adhere to the technical requirements for equating and the characteristics laid out in the test specifications. The form(s) must also be reviewed by a content expert to ensure that there are no content-related issues or clue-ins on the test form, the items are grammatically correct and accurately worded, and each multiple-choice item has only one correct response. When all the content and measurement reviews are complete, the test form must be proofread to make certain that no errors were introduced into the items or test directions during the review process.

Storing test items in a commercially marketed electronic bank is a step which can occur at anytime during the test development process – after item writing or pilot testing. Item banking is also often a step that is often overlooked by test developers because uploading test items into the electronic bank and training users to operate the bank can be a time consuming process. However, these challenges may be overcome by selecting an electronic bank that has a user-friendly interface. There are several benefits to ensuring that items are properly stored in an electronic bank. First, test items that are centrally stored in an electronic bank have a lower chance of becoming lost or “misplaced”. Second, all test items and relevant item-level information (e.g., statistical information, item rationales) are stored in the same location. Many electronic item banks

allow users to upload statistical data about the item into the bank; so when an item is queried, users can see relevant statistical information about the item before deciding whether to include it on a test. Third, an electronic item bank provides safe storage of high-stakes test items that require more security. Electronic item banks that require passwords before access is granted limit the number of people with access to the items. Storing high stakes test items through a password-protected bank ensures that some students do not have access to the test items before other students prior to test administration. Fourth, electronic item banks facilitate test form construction. All electronic banks let users compile items by standards, which later allow for item searches or queries by standard. So, if users are interested in constructing forms with specific standards, all they would need to do is perform a search of all items that have been compiled under those standards, select items they want on the test form, and generate the test. Developing good quality items is an expensive and lengthy undertaking, which often involves hundreds of thousands of dollars over nine to twelve months. Therefore, it is important to properly store each test item that “survives” development so that none is lost because it was misplaced or accessed by over-zealous educators keen on helping their students perform well on the high stakes assessments.

The development process described above is fairly comprehensive. Some of the steps are compulsory, while others are discretionary. Some steps are compulsory for developing summative assessments but discretionary for formative

assessments. Thus, during the assessment design phase, test developers need to determine, on the basis of the test purpose, which steps are necessary. Figure 2 provides a brief overview of test development process and the test administration activities that follow after.

## **The assessments are unbiased and administered and used fairly.**

Not only must test scores be valid and the reliable, but conditions under which the tests are administered and the manner in which the results are used have to be fair. Test fairness is the “extent to which there is an absence of factors, unrelated to the intended purpose of a test, that advantage or disadvantage students” (Cizek, 2009, 10).” Fairness is an important concept in testing because the lack of fairness in testing threatens the valid interpretation of test scores. In other words, scores derived from a test that provides an unfair advantage to a specific group of students do not reflect the true abilities of these individuals. Test experts agree that there are two principal ways in which fairness can be compromised during testing: through biased test items and through the lack of equal treatment during testing or scoring.

To limit or eliminate bias in the test items, test developers typically conduct sensitivity reviews of the test items or run statistical analysis of the field test data. A sensitivity review is “a generic term for a set of procedures for ensuring (1) that

stimulus materials<sup>3</sup> used in the assessment reflect the diversity in our society and the diversity of contributions to our culture, and (2) that the assessment stimuli<sup>4</sup> are free of wording, and/or situations that are sexist, ethnically insensitive, stereotypic, or otherwise offensive to subgroups of the population” (Bond, Moss, & Carr, 1996, p. 121). On a superficial level, a sensitivity review may seem superfluous and unnecessary. However, since the main goal of a sensitivity review is to remove any distracting or offensive language or references on the test that may affect the stress levels of specific students and the way they respond to the test items, efforts to organize such a review become a necessary step for ensuring validity. Text Box 2 provides a brief description of putting together a sensitivity review.

The second method for detecting test bias is through statistical analysis. One statistical analysis that is commonly used is that of *differential item functioning* (or DIF analysis). DIF occurs when examinees of equal ability but with different group membership have unequal probabilities of success on an item (Angoff, 1993). For example, an item would exhibit DIF if it is far easier for boys to solve than it is for girls, when comparing boys and girls of the same ability levels established through their overall test result. A significantly lower performance by the girls than the boys does not necessarily mean that bias is present; it

---

3 Stimulus materials on a test may include, but are not limited to, text passages in a language test, diagrams, pictures, or graphics.

4 Assessment stimuli refer to test items, questions, or problems that elicit student responses.

### ***Text Box 2. Considerations for a Sensitivity Review Panel***

When developing items, item writers typically consider fairness and take care not to introduce wording and graphics that may bias the test item. However, as a rule, item writers should not provide a sensitivity review of their work. Generally, a sensitivity review is intended to be an independent review to identify language and biases that may make a test unfair. As a result, a panel of 5 to 10 reviewers (independent from the development process) are brought together to review the test material. Panelists may either be recruited or self-nominated for participation on the panel. These individuals typically have professional or instructional content expertise in the test area. In addition to content expertise, panelists must be able to understand and represent a range of diverse cultural and ethnic perspectives. Thus, oftentimes these panels are made up of men and women who represent different races, ethnicities, religious backgrounds, and geographic regions.

Camilli, 2006

does, however, mean that the item is performing statistically differently from other items on the test for these comparison groups and warrants a further review for *potential* bias. Under these circumstances, the test developer may choose to do one of two things. If the test developer has sufficient items to create the test form(s), then he or she may choose to remove the item from the

pool. If, however, not enough items are available to put a test form(s) together, then he or she may have the sensitivity review panel provide a recommendation on whether to keep the item as is because it is not biased, revise the item to rectify the bias, or delete the item from the item pool.

The lack of equal treatment among students during testing can also reduce test fairness. To limit the effects of unequal treatment of examinees, test developers will standardize procedures for test administration and scoring. In so doing, test developers ensure that all students take the tests in such a manner that the results for all students have “the same meaning across all forms and administrations” (Cohen & Wollack, 2006, p. 358). When testing conditions, directions, and scoring procedures are not applied consistently across test administrations for all examinees, there is no assurance that all individuals taking the test have the same understanding of what to expect. For example, if the test directions specifying the way responses should be indicated are clearly provided in one class but not another, examinees in the latter class may perform poorly on the test, not because they did not know the content but because they did not know how or where to write their responses to the test items. Therefore, to avoid introducing unequal treatment in testing, test developers must carefully consider the test directions to the teachers and students, the conditions for testing, and scoring procedures (see Text Box 3 for examples of what is typically included or considered in the directions for testing, conditions for test administration, and scoring protocols).

## Performance standards are used to determine students' learning proficiency.

The two salient features that make standards-based assessments different from other forms of testing (e.g., diagnostic tests, aptitude tests, achievement tests) are (a) the alignment between tests and content standards, and (b) the reporting of test scores according to performance standards (Koretz & Hamilton, 2006). Reporting student achievement against performance standards allows for comparisons to be made between student performance and a set of predetermined expectations. Reporting student performance using raw scores (e.g., 28 items correct out of 40 items) or percent correct is not a meaningful way to characterize student performance. In fact, experts argue that the best way to report test results in education systems that are trying to raise student achievement is through performance standards. Simply reporting student achievement as percent correct or as raw scores is not enough to raise performance because the level of test difficulty may be low to begin with; thus, even the students with the highest performance on the test may still be underperforming relative to what is considered acceptable or proficient (see Figure 3 for an explanation).

To categorize test takers in the various performance levels (see Figure 4), judgments have to be made by individuals with relevant professional and instructional content expertise about scores that define the upper and lower end of performance for each level. In other words,

### *Text Box 3. Test Directions, Conditions of Administration, and Scoring Procedures*

According to Cohen and Wollack (2006, p. 358),

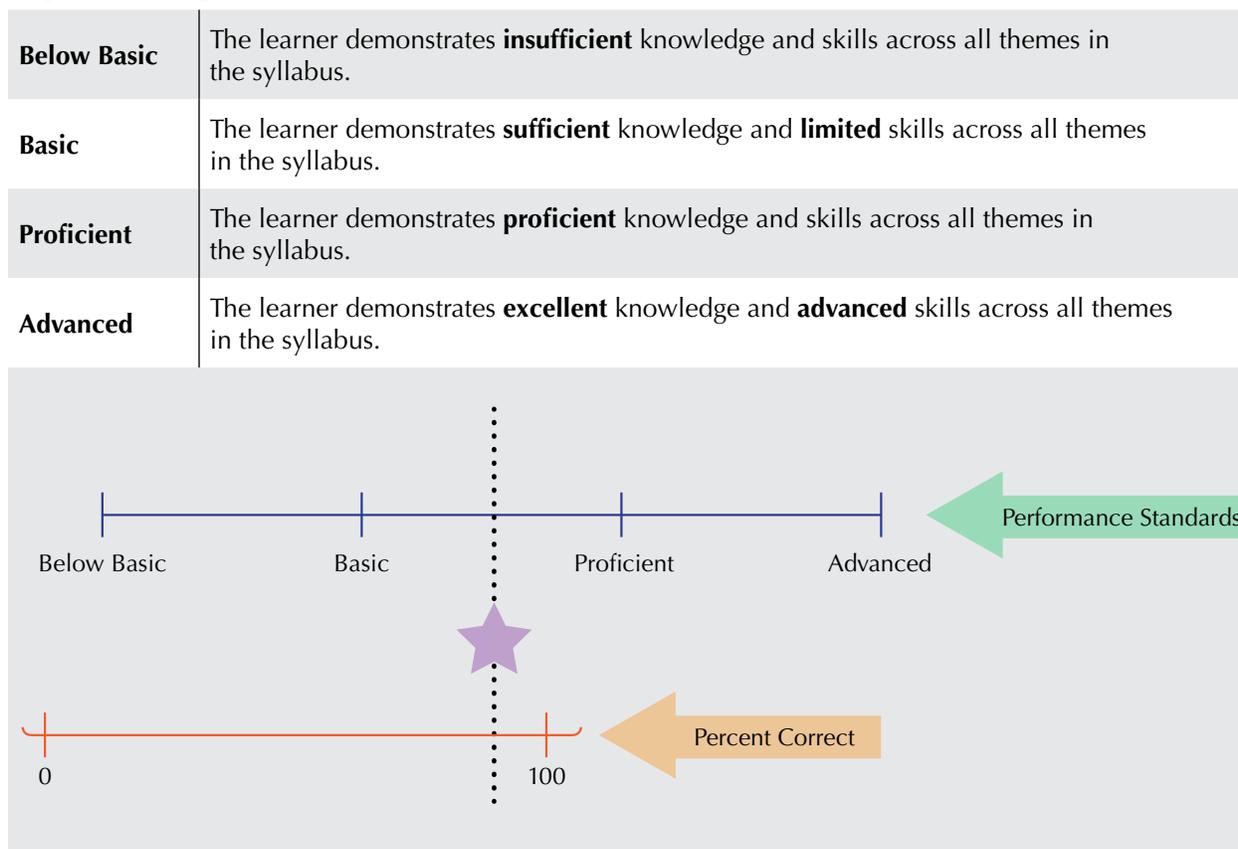
**Directions [for testing]** should normally include such things as how to answer the questions on the test, whether or not to write in the test booklet, whether ancillary materials such as a calculator are allowed, applicable time limits, how much help the proctor can be expected to provide, or whether guessing is discouraged (e.g., there is a penalty for guessing), encouraged (e.g., no penalty for guessing), or required.

**Conditions of [test] administration** include such things as method of administration of the test, training of administrators, training of examinees to respond appropriately, special instructions for registration, examinees check-in, room lighting and temperature, and seating.

**Scoring rules** can include machine or hand scoring, training of raters if appropriate, and impact of particular features of the response, such as spelling, handwriting, showing work, length, and so on.

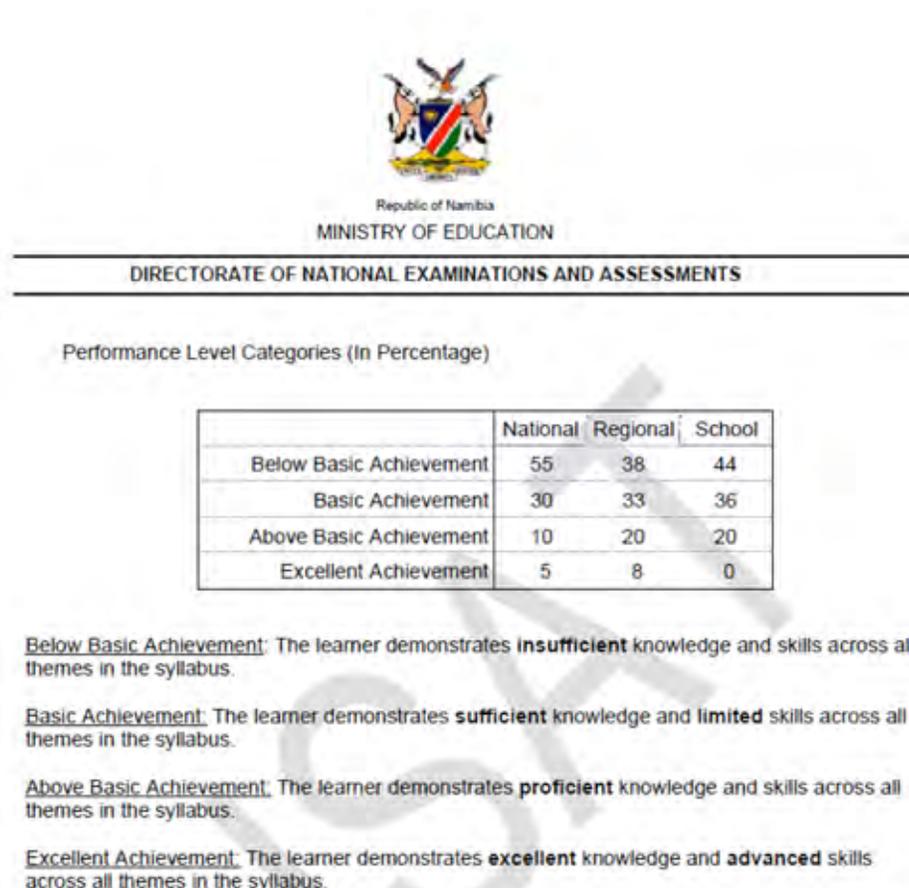
these content experts judge, through a systematic process, the minimum scores students need to attain to be placed in each performance level (e.g., Below Basic, Basic, Proficient, and Advanced).

**Figure 3. Using Performance Levels To Raise Student Achievement**



This figure illustrates the way the same student achievement (represented by the star) can be characterized differently when it measured against performance standards (represented in blue) and when it is measured as percent correct (represented in red). Against the performance standard, this student’s performance is classified as Basic; however, when reported as percent correct this student’s performance falls in the upper quartile. Therefore, if education officials want to raise student achievement, it is recommended that they create tests that measure what student must know and be able to do, and compare student performance against these expectations.

**Figure 4. A Section From the Grade 5 School Report for the Namibian English NSAT**



The example presented above is a section taken from the Namibia Grade 5 school reports for the English NSAT. The table summarizes the percentage of students at a school who achieved Below Basic Achievement status (44%), relative to the regional average (38%) and national average (55%). This table gives educators at the school and ministry an idea of student performance at this school in relation to that at other schools.

*Note.* From Ministry of Education & Directorate of National Examinations and Assessments, Namibia. (2010b). *Grade 5 school report for Namibian English National Student Achievement Test*. Windhoek, Namibia: Namibia Department of Education.

---

The scores that distinguish one performance level from another are referred to as *cut scores*. There are several methods for setting cut scores, including the Angoff method, the Bookmark method, and the Body of Work method; and each of these methods follows different procedures for establishing cut scores. There is no perfect method for setting cut scores, but some methods are more appropriate than others in specific circumstances (Zieky & Perie, 2006).

Although there are several ways of setting performance standards (also referred to as standards setting), the major steps that must be followed to set reasonable cut scores are consistent across methods (Zieky & Perie, 2006). First, it is important for policy makers and ministry officials to determine the performance levels (e.g., Pass/Fail, Below Basic/Basic/Above Basic/Excellent) to be reported. The decision regarding the type and number of levels of performance is typically reached through consensus. It is generally recommended that no more than three or four performance levels be established because, beyond four levels, it becomes increasingly difficult to distinguish among the levels. Next, descriptions of what students need to be able to do to reach each performance level must be developed. This is accomplished by convening groups of educators familiar with students in the targeted grades and the content area to describe the desired student performance. It is recommended that policy makers approve these performance levels before taking steps to set the cut scores for each level. In addition to writing descriptors for each performance level

for specific grades, it is important for education leaders to develop descriptors that apply to all grades for each performance level. Standardizing the performance levels across grades makes the expectations for performance consistent across grades, so that what is expected of Grade 4 students at the Basic performance level is consistent with the expectations for Grade 5 students at the same performance level. Going back to the example of performance standards in Namibia, which was presented in Table 2a (and will be presented again, below), the performance descriptors apply to all grades and across all subjects, while the performance descriptors in Table 2b are specific to each grade and subject.

*Provisional* cut scores are ready to be set after the performance-level descriptors are approved by education officials. As described by Zieky and Perie (2006), “performance level descriptors focus on what students should know and be able to do, [while] cut scores focus on *how many* score points students have to earn to demonstrate they have reached the level of knowledge and skill indicated by a specific performance level descriptor” (p. 5). A panel of between 10 and 15 judges is needed to set cut scores per subject per grade. These individuals can be the same content experts who developed the performance descriptors, but this is not a necessity. Although the different standards-setting methods employ different procedures for setting cut scores, all methods require the panel to envision the typical student at each performance level and use their expert judgment to determine the probability of this typical student’s answering each item on the test

**Table 2a. Namibia’s Performance Standards–Description of Each Level of Performance**

<b>Below Basic Achievement</b>	The learner demonstrates <b>insufficient</b> knowledge and skills across all themes in the syllabus.
<b>Basic Achievement</b>	The learner demonstrates <b>sufficient</b> knowledge and <b>limited</b> skills across all themes in the syllabus.
<b>Above Basic Achievement</b>	The learner demonstrates <b>proficient</b> knowledge and skills across all themes in the syllabus.
<b>Excellent Achievement</b>	The learner demonstrates <b>excellent</b> knowledge and <b>advanced</b> skills across all themes in the syllabus.

correctly. Therefore, all individuals recruited for standard setting *must* have a solid understanding of what students should know and be able to do following instruction in the topic area. Like all educators recruited for test development, all individuals on standards- setting panels should come from a variety of regions and represent the various subpopulations in their country.

When the provisional cut scores are set by the standards-setting panel, the next step is to establish operational cut scores. Establishing the operational cut scores consists of obtaining the approval of policy makers to apply the recommended cut scores to the actual student performance. While reviewing the cut scores, policy makers may adjust these scores one or two standard errors of measurements upward or downward to meet specific policy needs. By adjusting the cut scores upward, policy makers make the test more difficult, which lowers the pass rates. On the other hand, adjusting the cut scores downward lowers test difficulty and increases the pass rates. There are many reasons for policy

makers to choose to adjust the pass rates. For example, recognizing that tests are merely a “systematic sample of a person’s knowledge, skill, or ability” (Cizek, 2009, p. 10) and that all test inferences about what students know and are able to do are tentative,<sup>5</sup> depending on the level of rigor applied to the development process, it is not uncommon for policy makers to lower the cut scores to pass students who fall on the cusp of the pass/fail cut score (Zieky & Perie, 2006). Or in brand-new testing programs, by which an education system with poor student performance is introducing standards-based assessments for the first time, the pass rates may be disappointingly low. For the sake of getting local buy-in for a new assessment program, policy makers may, under this circumstance, choose to lower the cut scores so that fewer students will fall within the lower

<sup>5</sup> It is important to note that all test inferences about what students know and are able to do are tentative, with some being made more confidently than others—that is, the more items on a test to measure what students have learned, the greater the confidence concerning the inferences about what students know about that specific topic (Cizek, 2009, p. 10).

**Table 2b. Expected Level of Performance for Each English Competency for Grade 5 Students in Namibia**

Theme	Topic	Competency <sup>a</sup>	Below Basic For example, a student who is considered BELOW BASIC	Basic For example, a student who is considered BASIC	Above Basic For example, a student who is considered ABOVE BASIC	Excellent For example, a student who is considered EXCELLENT
Reading and responding	01. Read intensively a range of texts across the curriculum, for example, reading texts on HIV and AIDS, Population, Education, Environmental Education, Human Rights and Democracy <i>- for pleasure</i> <i>- for information</i> <i>- to complete a task</i> <i>- to give personal opinions</i>	<b>1.01.01:</b> Predict outcomes	Cannot predict outcomes	Can make limited predictions based on simple texts	Can make some predictions using evidence from texts	Can locate and use words/phrases to support predictions and inferences
		<b>1.01.02:</b> Distinguish chronological order or sequence of events	Cannot distinguish chronological order or sequence of events	Can distinguish some sequential events but is unable to identify chronological order	Can distinguish most sequential events and chronological order	Can distinguish almost all sequential events and chronological order
		<b>1.01.03:</b> Identify main idea	Demonstrates little to no understanding of main idea	Understands the main ideas of simple text	Understands and can identify the main ideas of moderately difficult text	Understands and can identify the main ideas of complex text

<sup>a</sup> In Namibia, competencies are synonymous with content and process standards.

*Note.* From the Ministry of Education & Directorate of National Examinations and Assessments, Namibia. (2010a). *National Performance Standards for Grade 5 English*. Windhoek, Namibia: Namibia Department of Education.

categories of performance. The same policy makers may choose to revert back to the panel-recommended cut scores after a few years, when the program is more established.

### **Test scores are organized in a manner that is useful.**

Essentially, the purpose of the test is a key determiner of the way the scores are reported. For example, if a test is intended to make decisions about whether a student is eligible for graduation from high school, then reporting the test results in terms of the student's Pass/Fail status would suffice. However, if the test is intended to provide educators with a profile of students' academic strengths and weaknesses, then reporting by Pass/Fail status would be inadequate. For the test results to be useful for instructional purposes, student performance needs to be reported analytically by content domain, such as themes or standards. Providing teachers with reports containing an overall test score, or a holistic score, as overall Pass/Fail status or performance levels does not offer sufficient diagnostic information about student performance for teachers to identify academic areas in which students require added support. Likewise, if the test is intended to provide ministry officials with guidance about resource allocation and program effectiveness, then the test score reports should be organized in a way that helps these educators answer these questions. Table 9 summarizes the different types of score reporting and the appropriate application of the information derived from specific types of scores.

The other considerations for designing score reports include the relation of the unit of reporting to the intended audience for the report (Fretchling, 1989). There are many ways to organize data so that they are meaningful for different audiences, and generally, it is beneficial to tailor the presentation of test results to audiences' specific interests or needs. For example, it is more appropriate for the test results to be reported at the student level in a report to a parent than to a school board member. A school report may also be appropriate for parents who want to know how their child's school performed relative to other schools in the region. Another consideration for designing score reports is the type of data to include. In other words, should comparison data for the unit of reporting be included in the report? Should a report for school principals include test results for their school alongside data from other schools in the same district or region, or alongside schools with similar student demographics from around the country? Would such comparison data allow principals to gauge how their schools are performing relative to other schools with similar characteristics? These examples illustrate how the audience and the information one wants to communicate determine the unit of reporting (e.g., student, classroom, school, district, or region).

Figures 4 and 5 illustrate score reports that serve very different purposes. Figure 5 is a score report for a nationally administered summative assessment, and Figure 6 presents an example of a score report for a formative assessment administered locally by teachers.

---

**The assessment results are used to guide policy analysis, programmatic decision making, instructional planning, and resource allocations.**

All resources and efforts to create a reliable test that provides valid scores, guarantee fairness throughout the testing process, ensure that test reports communicate results effectively, and set performance standards are lost if the test results are not used to guide decision making at all levels of the education system. Measuring student learning outcomes through assessments allows educational leaders to examine and monitor whether the system is operating as it should and achieving its intended learning goals. If the results indicate that there are still key areas of weaknesses, then decisions regarding the appropriate course of action must be made to strengthen these areas of poor performance. For example, if the test results show that students are performing worse in geometry than in measurement, then ministry officials may want to evaluate further the potential root cause(s) of this poorer performance. The lower performance on geometry could be due to a myriad or combination of factors, such as poorly designed instructional materials, the lack of teacher knowledge or instructional expertise in geometry, and the inaccessibility of materials and activities. Thus, test data, when organized effectively, can shed light on areas in need of more attention.

From the central level in the ministry of education to the schools and communities, a degree of technical understanding is necessary for the correct interpretation and use of data for improved educational services and support. Therefore, technical assistance and training from experts need to be available at all levels of education, so that teachers, principals, school administrators, policy makers, and community leaders understand how to interpret and use the data appropriately for effective decision making. For this to happen, there must be collaboration between the technical experts and educators, both within and across all levels of the system, to establish effective training systems that focus on providing guidance and support around data interpretation and use. It is also imperative that the direct training of these educators be concentrated not only at the ministry level but at the school level, so that teachers and principals are the direct recipients of the trainings. The maximum benefit of implementing a system of standards and assessment will be reached only if each level receives the training and support it needs for using data to make decisions.

**Figure 5. A Section From the Grade 5 School Report for the Namibian English National Student Achievement Test**



Republic of Namibia  
MINISTRY OF EDUCATION

---

**DIRECTORATE OF NATIONAL EXAMINATIONS AND ASSESSMENTS**

---

Report on the National Standardized Achievement Test (NSAT), 2009

Grade 5

School Demographic Information:

School Code:	School Name:
Rural/Urban: <b>Rural</b>	Region:
No. of Learners Taking the Test: <b>102</b>	

Percentage Correct Scores:

ENGLISH TEST	Cognitive Level	National	Regional	School
<b>THEME 1: Reading and Responding</b>		42	50	46
<b>Topic: Read intensively a range of texts across the curriculum</b>		45	51	47
Predict outcomes	Application	44	52	47
Distinguish chronological order or sequence of events	Comprehension	27	36	26
Identify main idea	Comprehension	34	45	47
Distinguish fact from opinion	Comprehension	45	50	44
Find basic information in texts	Comprehension	26	26	15
	Knowledge	54	62	59

Note. From a section of the Grade 5 school report for the Namibian English National Student Achievement Test (NSAT). The NSAT is a summative assessment that is administered at the end of the academic year to all Grade 5 students every other year. Namibia administers a similar assessment to students in Grade 7.

From Ministry of Education & Directorate of National Examinations and Assessments, Namibia. (2010b). *Grade 5 school report for Namibian English National Student Achievement Test*. Windhoek, Namibia: Namibia Department of Education.

The intended audience of this report is the school's principal, its leadership team, and teachers. The unit of reporting is the school. The test scores are reported by theme and by standard. Specifically, the report indicates the percent number of students who answered correctly the questions aligned with each theme and standard. This percentage is provided at the national, regional, and school level.

According to the report above, of the 102 students who took the test, the average percentage of items answered correctly for Theme 1: Reading and Responding at the school was 47%. This school's average is lower than the regional average of 51% but higher than the national average of 45%.

One of the inferences that readers can make from this report is that predicting outcomes is an area of relative weakness for Namibian students throughout the country. This is an area of instruction that needs to be strengthened. At the programmatic level, ministry officials may want to focus more teacher-training and/or strengthen instructional materials for this topic area.

**Figure 6. A Formative Assessment Diagnostic Class-Level Report**

Teacher:												
Subject:			Exhibit an understanding of the base 10 numbering system by reading, modeling, and writing whole numbers to at least 100,000; demonstrating an understanding of the values of the digits; and comparing and ordering the numbers.				Represent, compare, and order numbers to 100,000 using various forms, including expanded notation.		Multiply and divide numbers written in scientific notation.			
Student	Overall raw score											
	Test item	Overall % correct	1	2	3	4	5	6	7	8	9	
First name	Correct	resp.	b	d	a	b	d	c	a	b	d	
Jose	16	84%	b	d	a	!	d	c	a	b	d	
Michael	6	32%	b	d	!	!	"	c	a	#	d	
Kiara	10	53%	b	d	\$	!	"	c	a	"	d	
Lequite	7	37%	"	#	"	#	d	!	a	"	d	
Robert	18	95%	b	d	a	b	d	\$	a	b	d	
Ashley	10	53%	b	d	\$	!	d	c	!	"	d	
Dawan	10	53%	#	d	\$	b	d	c	!	#	d	
Stephanie	7	37%	b	d	\$	b	d	c	!	#	#	
Jon	10	53%	b	\$	a	!	"	c	a	#	d	
Myesha	8	42%	b	d	"	!	"	c	a	!	d	
Dashia	19	100%	b	d	a	b	d	c	a	b	d	
Kim	5	26%	#	d	!	!	d	c	\$	!	#	
Juan	7	37%	#	#	\$	!	d	c	!	b	d	
Kristina	11	58%	b	d	\$	b	d	\$	a	#	#	
Chloe	9	47%	b	d	a	#	"	c	!	#	d	
Ben	11	58%	b	d	!	#	d	c	a	"	d	
Patrice	7	37%	b	d	"	!	"	\$	a	"	d	
Anthony	6	32%	b	#	!	!	d	c	a	#	#	
Rasheeda	15	79%	b	d	a	!	d	c	a	!	d	
Naquan	7	37%	b	d	"	!	"	c	a	!	d	
Score	10	25%	80%	80%	30%	25%	65%	80%	70%	20%	80%	
% Correct by standard	54%				65%		50%					
Most common incorrect response	a	a	b	d	c	b	d	a	a			

The above report was designed to provide teachers with feedback on students’ performance on classroom formative assessments. This report is more detailed than the one in Figure 4. The primary audience for this latter report is teachers. The unit of reporting is at the student level. The report organizes test scores for each student by test item, grouped by standard. Student answers shaded in pink represent incorrect responses. At the bottom of the report is a summary indicating the percentage of students responding correctly to a test item, as well as the average percent

Standard									
Select and use appropriate operations—addition, subtraction, multiplication, division, and positive integer exponents—to solve problems with rational numbers, including negative rationales.	Select and use appropriate operations (addition, subtraction, multiplication, and division) to solve problems, including those involving money.			Select, use, and explain the commutative, associative, and identity properties of operations on whole numbers in problem situations, e.g., $37 \times 46 = 46 \times 37$ , $(5 \times 7) \times 2 = 5 \times (7 \times 2)$ .	Estimate and compute the sum or difference of whole numbers and positive decimals to two places.	Select and use a variety of strategies (e.g., front-end, rounding, and regrouping) to estimate quantities, measures, and the results of whole-number computations up to three-digit whole numbers and amounts of money to \$1,000 and to judge the reasonableness of estimates.			
10	11	12	13	14	15	16	17	18	19
d	d	b	c	b	d	c	a	d	c
d	d	b	!	b	d	c	a	d	!
d	d	#	!	"	#	c	a	"	!
d	\$	"	#	b	d	c	a	#	!
d	d	#	!	#	\$	c	a	"	\$
d	d	b	c	b	d	c	a	d	c
d	d	!	\$	b	d	c	!	"	\$
d	d	"	\$	b	"	c	a	#	!
d	\$	#	!	#	"	#	a	"	!
d	d	#	\$	b	d	#	a	"	!
#	\$	!	\$	!	d	c	a	#	\$
d	d	b	c	b	d	c	a	d	c
#	\$	b	!	"	#	\$	!	d	!
#	d	!	!	!	"	c	a	"	!
d	d	!	#	b	d	c	a	"	!
d	d	"	#	"	\$	c	a	"	!
d	d	#	#	b	#	c	a	"	\$
"	\$	#	#	"	d	c	!	d	!
"	"	!	!	!	#	#	a	d	!
d	\$	b	!	b	d	c	a	d	c
d	\$	#	!	!	"	c	!	"	\$
75%	60%	25%	10%	50%	50%	80%	80%	35%	15%
75%	32%			50%		80%	43%		
a	b	a	d	cd	ac	a	d	c	d

correct by standard. In the very last row, the report provides teachers with an idea of which distracters were most frequently chosen by his/her students. By examining the incorrect responses most frequently chosen by students, teachers have a better sense of the misconceptions that their students have. This information allows teachers to modify their instruction accordingly, to address this issue.

**Table 9. Types of Information Derived From Different Forms of Score Reporting**

	Type of score reporting	Information derived
	<p><i>Overall test score</i> Student's overall score (e.g., 30/40) or percent correct (e.g., 75%)</p>	<p>This score reporting provides information about a student's performance relative to the number of questions on the test. It does not provide diagnostic information—i.e., it does not indicate areas in which the student has performed well and areas in which he or she did not do well.</p>
Holistic score	<p><i>Pass/Fail status</i> Student's Pass or Fail status on the test</p>	<p>This score reporting provides information about a student's overall Pass or Fail status, an indication of whether the student has met the minimum requirement for his or her performance to be considered acceptable. This reporting method does not provide diagnostic information—i.e., does not indicate areas in which the student has performed well and areas in which he or she did not do well.</p>
	<p><i>Performance levels</i> Student's performance level (e.g., Unsatisfactory, Needs Improvement, Satisfactory, or Advanced)</p>	<p>This score reporting provides information about a student's overall performance level and how far the student is from a satisfactory/acceptable level of performance. This reporting method provides some detail about a student's performance but does not specify the content areas in which he or she performed well and those in which he or she did not do well.</p>
Analytic score	<p><i>Theme</i> Student's overall score or percent correct, by content theme (e.g., Number Sense, Operations, Geometry, Measurement, Statistics, and Probability)</p>	<p>This score reporting provides information about a student's performance by content theme and allows the teacher to understand student strengths and weaknesses by content theme, but does not specify the knowledge and skills he or she performed well on and those in which he or she did not do well. This score reporting does allow educators to plan for general programmatic improvements.</p>
	<p><i>Standard</i> Student's overall score or percent correct, by standard</p>	<p>This score reporting provides information about a student's performance by standard and specifies the knowledge and skills in which he or she performed well and those in which he or she did not do well. This score reporting allows educators to plan for specific programmatic improvements.</p>

# What are some of the indicators of successes, challenges, and limitations of an assessment program?

Based on the procedures and considerations described above, it is evident that establishing an assessment system is a complex process that requires a considerable investment of time, resources (human and financial), effort, and commitment (see a case study on the Honduran MIDEH project (Drury, 2011) a discussion on how standards-based assessments were implemented in Honduras). In light of the level of investments that are made in developing and implementing an assessment program, monitoring efforts must focus on indicators that gauge the success of the program (these indicators are presented in Text Box 5 and have been organized around each first principle).

The following list of challenges is intended not to be exhaustive but to give an idea of some of the common challenges and limitations experienced when developing and implementing assessment programs. For an in-depth discussion of challenges and limitations, please refer to the article by Kuan (2011). *“EQUIP2 Lessons Learned: Designing, Implementing, and Evaluating Programs Focused on Student Assessments: A Review of EQUIP2 Associate Awards in Egypt, Ghana, Honduras, and Namibia* (Kuan, 2011).

*Challenge #1.* When designing and establishing a new assessment program, ministry officials, donors, and test developers must ensure that there are *sufficient funds* to develop the assessments and sustain the program long term. The funds set aside for the program should cover the cost of developing the assessments, as well as recurring costs associated with administering

the tests in different academic years. In addition to development and administration costs, funds should be made available to provide training for educators in the interpretation and use of data. Until educators understand and know how to use test data to inform their daily practice and decisions, the greatest benefits for implementing an assessment program remain untapped.

*Challenge #2.* Ministry officials, donors, and test developers must ensure that sufficient time is set aside for developing the assessments. It is tempting to rush the test development process because of the needs of the education system. However, if the assessments are properly developed from the outset, they can provide high-quality feedback about student learning over time with few or no major revisions or edits. Typically, it takes at least a year to develop quality assessments (and more if standards need to be developed first).

*Challenge #3.* The organization (e.g., department within the ministry, institutions of higher education) assigned to manage the assessment program has to have the appropriate qualifications and/or experience to conduct the day-to-day assessment activities related to test development, administration, scoring, and reporting. There is a set of steps and procedures that needs to be followed to produce high-quality tests. Hence, if the organization assigned to manage the assessment program requires additional training in test development, administration, scoring, and reporting, efforts must be made to ensure that the organization receives the appropriate support from experienced assessment specialists.

#### ***Text Box 4. First Principle and Indicators of success***

##### ***There is a clear purpose for testing.***

- Key education stakeholders (e.g., policy leaders, education officials, district administrators, principals, teachers, students, parents, community) understand the purpose for and importance of testing.
- Key education stakeholders support the assessment program.

##### ***The assessments are aligned with standards.***

- There is a clear and explicit correspondence between each assessment item and the education standards.

##### ***The test development process is technically defensible, and assessments produce valid and reliable test data.***

- The test development process is well documented.
- The procedures used to develop the test items are coherent and allow for a gathering of evidence that bolsters the validity argument according to the specific purpose of the test.

##### ***The assessments are unbiased and are administered and used fairly.***

- Procedures to limit or eliminate test bias are built into the development and administration process.
- The results from the assessments are used in a manner that is responsible and fair.

##### ***Performance standards are used to determine students' learning proficiency.***

- The performance standards were established in a systematic manner by multiple educators (teachers, subject specialists) familiar with student performance in the target grades.
- The performance standards were reviewed and approved by education and/or policy leaders who considered the impact of their decision on student performance.

##### ***Test scores are organized in a manner that is useful.***

- The results are aggregated and disaggregated in ways that are meaningful to key education stakeholders.
- Test reports are clear to the audience they are intended for.

##### ***The results from the assessments guide policy analysis, programmatic decision making, instructional planning, and resource allocations.***

- Guidance is provided around the interpretation of the assessment data.
- The intended audience groups understand how to interpret and use the assessment data for short- and long-term decision making.

---

*Challenge #4.* The ministry of education must make a sustained commitment to align all components of teaching and learning with standards. As mentioned throughout this paper, assessment programs work in conjunction with other components of teaching and learning (e.g., curriculum, instructional materials, and teacher training) to improve the quality of education for students. It may be difficult to improve student learning if any of these components is not aligned with the standards.

*Challenge #5.* In designing the assessment program, ministry officials must be realistic about the types of assessments that are included in the program and ensure that the implementation timeline for these assessments is practical. While it is important to include multiple forms of assessments in an assessment program, an implementation plan that tries to roll out too many assessments at the same time can overwhelm the education system and undermine all efforts.

*Challenge #6.* If ministries of education decide to link rewards and sanctions to test performance, they need to be aware that this policy may inadvertently alter the way instruction is delivered in the classrooms. In classrooms where attaining high test scores is the main goal, teachers may begin to focus instruction only on concepts that are covered in the test. Teaching to the test is a practice that should be discouraged because tests typically assess only a subset of the knowledge and skills outlined in the standards.

# In conclusion

---

Assessments not only provide a systematic way to inform education stakeholders (e.g., students, teachers, parents, administrators, and the public) about student performance, but are also important drivers of change in an education system (Weiss et al., 2001). Assessments can play a critical role in communicating the learning goals of a school system, setting targets for teaching and learning, and transforming the behaviors and performance of teachers and students (Linn & Herman, 1997). The results obtained from assessments can be instrumental in shedding light on areas of weaknesses within a school system and bringing about the necessary improvements to teaching and learning, resource allocation, and teacher professional development (Weiss et al.). Over time, assessments help educators track the progress of individual students, classrooms, schools, and districts, as well as allow policy makers and the public to determine the effectiveness of programmatic decisions within an educational system (Pellegrino et al., 2001). For an assessment program to be effective and bring about the desired improvements to student learning, it needs to employ both formative and summative assessments. Both types of assessments need to be mutually aligned with the same learning goals and measure a broad range of tasks and problems that allow students to apply their learning in different contexts (Shepard, 2006).

# Suggested reading

Below is a list of suggested reading to assist readers who want to gain more insight to some of the topics discussed in this paper.

## About assessments in general

- Brennan, R. L. (Ed.), *Educational measurement* (4th ed.). Westport, CT: American Council on Education and Praeger Publishers.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.

## About summative assessments

- Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for national assessment of educational achievement*. Washington, DC: The World Bank.
- Greaney, V., & Kellaghan, T. (2008). *Assessing national achievement levels in education*. Washington, DC: The World Bank.
- Drury, B. (2011). *USAID/EQUIP1 Honduras Education for All—Fast Track Initiative (EFA-FTI): MIDEH case study*. Washington, DC: USAID.
- Kellaghan, T., Greaney, V., & Murray, T. S. (2008). *Using the results of a national assessment of educational achievement*. Washington, DC: The World Bank.

- Kuan, L. (2011). *EQUIP2 lessons learned: Designing, implementing, and evaluating programs focused on student assessments: A review of EQUIP2 Associate Awards in Egypt, Ghana, Honduras, and Namibia*. Washington, DC: USAID.

## About formative assessments

- Stiggins, R. (2008). *An introduction to student-involved assessment for learning* (5th ed.). Upper Saddle River, NJ, and Columbus, OH: Pearson and Merrill Prentice Hall.
- Popham, W. J. (2011). *Classroom assessment: What teachers need to know* (6th ed.). Upper Saddle River, NJ: Pearson.

## About using assessment data

- Bambrick-Santoyo, P. (2010). *Driven by data: A practical guide to improve instruction*. San Francisco: Jossey-Bass

# References

- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, D.C.: American Research Association.
- Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3–23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.
- Bond, L., Moss, P., & Carr, P. (1996). Fairness in large-scale performance assessments. In G. W. Phillips & A. Goldstein (Eds.), *Technical issues in large-scale performance assessments* (pp. 117–140). Washington, DC: National Center for Education Statistics.
- Briars, D., & Resnick, L. (2000). *Standards, assessments—and what else? The essential elements of standards-based school improvement*. Los Angeles, CA: CRESST/University of Pittsburgh.
- Broadfoot, P. M., Daugherty, R., Gardner, J., Gipps, C. V., Harlen, W., & James, M. (1999). *Assessment for learning: Beyond the black box*. Cambridge, UK: University of Cambridge School of Education.
- Camilli, G. (2006). Test fairness. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 221–256). Westport, CT: American Council on Education and Praeger Publishers.
- Cech, S. J. (2010). Test industry split over “formative” assessment. *Education Week Spotlight on Assessment*, 7–8.
- Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing concerns. *Theory Into Practice*, 48, 63–71.
- Cohen, A. S., & Wollack, J. A. (2006). Test administration, scoring, and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Drury, B. (2011). *USAID/EQUIP1 Honduras Education for All—Fast Track Initiative (EFA-FTI): MIDEH case study*. Washington, DC: USAID.
- Fretchling, J. A. (1989). Administrative uses of school testing programs. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 475–484). Phoenix, AZ: American Council on Education and Oryx Press.
- Ginsberg, A., Leinwand, S., Anstrom, T., & Pollock, E. (2005). *What the United States can learn from Singapore’s World-Class Mathematics System (and what Singapore can learn from the United States): An exploratory study*. Washington, DC: American Institutes for Research.
- Goertz, M. E., Floden, R. E., & O’Day, J. (1995). *Studies of education reform: Systemic reform*. New Brunswick, NJ: Consortium for Policy Research in Education.

- Haertel, E.H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 65–110). Westport, CT: American Council on Education and Praeger Publishers.
- Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Koretz, D. M., & Hamilton, L. S. (2006). Testing for accountability in K–12. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Kuan, L. (2011). *EQUIP2 lessons learned: Designing, implementing, and evaluating programs focused on student assessments: A review of EQUIP2 Associate Awards in Egypt, Ghana, Honduras, and Namibia*. Washington, DC: USAID.
- Linn, R., & Herman, J. (1997). *A policymaker's guide to standards-led assessment*. Los Angeles: National Center for Research on Evaluation, Standards and Student Testing.
- Millman, J., & Greene, J. (1989). The specification and development of tests of achievement and ability. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 335–366). Phoenix, AZ: American Council on Education and Oryx Press.
- National Academy of Education (2009). *Standards, assessments, and accountability*. Education policy white paper. Washington, DC: National Academy of Education.
- Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. Washington, DC: National Academies Press.
- Resnick, L. B., & Nolan, K. (1995). Where in the world are world-class standards? *Educational Leadership*, 52(6), 6–10.
- Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Shepard, L. A. (2006). Classroom assessment. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 307–353). Westport, CT: American Council on Education and Praeger Publishers.
- Thompson, S. J., Johnstone, C. J., & Thurlow, M. L. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved November 10, 2011, from <http://education.umn.edu/NCEO/OnlinePubs/Synthesis44.html>

---

Weiss, I. R., Knapp, M. S., Hollweg, K. S., & Burrill, G. (2001). *Investigating the influence of standards: A framework for research in mathematics, science, and technology education*. Washington, DC: National Academies Press.

Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessments* (pp. 18–40). New York: Routledge.

Zieky, M., & Perie, M. (2006) *A primer for setting cut scores on tests of educational achievement*. Princeton, NJ: Educational Testing Service. Retrieved November 10, 2011, from [http://www.ets.org/Media/Research/pdf/Cut\\_Scores\\_Primer.pdf](http://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf).







**USAID**  
FROM THE AMERICAN PEOPLE



AMERICAN  
INSTITUTES  
FOR RESEARCH®



*This report is made possible by the generous support of the American people through the United States Agency for International Development (USAID). The contents are the responsibility of the American Institutes for Research and do not necessarily reflect the views of USAID or the United States Government.*

U.S. Agency for International Development Cooperative Agreement

No. GDG-A-00-03-00006-00

© 2011