

# **WASHINGTON STATE DIAGNOSTIC ASSESSMENT GUIDE**

**JANUARY 2009**

PREPARED BY:

JOE STEVENS, PH.D.

MESA

UNIVERSITY OF OREGON

UNDER CONTRACT WITH THE OFFICE OF THE SUPERINTENDENT OF PUBLIC INSTRUCTION



## TABLE OF CONTENTS

EXECUTIVE SUMMARY .....	1
INTRODUCTION AND PURPOSE OF THE <i>DIAGNOSTIC ASSESSMENT GUIDE</i> .....	7
OVERVIEW OF THE WASHINGTON STATE LEGISLATION .....	8
ASSESSMENT PURPOSES .....	9
NORM REFERENCED, CRITERION REFERENCED, AND STANDARDS- BASED ASSESSMENT TOOLS AND PURPOSES .....	9
STANDARDIZATION IN ASSESSMENT .....	12
FORMATIVE AND SUMMATIVE ASSESSMENT PURPOSES .....	13
SUMMATIVE ASSESSMENT .....	13
FORMATIVE ASSESSMENT .....	14
SCREENING ASSESSMENT .....	17
DIAGNOSTIC ASSESSMENT .....	18
INTERIM ASSESSMENT .....	19
PROGRESS MONITORING .....	18
SUMMARY OF ASSESSMENT PURPOSES .....	21
COMBINING DIFFERENT ASSESSMENT PURPOSES .....	22
FORMATIVE ASSESSMENT PROCESSES: BACKGROUND AND FINDINGS	
FROM THE RESEARCH .....	24
THE IMPORTANCE OF FEEDBACK .....	25
INFORMAL ASSESSMENT AND CLASSROOM ASSESSMENT .....	26
QUESTIONING .....	27
OBSERVATION .....	27
PEER AND SELF ASSESSMENT .....	27
DESCRIPTIVE FEEDBACK .....	28
DIAGNOSTIC ASSESSMENT FOR STUDENTS WITH SPECIAL NEEDS .....	28
DIAGNOSTIC ASSESSMENT FOR STUDENTS IN SPECIAL EDUCATION PROGRAMS .....	28
DIAGNOSTIC ASSESSMENT FOR ENGLISH LANGUAGE LEARNERS .....	29
ASSESSMENT ACCOMMODATIONS .....	30
TEST DESIGN FOR STUDENTS WITH SPECIAL NEEDS .....	32
ISSUES IN THE USE AND INTERPRETATION OF DIAGNOSTIC ASSESSMENTS .....	33
CRITERIA FOR CHOOSING AN ASSESSMENT .....	33
TECHNICAL QUALITY .....	34
EVIDENCE FOR RELIABILITY .....	34
EVIDENCE FOR VALIDITY .....	36
TEST FORMS, SCORES, AND REPORTS .....	37
BIAS AND SENSITIVITY REVIEW .....	38

## TABLE OF CONTENTS (CONTINUED)

IMPLEMENTATION, USE AND INTERPRETATION . . . . .	39
DIFFICULTIES, PROBLEMS, AND PITFALLS . . . . .	39
SUGGESTIONS FOR SUCCESSFUL IMPLEMENTATION . . . . .	40
CONCLUSION . . . . .	42
REFERENCES . . . . .	43
APPENDIX: RESOURCES FOR EDUCATORS INTERESTED IN FORMATIVE ASSESSMENT . . . . .	48

## EXECUTIVE SUMMARY

This summary provides a brief overview of the issues discussed in the remainder of this guide. The purpose of *The Washington State Diagnostic Assessment Guide* is to provide Washington State educators with information that will support the selection, use, and interpretation of formative and diagnostic assessments. Recent legislation in Washington provides support for educators to purchase and use diagnostic assessments. This report provides a clear definition of the concept of assessment as well as background and general information on formative and diagnostic assessment including:

1. A brief review of the Washington state diagnostic assessment legislation (ESHB 6023).
2. Definitions of different assessment purposes and how they relate to diagnostic assessment.
3. A description of some of the major findings from the research on formative and diagnostic assessment.
4. A discussion of the policy issues related to the implementation of diagnostic and formative assessment processes as well as the use and interpretation of assessment results.

The 2007 Washington State Legislature appropriated \$4.8 million to school districts to purchase and implement diagnostic assessments during the 2007-2009 biennium. School districts were eligible to receive \$5 per student for the purchase and implementation of diagnostic tools during the 2007-08 school year. During the 2007-2008 session, the legislature changed the way the original \$4.8 million were to be used. Approximately \$2.3 million were to be allocated to districts for purchasing and administering diagnostic assessments. The remaining \$2.5 million were to be used to develop and implement diagnostic and formative assessments. During the 2007-2008 school year, approximately \$1.8 million of the \$2.3 million were distributed to 116 of the State's 295 school districts, based on their iGrants applications for and proposed use of, the diagnostic assessment tools and results. School districts that receive funding are to report whether or not they expended the funds; remaining funds must be spent on diagnostic assessment purchase and use in future years.

ESSB 6023 defines a ***diagnostic assessment*** as an assessment that helps to improve student learning, identifies academic weaknesses, enhances student planning and guidance, and develops targeted instructional strategies to assist students before the high school WASL. According to the legislation, to the ***greatest extent possible*** the assessments tools had to be:

- a) aligned to the State's grade level expectations;
- b) individualized to each student's performance level;
- c) administered efficiently to provide results either immediately or within two weeks;
- d) capable of measuring individual student growth over time and allowing student progress to be compared to other students across the country;
- e) readily available to parents; and
- f) cost-effective.

The legislation also authorized the preparation of this *Washington State Diagnostic Assessment Guide* and the development of a *Formative Assessment Comparative Guide* that identifies and provides information on commercially available formative and diagnostic assessment instruments. This work was carried out by Measurement, Evaluation, and Statistical Analysis (MESA) Associates. Questions about the Comparative Guide should be addressed to Joseph Stevens, [jstevens.mesa@comcast.net](mailto:jstevens.mesa@comcast.net).

Throughout this *Guide*, the term *assessment* takes on a broad array of meanings. The term might refer to a particular *assessment tool*, such as the *Early Diagnostic Mathematics Assessment* (EDMA). The term is also used to describe *assessment results* (scores, reports, and descriptive information) derived from students' responses to an assessment tool. The term *assessment* may be used to refer to an *event* such as screening at the beginning of a school year. Finally, the term may be used to refer to an *assessment process* – using assessment tools to gather assessment information as well as summarizing, interpreting, and acting upon information obtained from one or more assessment tools. Given the variety of meanings throughout this guide, we indicate whether or not we are discussing an assessment event, an assessment process, an assessment tool, or assessment results.

In addition to the array of meanings of *assessment*, there are many assessment purposes. This *Guide* defines each of them so that the purposes of *diagnostic* and *formative* assessments can be distinguished from the purposes of large-scale tests, interim assessments, etc. Educators must be clear about their needs so they can select one or more assessment tools that provide the information they need. When educators are clear about their assessment purposes, they are more likely to use the assessment results in a process that helps them achieve their goals. Finally, if educators are clear about their purposes, they are more likely to time assessment events so that results are available when needed.

The Appendix gives resources for two major assessment purposes – formative and diagnostic – with diagnostic assessments being a subcategory of formative assessments. This *Guide* does not describe or suggest instructional interventions, even though it is well recognized that a strong link between assessment and instruction is a key component of educational effectiveness. This *Guide* does not describe or endorse specific assessment tools. There is a companion report in two parts: *The Formative Assessment Comparative Guide – Consumer Report* and the *Formative Assessment Comparative Guide – Technical Report*. These *Comparative Guides* provide information on most commercially available assessment tools in mathematics, reading, science, and writing for grades K-12. The consumer report provides quick summary of the purpose of the assessment, a summary of the focus of the assessment, contact information for the publisher, costs, and a technical rating. The technical report provides detailed information regarding content assessed, information on evidence for reliability and validity of the tests, and additional details on scores, reporting, and administration procedures. These *Comparative Guides* are intended to help teachers, schools, and districts select the most appropriate tools for their assessment purposes.

Research on the use of formative assessment processes shows positive impacts on a number of aspects and outcomes of educational practice, including: a) increases in student motivation and attitude, b) improved student attention, and c) more active and deeper learning. One of the most important results from the research on formative assessment processes is the finding that regular use of a formative assessment process results in substantial gains in student achievement (Black & Wiliam, 1998b). Many studies have found that the use of a formative assessment process improves achievement for all students, sharply increases the performance of lower achieving students, and narrows the achievement gap between lower achieving and higher achieving students.

This *Diagnostic Assessment Guide* also defines four specific formative assessment purposes: screening, diagnosis, interim measurement, and progress monitoring. Although some authors consider these to be distinct, we consider them as subcategories of formative assessment. The purpose of **screening** is to make an early identification of a student's strengths or weaknesses to allow classification, placement, or intervention. Screening assessment tools are designed to rapidly identify those individuals who need specific placement, attention, or instructional intervention. **Diagnosis** is another subcategory of formative assessment – designed specifically to identify the causes of student weaknesses, usually with intent to guide or modify instruction or to design differentiated instruction. **Interim measurement** takes place two or three times per year to determine where students are in relation to achievement of specific academic standards. Finally, **progress monitoring** is a specific type of interim assessment event, characterized by frequent, repeated use of assessment tools, to determine whether or not students are responding well to particular instructional interventions. Progress monitoring is usually conducted in conjunction with the delivery of an instructional intervention so that the student's response to intervention can be observed and evaluated.

In addition to the definition of formative assessment purposes, we define the **summative assessment purpose** as evaluation for the purpose of judging performance at a particular point in time. Summative assessment instruments are primary tools in accountability testing and in efforts to evaluate the performance of students, schools, and states. Summative assessment events occur at or near the end of a course of study, a class or an instructional unit, or a school year, rather than during the period of instruction. Summative assessment results are inherently evaluative and typically express results as grades, judgments of proficiency, or measures of attainment. Summative assessments are generally high stakes events, often being used to determine eligibility for matriculation to the next grade, graduation, or other significant decisions.

The timing of assessment events is one key difference between formative and summative assessment purposes. While summative assessment events occur at the end of an instructional period, formative assessment events occur before and during the instructional process. Formative assessment tools are designed to be more closely linked to learning and instruction; therefore, they are used more frequently and are interlaced with instructional activities. Another key

difference between summative and formative assessment purposes is the relative emphasis on evaluation or grading. While evaluation is at the core of summative assessment, there may be no evaluation or grading per se in a formative assessment process. Rather, information from formative assessment tools is used to provide feedback and guidance on learning in progress.

An important topic discussed in this *Diagnostic Assessment Guide* is the role of feedback in the assessment process. For summative assessment results, feedback is provided in the form of final evaluative judgments (e.g., a final course grade), which can include information about mastery and level of attainment. On the other hand, feedback that is directly linked to instructional change to improve student achievement is a distinguishing attribute of feedback from formative assessment events. Formative feedback provides immediate information to students and teachers, that focuses on how instruction can be adjusted to achieve improvement in student performance.

This *Guide* also presents a discussion of the use of diagnostic and formative assessment in the identification and instruction of students with special needs. This includes the use of progress monitoring methods to evaluate students' responses to interventions (RTI) to help in determining whether or not students need special education services. Diagnostic assessments play a critical role in the identification of students in need of special education services. Many students who struggle in academic content areas have inconsistent response patterns that make it difficult to diagnose causes using typical classroom formative, district interim, and state level assessments. To provide instructionally relevant information, well developed diagnostic and formative assessment tools can be used to more carefully determine whether or not students are learning targeted knowledge and skills and, if not, to determine sources of students' learning needs.

Formative and diagnostic assessment tools must be designed and administered in such a way that differences in language ability do not impede the evaluation of students' skills and content area knowledge. The key challenge in assessment of English language learners (ELLs) is making sure that the targeted knowledge and skills are being measured, and not some other aspect of language knowledge or language ability. It is recommended that the reading and language requirements of science, social science, and mathematics assessment tools be made as simple and accessible as possible. The use of simplified language in content area assessments has been shown to help both English language learners and native English speakers.

Any time an assessment is administered, some test-takers may have cognitive, sensory, physical, or language characteristics that interfere with interpretation of the assessment results. As such, test scores may not accurately reflect the student's understanding (or misunderstanding) in the domain. To ameliorate this problem, accommodations should be provided during assessment events. Accommodation decisions should be matched to the intended purpose of the assessment results. For example, if the assessment results will be used to predict later achievement and track student progress toward achieving the standards on state tests, the policies and methods used for accommodations should closely match those used for the state test. On the other hand, if the

purpose of the formative assessment is more directly focused on learning improvements, then greater flexibility in the choice and application of accommodations may be warranted.

Choice of an assessment tool is complex. The companion *Comparative Guides* provide suggestions and recommendations for how to choose an assessment. These issues are also described in this *Guide*. Resources for locating assessment instruments are listed in the Appendix of this *Guide*.

The *Standards for Educational and Psychological Testing* (AERA, et al, 1999) provide extensive guidelines for the effective and responsible use of assessment tool and processes, including discussion of best practices and detailed information on technical adequacy of tests and assessments. Test users should review information on the stated purpose and development of an assessment tool to determine whether or not it matches users' purpose(s). Examination of evidence for the reliability and validity of the use and interpretation of assessment results should be a paramount concern for all those who use assessments to ensure that the instrument works effectively in the ways intended.

The final section of the *Diagnostic Assessment Guide* discusses the implementation, use, and interpretation of diagnostic and formative assessment results. A number of difficulties are briefly discussed – including problems and pitfalls that are common in current assessment practice or that may occur in the implementation of a new assessment system. Some of the challenges discussed at length in the research are aspects of teacher practice that do not conform to best practice in formative assessment processes. Research shows that teachers often apply summative assessment strategies borrowed from high-stakes tests to classroom assessment tools and predominantly focus on assessment for grading and evaluation purposes rather than using assessment processes to support student learning. The assessment tools used may not be designed to support diagnostic or formative applications. For effective diagnostic and formative assessment processes, it is important to select or develop a tool that provides an appropriate sampling of the content domain, is closely aligned with the instructional program, and that provides sufficient specificity to provide detailed descriptive feedback to support ongoing student learning.

The research on formative assessment also provides a number of suggestions for effective formative assessment processes. One recommendation is to ensure that there are clear linkages among assessment, curriculum, and instruction. Teachers should explicitly design feedback strategies that connect assessment results with instructional decision-making and planning for intervention.

As mentioned earlier, student involvement is a key component of formative assessment processes. Student involvement should be included as part of assessment and instructional activities including the use of self and peer assessment. Increased involvement enhances student engagement and increases student motivation and achievement. The research also recommends more integrated involvement of teachers in the design and use of assessment tools and results,

which requires increased professional development opportunities since many teachers may not know how to develop or select appropriate formative assessment tools, use assessment results formatively, or interpret assessment results to design responsive instruction.

Last, the research suggests changes in school or district level practices to support effective implementation of diagnostic and formative assessment processes. Policy should be adopted that communicates clear achievement expectations for students. Assessment systems should be coordinated across the district, and assessment results should be communicated in a timely and understandable way. To ensure assessment accuracy, investment must be made in fostering assessment literacy among participants, and in evaluating implementation of the assessment system.

## INTRODUCTION AND PURPOSE OF THE *DIAGNOSTIC ASSESSMENT GUIDE*

The purpose of the Diagnostic Assessment Guide is to provide educators with information that will guide their selection and use of diagnostic and formative assessment tools. Throughout this *Guide*, the term *assessment* takes on a broad array of meanings. The term might refer to a particular *assessment tool*, such as the *Early Diagnostic Mathematics Assessment* (EDMA). The term is also used to describe *assessment results* (scores, reports, and descriptive information) derived from students' responses to an assessment tool. The term *assessment* may be used to refer to an *event* such as screening at the beginning of a school year. Finally, the term may be used to refer to an *assessment process* – using assessment tools to gather assessment information as well as summarizing, interpreting, and acting upon information obtained from one or more assessment tools. Given the variety of meanings throughout this guide, we indicate whether or not we are discussing an assessment event, an assessment process, an assessment tool, or assessment results.

In addition to the array of meanings of *assessment*, there are many assessment purposes. This *Guide* defines each of them so that the purposes of *diagnostic* and *formative* assessments can be distinguished from the purposes of large-scale tests, interim assessments, etc. Educators must be clear about their needs so they can select one or more assessment tools that provide the information they need. When educators are clear about their assessment purposes, they are more likely to use the assessment results in a process that helps them achieve their goals. Finally, if educators are clear about their purposes, they are more likely to time assessment events so that results are available when needed.

This *Guide* also provides educators with information that will support the selection, use, and interpretation of results from diagnostic assessment tools. Recent legislation in Washington provides support for educators to purchase and use diagnostic assessment tools. This is an astute investment in that years of educational research link strong gains in student achievement, engagement, and motivation to the regular use and implementation of formative assessment tools and processes. Diagnostic assessment tools are a special type of formative assessment tool and process.

*“Formative assessment is central to good instruction in several ways, including focusing learning activities on key goals; providing students feedback so they can rework their ideas and deepen their understanding; helping students develop metacognitive skills to critique their own learning products and processes; and providing teachers with systematic information about student learning to guide future instruction and improve achievement.” (Lewis, 2006)*

This *Guide* provides background and general information on formative and diagnostic assessment tools and processes. The *Guide* briefly reviews the Washington legislation, defines a wide range of assessment purposes, describes some of the major findings from the research on formative and

diagnostic assessment, and discusses issues in the selection and use of diagnostic and formative assessment tools, as well as the interpretation of assessment results.

The Appendix presents resources for users of formative and diagnostic assessment tools and processes. It is not the purpose of this guide to describe or suggest instructional interventions, though it is well recognized that a strong linkage between assessment and instruction is a key component of educational effectiveness. This report also does not describe or support the use of specific assessment instruments. There is a companion report for this *Guide* that comes in two parts: *The Formative Assessment Comparative Guide – Consumer Report* and the *Formative Assessment Comparative Guide – Technical Report*. These *Comparative Guides* provide information on most commercially available assessment tools in mathematics, reading, science, and writing for grades K-12. The *Consumer Report* provides quick summary of the purpose of the assessment, a summary of the focus of the assessment, contact information for the publisher, costs, and a technical rating. The *Technical Report* provides detailed information regarding content assessed, information on evidence for reliability and validity of the tests, and additional details on scores, reporting, and administration procedures. These *Comparative Guides* are intended to help teachers, schools, and districts select the most appropriate tools for their assessment purposes.

## **OVERVIEW OF THE WASHINGTON STATE LEGISLATION**

The 2007 Washington State Legislature appropriated \$4.8 million to school districts to purchase diagnostic assessment tools and implement diagnostic assessment processes during the 2007-09 biennium. School districts were eligible to receive \$5 per student for the purchase and implementation of diagnostic tools. Districts that enrolled fewer than 100 students were to be allocated \$500 per school district. The number of students for each school district was determined using the October 2006 student count (See school district “October 2006 Student Counts” at: <http://reportcard.ospi.k12.wa.us/summary.aspx?year=2006-07>).

Applications were approved if the diagnostic assessment tools that were to be funded were consistent with the State’s definition of a diagnostic assessment and if funds were applied for an allowable use. Allowable uses included:

- a. purchase of assessments;
- b. costs of administering, scoring and reporting results; and/or
- c. training costs.

Funds were to be used for purchasing and administering the assessments to students. Funds could not be used for developing diagnostic assessments, although they could be used to administer and score previously developed diagnostic assessment tools.

During the 2007-2008 session, the Legislature changed the way the original \$4.8 million were to

be used. Approximately \$2.3 million were to be allocated to districts for purchasing and administering diagnostic assessment tools. The remaining \$2.5 million were to be used to develop and implement diagnostic assessment tools.

During the 2007-2008 school year, approximately \$1.8 million of the \$2.3 million were distributed to 116 of the State's 295 school districts, based on their iGrants applications for and proposed uses of the diagnostic assessment tools and results. School districts that received funding were required to report whether or not they expended the funds; remaining funds had to be spent on diagnostic assessment purchase and use in future years.

ESSB 6023 defined a *diagnostic assessment* as an assessment that “helps to improve student learning, identifies academic weaknesses, enhances student planning and guidance, and develops targeted instructional strategies to assist students” before the high school WASL. According to the legislation, to the *greatest extent possible* the assessment tools had to be:

- a) aligned to the State's grade level expectations;
- b) individualized to each student's performance level;
- c) administered efficiently to provide results either immediately or within two weeks;
- d) capable of measuring individual student growth over time and allowing student progress to be compared to other students across the country;
- e) readily available to parents; and
- f) cost-effective.

The legislation also authorized the preparation of this *Diagnostic Assessment Guide* and the development of a *Diagnostic Assessment Comparative Guide* to identify and provide information on commercially available diagnostic assessment instruments. This work was carried out by Measurement, Evaluation, and Statistical Analysis (MESA) Associates. Questions about the *Comparative Guide* should be addressed to Dr. Joseph Stevens, [jstevens.mesa@comcast.net](mailto:jstevens.mesa@comcast.net).

## **ASSESSMENT PURPOSES**

There are many different types of assessment tools. Various authors and users apply different terms to define the same or similar approaches to educational assessment. In this *Guide* we attempt to clarify assessment terms, using common-sense definitions that are consistent with the history of assessment practice and that draw important distinctions for application and practice. It is not the purpose of this *Guide* to discuss in detail all types of assessment tools. However, to provide clarity and contrast we briefly define and discuss a range of assessment purposes that may be distinct in important respects but often overlap with formative and diagnostic assessment purposes.

## **NORM REFERENCED- AND CRITERION REFERENCED/STANDARDS-BASED ASSESSMENT TOOLS AND PURPOSES**

Glaser (1963) distinguished between two types of information that can be provided from performance on an achievement test: 1) the relative position of one test-taker to others, or 2) the degree to which a test-taker has attained a particular criterion or level of achievement. This distinction has traditionally defined the essence of norm-referenced and criterion-referenced assessment purposes, respectively. More recently, with the advent of the standards movement, standards-based assessment is a name for a specific form of criterion-referenced testing.

### ***Norm-Referenced Testing***

The purpose of norm-referenced testing (NRT) is to compare one examinee's performance to the performance of a representative group of examinees of the same age or grade level, and who were administered the same assessment under the same standardized testing conditions. Judgments of performance are relative – they only describe a person's standing in comparison to the norm group rather than what students have learned. As an analogy, consider people running a foot race. If the results are reported in terms of order of finish (1<sup>st</sup>, 2<sup>nd</sup>, 3<sup>rd</sup>, etc.), then a norm-referenced interpretation has been made.

NRTs are designed and constructed to rank the test takers. Therefore there is a preference in test construction to select items that discriminate well among the test-takers and that represent a range of difficulty from below grade level to above grade level content. In this way, results are useful in comparing students' scores. Because of the way NRTs are constructed, scores tend to result in a normal or bell-shaped distribution of test scores. Scores are commonly reported as percentile ranks that report the relative ranking of an individual in comparison to the rest of the scores in the distribution. For example, a student score that results in a percentile rank of 75 means that 75% of test takers in the norm group received the same or lower score than the student.

The quality of the scores from an NRT depends on how well the norm group represents the population of examinees (e.g., how well the norm group represents all fourth grade students in the United States). While some norm-referenced scores are based on local comparisons (i.e., local norms or 'user' norms), generally scores are based on studies done by commercial test developers using nationally representative samples as the norm groups. To represent the population, professional test developers use careful sampling designs that ensure the norm group matches census information with respect to age or grade, gender, ethnicity, type of community, school size, and region of the country. The norm group is administered the NRT using the same standardized conditions that will be used for all test-takers. Because the process of sampling and testing of a norm group is complex, time-consuming, and resource-intensive, test publishers do not test norm groups every year. The results from a norm group may be used as the basis for comparison for five or more years. Students' scores from a local administration are then reported in relation to the performance of this norm group.

Many commercially available tests are NRTs, including the *California Achievement Tests* (CAT); the *Comprehensive Test of Basic Skills* (CTBS)-*TerraNova*; the *Iowa Tests of Basic Skills* (ITBS)

and *Tests of Academic Proficiency* (TAP); *Metropolitan Achievement Tests* (MAT); and the *Stanford Achievement Tests* (SAT), among others. Most NRTs are “battery” type tests that must cover an array of national content standards; therefore, there are usually only a small number of items within any specific area of the content domain. As a result, NRTs do not provide reliable information at levels more specific than general content categories (see for example, Stevens, 1995).

### ***Criterion-Referenced Testing***

The purpose of a Criterion-Referenced Test (CRT) is to determine whether or not students achieved a standard of mastery or competence in relation to the knowledge and skills students should learn at a particular grade level. There is no need to compare one student’s performance to the performance of other students; therefore, there is no need for below and above grade level content. Depending on the type of information needed, the passing score for a CRT may be set to indicate *minimum* competency or to indicate *mastery* of complex content. It is possible for nearly every examinee to earn a passing or a failing score on a CRT.

CRTs may be developed nationally (i.e., the National Assessment of Educational Progress or NAEP) or by states, school districts, schools, and/or classroom teachers. The test development processes for CRTs differ from test development processes for NRTs. CRT items are chosen to represent the content standards being taught. After a period of instruction on certain skills, the expectation is that the majority of students will perform well on items measuring those skills. A properly designed CRT contains multiple items for each learning target in the content domain allowing some evaluation of students’ strengths and weaknesses.

For a CRT used at district, state, or national levels, the passing score and all performance level cut-scores are most commonly determined by a committee of experts. In classroom applications, the passing score may be determined by the teacher. In either case, interpretations of performance on the CRT depend on subjective judgments about the proper location of the passing score and other cut-scores (Cizek & Bunch, 2007). The degree to which the subjective judgment is a reasonable judgment depends on the process used to set the cut-scores and the qualifications/expertise of the individuals who set the cut-scores.

### ***Standards-Based Testing***

Standards-Based Tests (SBTs) are one type of CRT. The central feature of an SBT is the alignment of test content to a particular set of content standards; reporting of assessment results describes performance in reference to proficiency levels. SBT is the name for a CRT that meets the accountability requirements of the 2001 Elementary and Secondary Education Act (also known as “No Child Left Behind”). There is substantial variation from one state to another in the fundamental construction of their SBT. NCLB requires the reporting of results in proficiency levels (i.e., “basic”, “proficient” or “advanced”). The “proficient” level is intended to represent

what students should know and be able to do in different content areas at a particular grade level.

Defining the proficiency categories requires a judgmental process for determining how test performance relates to expectations for student performance (see Cizek & Bunch, 2007). Because each state develops its own content standards and standards-based tests, individuals within each state often debate the appropriateness of academic content standards and associated performance levels or benchmarks. Debates focus on whether or not the content standards are too general or too narrow, too easy or too difficult, and whether or not appropriate levels of cognitive complexity are represented in the standards.

A key issue in the use of standards-based tests is the degree of alignment between the test content and state content standards. One of the challenges in constructing SBTs is how to fully represent content standards with a test of limited length. Often, many important standards or benchmarks are not assessed or the curricular alignment is only present at a general level, making it difficult to provide detailed diagnostic or formative assessment results.

States have constructed their SBTs in a variety of ways. Some states have constructed their SBTs directly from state content frameworks; others have used existing NRTs and simply set proficiency cut-offs on the NRT scores. For states that adopt NRTs, there is only a loose connection between the state's content standards and the content on its state tests. Finally, some states use an augmented NRT, wherein a core of items come from an existing NRT and supplemental items are added to create a stronger match to the state's particular content standards. It is important to recognize that simply attaching proficiency category descriptions to test scores does not eliminate important differences in test development and construction that can affect proper use and interpretation of results. Given that NRTs assess above and below grade level content, scores are very difficult to interpret in terms of grade level content standards.

## **STANDARDIZATION IN ASSESSMENT**

Standardization refers to the process of making the test content and structure, testing conditions, and test administration comparable or "standard" for all test takers. This process of controlling test content, structure, conditions, and administration is necessary if one person's performance is to be compared to another's. It is obviously an important and necessary feature of norm-referenced tests and of tests used for summative assessment purposes.

Standardization may also be important when using other types of tests for other purposes. Whenever direct comparisons are to be made from one test taker, school, district, or state to another or from one time to another, standardization is important. Some degree of standardization is important when administering standards-based assessment tools, to ensure that judgments of whether or not test takers have met proficiency is determined using the same conditions from one test taker to another.

Standardization may also be important for diagnostic and formative assessment tools and events depending on the purpose of assessment and how the assessment results will be used and interpreted. If information from a diagnostic or formative assessment is used to make comparisons across test takers, standardization is important.

For some assessment purposes, standardization is directly at odds with the need to provide accommodations that meet the needs of a particular student. In such cases, standardization of administration across individuals makes little sense. However, other aspects of standardization may be just as important. It may be necessary to use test forms that are equivalent from one assessment occasion to another, if the purpose of the assessment is to measure growth of skills or abilities over time. Standardization of test content, administration of test forms, and testing conditions ensure that observed growth is due to the skills and abilities of examinees and not due to fluctuations in test content, score meaning, or administration conditions.

### **FORMATIVE AND SUMMATIVE ASSESSMENT PURPOSES**

Scriven (1967, pp. 40–43) is credited with the first published use of the terms *formative* and *summative* as descriptions of two general functions of program evaluation. Later these terms were applied more narrowly to educational assessment. The distinctions between formative and summative assessment that we draw here are based primarily on assessment purposes, the timing of assessment events, the types of tasks given to students, the results produced by the assessment tools, and the ways in which assessment results are used and interpreted. We define screening, diagnosis, interim measurement, and progress monitoring as specific subcategories of formative assessment that have unique purposes. We also address whether or not different assessment tools can be used in tandem and whether or not one assessment tool can serve multiple purposes.

### ***SUMMATIVE ASSESSMENT***

The key purpose of a summative assessment tool is to summarize performance at a particular point in time. Summative assessment tools are primary tools in accountability testing and in efforts to evaluate the performance of students, schools and states. Summative assessment tools are commonly used to mark attainment of a benchmark and/or certify student performance. The delivery of a summative assessment is usually timed at or near the end of a school year, a course of study, a school term, or an instructional unit rather than *during* the course of instruction. Summative assessment events occur less frequently than formative assessment events and are designed to provide a snap-shot of performance at a particular point in time. Summative assessment purposes are inherently evaluative and the results are typically expressed as grades, judgments of proficiency, or measures of attainment. Summative assessment events are generally high stakes events, often being used to determine eligibility for the next grade, graduation, or other significant decisions.

Although not a requirement, many summative assessment tools are designed to yield results that

compare an individual's performance to other individuals and are therefore norm-referenced (see discussion above on Norm-Referenced Testing). When used in accountability applications like NCLB, summative assessment results emphasize group performance (e.g., "40% met proficiency") and may or may not include reporting of group comparison information (e.g., "percentile rank"). However, the main purpose of summative assessment events is the reporting of results that emphasize evaluative judgments (e.g., "grade of A", "course is passed", "meets proficiency"). Because of the inherent emphasis on evaluation in summative assessment, Harlen & Crick, 2003, found that the primary motivation for students taking such assessments is often extrinsic (e.g., to please others, to earn a diploma) rather than intrinsic (to self-evaluate attainment of a personal goals).

Summative assessment tools are often equated with standardized tests such as state accountability tests administered for NCLB reporting purposes; however, they are more commonly used for district and classroom assessment events. Local summative assessment tools include district benchmark tests, classroom end-of-unit or chapter tests, and final or end-of-term exams. Because summative assessment events occur after teaching, it is difficult to use summative assessment results to guide instructional interventions, to provide feedback to students, or to modify the course of learning. Instead the strength of summative assessment results is as a means to gauge the absolute level of student performance, to help evaluate the effectiveness of programs, teaching, school improvement plans, or the adopted curriculum.

Advantages of well-constructed summative assessment tools are the provision of *reliable* and *valid* snapshots of student knowledge and skills in a defined content area at the time of testing. Summative assessment tools can be a cost effective means for determining whether or not large groups of students have met learning targets on a broadly sampled representation of a content area.

Of necessity, summative assessment tools must measure a broad range of knowledge and skills in a relatively brief period of time. For this reason, developers of summative tests select test questions that are a sample of all that students should know and be able to do. Test development tends to emphasize the sampling of a breadth of content to represent the course of study being evaluated. Test development methods focus on measurement of overall level of skill and ability in the content area. In many summative assessments the ability to discriminate one performance level from another is the primary psychometric concern.

### ***FORMATIVE ASSESSMENT***

The Council of Chief State School Officers (CCSSO) has created an interstate consortium called the Formative Assessment for Students and Teachers: State Collaborative in Assessment and Student Standards (FAST SCASS). FAST SCASS defined formative assessment as:

“... a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students’ achievement of intended instructional outcomes.” (McManus, 2008, p. 3)

The two key characteristics of a formative assessment process are: 1) a purpose to enhance learning, inform instruction, or provide feedback, and 2) timing that involves the delivery of assessment at the beginning or during instruction or a course of study, class, or instructional unit (Black & Wiliam, 1998b; McManus, 2008; Sadler, 1989). The purpose of a formative assessment process is to guide and motivate learning, and to provide feedback to the student and teacher. Unlike summative assessment tools that provide a final evaluation of goal attainment, formative assessment tools are designed to provide an ongoing assessment of the progress of learners toward learning targets. To facilitate student learning progress, design and development of a formative assessment tool requires greater depth and representation of content, and a design that allows users to discover and reveal student strengths and weaknesses. Instead of a psychometric emphasis on discriminating among student performances, a formative assessment tool should be technically adequate in measuring achievement of clearly specified learning targets and in tracking learning over time. Scores and reports from formative assessment tools are intended to allow users to compare current performance to learning targets or goals. Reports are especially effective when assessment results provide information that can be used prescriptively to guide the design and delivery of subsequent instruction.

The timing of assessment events is a key difference between formative and summative events. While summative assessment events occur *at the end* of an instructional period, formative assessment events occur *during* the instructional process. Formative assessment tools are designed to be more closely linked to learning and instruction and, therefore, they are used more frequently, dynamically, and are interlaced with instructional activity. However, to be effective, formative assessment events must be given at a time that allows for instructional changes by a teacher, to promote changes in study activities by a student, or to facilitate changes in student motivation following assessment feedback. The intent of a formative assessment process is to provide the information needed to modify and guide teaching to improve its effectiveness and student achievement. This means that the information provided by formative assessment tools must occur during the time when learning is occurring so that both teacher and student can understand what adjustments need to be made so the student can progress toward learning goals.

Another key difference between summative and formative assessment purposes is the relative emphasis on evaluation or grading. While evaluation is at the core of summative assessment, there may be no evaluation or grading, per se, in the use of formative assessment tools. Rather, results from formative assessment tools are used to provide feedback and guidance; the assessment itself may be seen more as a form of practice than as a test. As a student learns, it is not expected that high levels of achievement or mastery will be immediately evident. Instead, a period of learning and engagement must occur during which the emphasis is on the assessment of

progress and determining the next steps to be taken along a pathway culminating in the learning goal. The purpose of a formative assessment process is to inform the student and the teacher about the progress being made as well as guiding the next steps to be taken to support learning.

Another difference between summative and formative assessments is the role of the student. While a formative assessment process requires and depends upon the involvement of the student, there is little involvement of the student in a summative assessment process beyond test-taking. In a formative assessment process, students should be involved in assessing their own learning and in using the feedback provided by each assessment tool to modify their own behaviors. The feedback loop among assessment, instruction, and learning (see next paragraph) is a critical component of an effective formative assessment process. Research shows that student involvement in assessment increases their motivation to learn (Natriello, 1987). Teachers may involve students in the assessment process by providing descriptive feedback, having students chart or monitor their own progress and performance, or by having students help assess and give feedback to peers. Direct student involvement also provides clear information about what the student knows and can do, what still needs to be learned, and how to reach next steps on the pathway toward learning goals.

It is also important to note that formative assessment processes may be particularly effective for lower performing students. Research shows that the use of formative assessment processes may narrow the gap between low and high performing students while raising the overall level of achievement for all students (Black & Wiliam, 1998b). The specific feedback provided by formative assessment tools is important, both for student understanding of how to learn, and for helping teachers make specific plans about the next steps for student progress and success.

Formative assessment tools may include observational checklists, homework, student self-evaluation guides, quizzes, and ongoing projects. To be effective, formative assessment tools must assess a few selected learning targets and provide results that guide instruction toward achievement of those targets. In the following sections, we describe several subcategories of formative assessment purposes that are relevant to diagnosis and intervention.

On the following pages, several types of formative assessments are described, including screening assessments, diagnostic assessments, interim assessments, and progress-monitoring. In the side bars, a health example is used to help readers better understand the distinctions between these assessment purposes.

## Screening Assessment

The purpose of a screening assessment tool is to make an early identification of students' strengths or weaknesses, to allow classification, placement, or intervention. Screening assessment tools are a subtype of formative assessment tools, but they are not designed to result in an in-depth understanding of student skills and abilities. Instead, screening assessment tools are designed to rapidly identify those individuals who need specific forms of placement, attention, or instructional intervention.

As a result, an assessment tool used for screening may be characterized by less depth of content and by less accuracy or detail in the assessment information provided. One would also expect a well designed screening assessment tool to focus on a narrow range of skills, knowledge, or performance at a particular grade level rather than attempting to measure a large range of ability. For example, to identify children in need of reading intervention for basic skills, a good screener would concentrate on the identification of basic skill deficits; an instrument designed to screen children for a talented/gifted program would focus on other ranges of performance and ability.

A key distinction between screening assessment and other formative assessment events is the timing of administration. Unlike other forms of assessment, screening assessment occurs before instruction or placement. The results from a screening assessment may also suggest the need for additional assessment events or samples of student work to help determine what areas of the student's knowledge and skills are truly problematic, most in need of remediation, and are amenable to instruction.

### Screening Assessment

Kale has taken gymnastics classes for a year. He is not making progress and is easily fatigued. The teacher thinks he needs to make more of an effort. During a routine school health screening, the doctor finds that Kale has low blood pressure and a slow pulse. Kale also appears to have some breathing problems. The doctor recommends that Kale have some further tests.

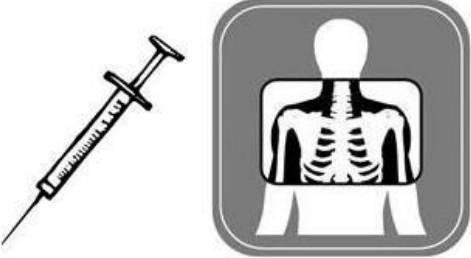
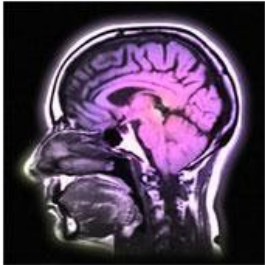


## Diagnostic Assessment

Diagnostic assessment tools are a subcategory of formative assessment tools that are designed specifically to identify the causes of students' learning problems – usually with the intent to guide or modify instruction or to design differentiated instruction. Many consider diagnostic assessment to be a distinct category of assessment (Kellough & Kellough, 1999; McMillan, 2001); however, our view is that much of the purpose, practice, and application of formative and diagnostic assessment overlap.

An effective diagnostic assessment process will focus on the identification of specific student weaknesses that will lead to remediation through additional instruction. A diagnostic assessment process can be viewed as a decision-making strategy for determining when and how to deliver instructional remediation. For a diagnostic assessment tool to be useful in this process, it must provide detailed analysis of student performance that allows specificity in diagnosis, and that provides sufficiently rich detail so that intervention can be planned and implemented. To be instructionally relevant, diagnostic assessment tools must also be sufficiently aligned and representative of content being taught, or soon to be taught, in the classroom.

While other forms of testing (e.g., norm-referenced or standards based) may identify students who are performing well or poorly, diagnostic assessment tools are designed to provide a bridge between identification of the proficiency level and instruction by illuminating the *reasons* for the level of performance. Diagnostic assessment tools may also help users determine whether or not the student is ready to move on to the next skill or concept. For example, instruction on interpreting a character's motives may be ineffective if students struggle with literal comprehension or with following the events of a story.

Diagnostic Assessment
<p>The doctor does several diagnostic tests to figure out why Kale's breathing is labored and his blood pressure and pulse are low. She checks Kale's lungs, thyroid, and blood. She finds that Kale has an enlarged thyroid, a low thyroid hormone count, and anemia (a low hemoglobin count).</p>



The results of a high quality diagnostic assessment tool help ensure that instructional activities are tailored to a student's identified needs. Diagnostic assessment purposes may be contrasted with formative assessment purposes by a greater focus on those in need of remediation and by the presumption that *individualized intervention* will be linked to diagnosis. Therefore, diagnostic assessment tools typically focus on the assessment of basic, underlying skills rather than higher order thinking skills. However, once students have mastered basic skills, tools are needed to determine why some students struggle with higher order thinking skills.

### *Interim Assessment*

The purpose of an interim assessment tool is to provide a measure of students' progress toward achieving proficient performance on a standards-based summative test or to measure their growth on a measurement scale as they move toward a final summative assessment event. To be effective, interim assessment tools measure the same knowledge and skills as are measured on the summative test and indicate students' level of performance on an interval scale. Interim assessment events occur several times each year. Interim assessment tools provide sub-scores related to areas of tested knowledge and skills (e.g., number sense, measurement, literal comprehension) so that teachers know how to focus their teaching. If students are not demonstrating adequate growth, teachers can reteach important skills and concepts in areas of weakness. However, interim assessments are unlikely to provide sufficiently detailed results to diagnose learning problems.

Interim assessments may be *computer adaptive tests*. Computer adaptive tests use students' responses to 'locate' the student on an underlying scale (similar to a ruler), so that the students only see and respond to test items targeted to their current level of performance. The development of computer adaptive interim assessments requires a large pool of test items that are all calibrated to the same achievement scale. The computer program must select items for each student that represent the same content standards as those that are assessed on the summative, standards-based test. The computer program selects items for each content standard that are at the appropriate level of difficulty for each student's ability level.

### **Individualized Intervention**

Kale's doctor prescribes a thyroid hormone and iron tablets.




### **Interim Assessment**

Kale goes back to the doctor every three months so that the doctor can check his blood pressure, pulse, thyroid hormone level and hemoglobin level.



## ***Progress Monitoring***

Progress monitoring is a special type of interim assessment process that is characterized by frequent, repeated assessment. Screening, diagnosis, intervention, and progress monitoring are used in combination in a process called “response to intervention.” Progress monitoring is generally used in special education programs to determine whether or not students should receive special education services. A progress monitoring assessment process can be implemented with individual students or groups of students. Progress monitoring is generally used in combination with specific instructional interventions so that the student’s response to interventions can be observed and evaluated, to determine whether or not the interventions are successfully addressing students’ learning needs. Progress monitoring provides a means to determine whether or not a student is showing adequate progress or needs additional forms or methods of instruction.

<b>Progress Monitoring</b>
<p>Kale’s progress is also monitored for his performance in gymnastics. The teacher’s weekly assessment of Kale shows he has more energy during class. His performance is also improving every week.</p>


In typical practice, a progress monitoring process is used to determine a student’s current level and, over time, to determine rate of improvement and establish learning goals. Frequent assessment is conducted to monitor progress toward the learning goals. If there is not adequate progress and learning goals are not met, additional or alternative forms of instruction are implemented. The progress monitoring assessment results are also useful for evaluating the relative efficacy of multiple approaches to instruction or intervention.

Since progress monitoring depends on frequent assessment events (perhaps weekly or monthly) and the tracking of student performance over time, there are technical requirements for a progress monitoring tool to be effective. These requirements may be different than those for other types of assessment instruments. First, a key requirement of a good progress monitoring tool is the availability of multiple forms of the tool. To allow the intensive repeated assessment necessary for some applications of progress monitoring, an assessment tool may need to have 20 or more test forms. A progress monitoring process depends on the ability to make valid comparisons of student performance over time. As a result, tests and their administration conditions need to be standardized. It is also important that test forms are designed so that the content and difficulty of each form are equated and scaled, to allow valid comparisons from one form to another.

There are two common but distinct approaches used to monitor student learning when students

are served by special education programs: curriculum based measurement (CBM) and mastery measurement. Most classroom assessment tools used in special education programs assess students' mastery of a single skill or a small set of skills. When mastery is demonstrated, instruction and assessment focuses on the next set of skills. As a result, each assessment tool references different concepts and skills at different times of the school year. Student progress is difficult to track over time because different content is being assessed on each testing occasion.

In contrast, CBMs can be effectively used to monitor progress. The CBM approach to measurement depends on the construction of tests that sample all skills/knowledge in one curriculum area (e.g., reading) on each assessment occasion. Each test form is designed to be an alternate form with different items but the same representation of the annual content and equivalent difficulty of each form. Thus, scores received by a student on one occasion can be compared to scores received at other times of year so that progress can be evaluated validly.

### ***SUMMARY OF ASSESSMENT PURPOSES***

The foregoing discussion may suggest that all assessment tools fall neatly into one of the assessment purposes described above; however, there can be a great deal of overlap among the different assessment purposes, and some tools, if developed appropriately, may be used for more than one assessment purpose. For example, a well developed progress monitoring tool might provide diagnostic information.

The foregoing also suggests that all types of formative assessment tools fall into one of four categories: screening, diagnosis, interim evaluation, or progress monitoring. Classroom teachers use many different types of formative assessments to evaluate both student learning and the success of their instruction. These tools may be developed by the teacher or embedded in published instructional materials. It is beyond the scope of this *Guide* to describe the full range of formative assessment tools, processes, and events. Three recommended classroom assessment texts (Shepard, 2006) are *Student Centered Classroom Assessment* (Stiggins, 20xx), *Understanding by Design* (McTighe & Wiggins, 20xx), and *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms* (Taylor & Nolen, 2007). These texts are intended to guide classroom teachers in the selection, development, and use of classroom assessment tools, results, and processes. Information about other types of classroom-based assessment is given on page 26 in this *Guide*.

One of the challenges faced by educators and policy-makers is the inconsistency with which the terms describing these assessment tools and purposes are used by test publishers, test users, and researchers. One clear distinction is between summative and formative assessment. Summative assessment events tend to occur after instruction has occurred while formative assessment occurs before or during the instructional period. The emphasis in summative assessment is on evaluation while the emphasis in formative assessment is on enhancement of learning.

Within formative assessment there is a great deal of overlap among different subtypes. Screening assessments are brief, occur prior to instruction, and serve to aid placement or classification decisions. Diagnostic assessment can be characterized by a greater emphasis on discovering weaknesses and reacting with remedial instruction. Interim assessments are those that give educators a sense of whether or not students are progressing toward proficiency on a standards-based test. Progress monitoring is characterized by more frequent, repeated assessment to track the course of learning and evaluate the effectiveness of instructional interventions.

Effective use of assessment results depends on selecting the tool(s) that are likely to provide the information needed. An interim assessment tool that provides sub-scores related to broad state standards is unlikely to provide adequate information to determine the causes of students' learning difficulties; therefore, such a tool will not provide adequate diagnostic assessment information. If an assessment tool designed to be a screening tool is used for diagnostic assessment, it is unlikely to provide sufficiently specific information about students' strengths or weaknesses or to assist teachers in designing adequate instructional interventions. Diagnostic assessment tools may not provide sufficient breadth of coverage of the content standards to determine whether or not students are making adequate progress toward standards. In selecting assessment tools, users should carefully examine the content assessed and the types of reports generated, to see whether or not the information provided will meet users' needs.

### ***COMBINING DIFFERENT ASSESSMENT PURPOSES***

Knowledge of the distinctions in purposes of assessment is important for correctly matching an assessment tool to the intended purpose and use of the assessment results. Assessments of one type seldom can be substituted for an assessment of another type (Popham, 1999). Because of the different purposes of formative and summative assessments, the design, construction, and development of the instrument will often differ. The timing of assessment, administration conditions, scoring, and reporting are also likely to be different depending on whether or not an assessment is designed to be formative or summative. Of particular importance in a diagnostic assessment tool is design that provides the level of detail needed for identification and diagnosis of specific causes of weaknesses. Of particular importance in a formative assessment is a design that provides results that can be directly linked to instruction. To accomplish these tasks, diagnostic and formative assessment tools involve different item types, scores, and score reporting methods than summative assessment tools. Because of these fundamental differences in test purpose, design, and reporting, a test designed for one purpose may not function well for another. We caution users to carefully evaluate and determine whether or not an assessment tool considered for use has been designed and developed to effectively meet his or her needs and goals.

Some believe that any assessment can be used in either a formative or summative way. However, for an assessment to work well, it must be designed and constructed to fit its intended purpose.

For example, the kind of standards-based tests (SBTs) used in state NCLB testing may be used to provide formative feedback but with limited success, since they do not have enough items or the appropriate kind of items and tasks to provide diagnostic detail, and because the timing and infrequency of assessment events will not be suited to instructional monitoring and intervention. We urge caution in attempting to use an assessment tool for applications other than the primary purpose for which the assessment tool was developed unless there is independent research validating the additional uses.

Even when assessments are correctly categorized as serving different purposes, there is some debate as to whether or not different assessment types can be used together in the same assessment or accountability system. Crooks (1988) examined whether or not formative and summative assessment use can be compatible. His view was that the functions served by the two types of assessment were distinct (feedback versus grading, for example) and that the summative function has been too dominant. Crooks argued for separating formative and summative functions. In contrast, Brookhart (2001) and others argue that each kind of assessment can be seen as parts of the same whole. Biggs (1998) suggested that we need to make use of both kinds of assessments but this marriage works best if formative and summative assessments are both criterion referenced.

Some argue (e.g., Biggs, 1996) that there is a powerful interaction between formative and summative assessment purposes that could be profitably considered together within a common framework. Such a synthesis could provide support for learning that contextualizes the results of summative assessment events to ensure their more positive application, and allows the results to support feedback from formative assessment. However, when feedback from a summative assessment tool cannot be used to lead to appropriate adjustments to teaching and learning, a key component of formative assessment (Sadler, 1989), the two assessment types are seen in effect as mutually exclusive.

Whether or not they are mutually exclusive depends on the model of assessment adopted. Feedback from summative assessment events (“backwash”) is generally agreed to be negative, focusing on individual characteristics of the learner instead of the learning process and task, and leads to a shallower approach to learning. Feedback from formative assessment, on the other hand, is oriented directly toward the learning task and facilitates deeper learning (Biggs, 1998).

## **FORMATIVE ASSESSMENT PROCESSES: BACKGROUND AND FINDINGS FROM THE RESEARCH**

This section of the *Guide* presents research on formative assessment tools and processes. This research represents studies conducted over many years to illuminate the characteristics of effective formative assessment processes, and to better understand how these processes work to support teaching and student learning.

Early research helped define and characterize formative assessment, and what practices and processes are involved in making assessment formative. Bloom et al. (1971) borrowed the term “formative evaluation” from Scriven’s (1967) description of different kinds of program evaluation. Bloom and colleagues were concerned with the use of brief tests for the evaluation of mastery learning. Their model consisted of a) the diagnosis of learner characteristics, b) the analysis of learning tasks to determine the next instructional steps, c) feedback and corrections, and d) summative evaluation of attainment. Sadler (1983, 1989) described the importance of a feedback loop in the use of formative assessment. In this model, formative assessment entailed a) attending to learning goals, b) developing strategies to meet goals, and c) monitoring performance to determine goal achievement. Both of these early models of formative assessment emphasize the use of feedback and explicit attention to the discrepancy between student performance on a current assessment tool and the attainment of learning goals.

The research on formative assessment establishes its positive impact on a number of features and outcomes of educational practice. Natriello (1987) found that student motivation and achievement were impacted by several features of formative assessment practice including a) a focus on tasks rather than comparison of student performance, b) use of clear criteria for achievement, c) setting challenging standards, and d) provision of differentiated feedback to students. Crooks (1988) documented a number of positive effects of formative assessment on students. He found that formative assessment served to consolidate students’ prior skills and knowledge before new material was introduced, helped to focus students’ attention, encouraged active learning, and provided greater opportunities for practice. Some other important features of formative assessment noted by Crooks were the provision of corrective feedback, development of students’ self-monitoring, guidance of further instruction, and the creation of feelings of mastery and accomplishment for students.

One of the most important results from the research on formative assessment is the finding that regular use of formative and diagnostic processes results in substantial gains in student achievement on large scale tests. In an extensive review of the research, Black and Wiliam (1998b), found that the use of formative assessment resulted in improvements in learning achievement ranging from .40 to .70 of a standard deviation. They found that the use of a formative assessment process raised student achievement overall, closed the achievement gap between lower achieving and higher achieving students, and positively affected student

motivation and self-esteem. The research also documents that well-designed formative and diagnostic assessment tools can provide detailed, individualized, and instructionally relevant information that can guide and foster both teaching and student learning (Black, Harrison, Lee, Marshall, & Wiliam, 2004). Thus, in contrast to commonly used summative tests, formative assessment tools provide a direct and effective link between assessment and instruction.

## THE IMPORTANCE OF FEEDBACK

Feedback is an integral component of any assessment process. Whenever assessment events occur, feedback is provided. Close examination of how feedback is provided and used reveals a great deal about the purpose and utility of an assessment tool or system of tools. For example, summative feedback in the form of a course grade provides the student with information about achievement of course goals and communicates similar information to other consumers of the grade report (e.g., parents, teachers in the next course). Summative assessment feedback may also shape future learning by influencing student enrollment decisions, or by motivating a student to work harder during the next grading period. Most commonly, however, both the timing and the level of detail in the report of summative feedback prevent its effective use to guide instruction or alter specific trajectories of student learning.

On the other hand, feedback that is directly linked to instructional improvement is a distinguishing attribute of formative assessment. Formative feedback can provide immediate information to students, teachers, or administrators. The focus in formative feedback is on how assessment information can inform instructional improvement. Formative feedback has been defined as:

“...information about the gap between the actual level and the reference level of a system parameter *which is used* to alter the gap in some way.” (Ramaprasad, 1983, p. 4, emphasis added)

There are several noteworthy features of this definition. First, there is an implicit learning goal defined (i.e., reference level). Second, assessment results are used to reveal the discrepancy between current level of performance and the learning goal (i.e., the gap). But last and perhaps most important is the idea that the assessment results are used to alter the gap. Thus a key feature of a formative assessment process is use of information about gaps in desired performance to alter or change instructional practice. This might occur for a student by having different instructional activities assigned to improve mastery. For a teacher, formative feedback might result in a change in curriculum design for the whole class if the teacher found a gap in performance for many students in a class. The expectation of the feedback provided by formative assessment results is that it will help students improve their performance relative to learning goals. However, for formative assessment events to result in effective use of results, they must occur repeatedly during the learning process. When a formative assessment event occurs during learning, feedback can be provided while there is still time for the teacher to take action and for the student to benefit from

feedback

Effective descriptive feedback focuses on the learning process, identifies specific strengths and accomplishments, identifies weaknesses for improvement, and describes the pathways students can take to close the gap between current performance and learning targets. Effective feedback also provides scaffolding that helps students and teachers understand next steps to be taken to move forward in their learning. The most helpful feedback provides specific information about current levels of understanding, suggests means for improvement, and motivates students to focus their attention on learning goals, rather than on getting right answers on tests (Bangert-Drowns, Kulick, & Morgan, 1991). Further, to effectively use diagnostic and formative assessment results, feedback to teachers must provide some degree of prescription about what instructional interventions are needed. To be most effective, the information must relate to a developmental model of cognitive growth that helps to guide the course of learning in developmentally valid ways (i.e., construct-relevant; Messick, 1975). Clearly, given these critical purposes for feedback, assessment results are only part of the feedback. Information regarding effective instructional practices in response to learning challenges is essential.

In summary, it is clear from an abundance of research that one of the central characteristics of a formative assessment process is the provision of feedback. Feedback is the critical link between assessment and instruction that fosters the benefits of formative assessment. In planning the implementation of formative assessment systems, users should explicitly consider the match between curricular goals and the assessment instrument to ensure that feedback information will be matched to assessment purpose. Users should also explicitly design methods and procedures to enhance the use and impact of feedback information to motivate students and to guide instruction and curricular planning and design.

### **INFORMAL ASSESSMENT AND CLASSROOM ASSESSMENT**

The focus of the *Washington Diagnostic Assessment Project* is on commercially available formative and diagnostic assessment instruments (see review in the *Washington State Diagnostic Assessment Comparative Guide*). However, research shows that informal assessments and locally developed classroom assessments can be very effective for some types of formative assessment purposes. Such assessment strategies may include question and answering techniques used by a teacher with students, observations during small group work, homework, quizzes, projects, and other techniques. Effective teachers can use a range of assessment strategies and techniques to gather valuable formative information from students. This information can be applied to modify instruction and to guide the delivery of instruction and the course of student learning. In such usage, assessment is closely intertwined with instruction.

A number of instructional strategies suggested in the research can be used in support of classroom assessment. These include involving students in setting goals and having clear expectations for learning. When students participate in goal setting they develop a better understanding of what is

expected, as well as the criteria for meeting goals. Students can be included in the definition and description of what quality work looks like, what criteria should be used to judge goal attainment, and the processes to move toward learning goals. Assessment tools, assessment results, and examples of assessments that demonstrate goal attainment can all be used and discussed with students to support progress.

### *QUESTIONING*

Questioning is an integral part of pedagogy. The strategic use of questioning should be viewed not only as an instructional strategy but as a formative assessment activity. Well framed questions allow the teacher to quickly determine the level and nature of student understanding. Questioning can make almost immediate instructional adjustment and adaptation possible. The adroit use of questions can encourage metacognitive thinking in students and can help model learning strategies and problem solutions. Effective questioning can also engage students in the classroom and help motivate students. Another effective aspect of questioning strategies concerns helping students learn how to frame their own questions effectively, either for use with the teacher or in peer activities with other students (Johnson & Johnson, 1990; Rosenshine et al., 1996).

### *OBSERVATION*

Observation is another classroom assessment strategy that can provide formative assessment results. Direct observation of student work and activities is an important mechanism for gathering formative assessment information. The teacher may be able to observe processes or procedures being used by students that can reveal misconceptions, weaknesses in skills, and other information that can be used to make adjustments to improve teaching and student learning. Teachers can also encourage students to observe and assess how peers complete work or solve problems as a way to make the learning process more explicit and to develop a learning community.

### *PEER AND SELF-ASSESSMENT*

Peer and self-assessment processes have also been shown by research to be effective formative assessment strategies and to be motivating for students (Biggs, 1999; Black & Wiliam, 1998b; Brown, Rust & Gibbs, 1994; McManus, 2008). Peer assessment activities help to create a learning community within a classroom. Self assessment activities can increase student understanding of their progress and how learning targets can be achieved. When students are involved in goal setting, self assessment provides an important opportunity for students to monitor their own progress and develop metacognitive skills in support of learning.

## ***DESCRIPTIVE FEEDBACK***

Descriptive feedback is an integral part of effective formative assessment. Information gathered by the teacher in questioning, observation, and other classroom activities can be used to guide student learning through detailed feedback. It shows students how they are currently performing, how that level of performance relates to learning targets and goals, and how they can make progress toward their learning targets.

Deeper discussion of these valuable alternative methods of classroom instruction and assessment are beyond the scope of this *Guide* but the reader is encouraged to consider these methods as additional alternatives for supporting the use of formative assessment and fostering student learning (see, for example, Stiggins et al., 2007 and Taylor & Nolen, 2007).

## **DIAGNOSTIC ASSESSMENT FOR STUDENTS WITH SPECIAL NEEDS**

Diagnostic assessments play a critical role in the identification and instruction of students with special needs (Fuchs & Fuchs, 1986). For convenience in some of the following discussion we group students in special education programs and English language learners (ELLs) together because, although assessment accommodations may differ for these groups of students, several diagnostic assessment issues, procedures, and recommendations can be generalized across these groups of students.

### **DIAGNOSTIC ASSESSMENT FOR STUDENTS IN SPECIAL EDUCATION PROGRAMS**

Diagnostic assessment tools play a critical role in the identification of students in need of special education services. The recent reauthorization of the Individuals with Disabilities Education Act (IDEA; 2004) recognized a strategy called “response to intervention” (RTI) as a potential procedure for identifying students in need of special services. RTI relies on an integrated assessment and instruction strategy to deliver and monitor the effects of precisely designed instruction to students at-risk for failure. Diagnostic assessment tools provide the necessary information for determining the instructional needs of these students.

RTI is a process of systematically using assessment results to design, monitor, and adjust instruction to meet students’ needs. Screening tests are administered to *all* students to determine whether or not they are at risk. Those students whose performance indicates that they are not on target for achieving instructional benchmarks are given diagnostic assessments, to determine their misconceptions or skill deficits in a content area. Because these students may have significant deficits that are not easily remedied by typical classroom instruction, diagnostic assessment tools provide valuable information about students’ misconceptions or skill deficits. Teachers can use this information to develop varied instructional interventions that are tailored to each student’s needs (Fuchs, Fuchs, Hosp, & Hamlett, 2003).

Determining the instructional interventions or strategies students need to compensate for misconceptions or skill deficits is the primary purpose of the RTI diagnostic assessment process. Diagnostic assessment results should differentiate between students' *slips* in thinking and persistent *bugs*. Slips are random errors in students' declarative or procedural knowledge that are not the result of inherent misconceptions or skill deficits in the content area. Bugs, however, represent persistent misconceptions about domain specific knowledge or skill deficits that consistently interfere with students' learning. Identifying bugs in student thinking or skills is the intent of the RTI diagnostic assessment process.

Many students who struggle in academic content areas have inconsistent response patterns that make it difficult to diagnose causes. To provide instructionally relevant information, diagnostic assessment tools should be strategically designed to adequately reflect students' conceptual understanding and skills in the domain. Essential prerequisite knowledge and skills should be sufficiently sampled to provide a clear representation of what students know and are able to do. Items should be written to provide detailed information about students' persistent misconceptions or skill deficits. These technical requirements make several assumptions about the diagnostic assessment tool: a) content aligns with a cognitive model,<sup>1</sup> b) sub-score reliability is sufficient to be able to depend upon students' scores, and c) item responses provide information about misconceptions or skill deficits patterns.

In an RTI model, once instructional intervention decisions have been made and implemented for at-risk students, their responses to the instruction are monitored. Progress monitoring assessment tools are administered to determine whether or not the instructional design and delivery decisions are appropriately aligned with students' needs, as evidenced by their growth rates. If students are not making adequate progress, additional diagnosis is done and additional interventions are planned, implemented, and monitored. Diagnostic assessment tools used for RTI provide information about students' progress, as well as the effectiveness of interventions.

## **DIAGNOSTIC ASSESSMENT FOR ENGLISH LANGUAGE LEARNERS**

Formative and diagnostic assessment tools and processes must be designed and administered in such a way that differences in language ability do not impede the evaluation of students' skills and content area knowledge. The key challenge in assessment for ELL students is making sure that the content of interest is being measured and not some other aspect of language knowledge or ability. It is critical to avoid confusing language learning with issues of academic knowledge and achievement. Language issues may be particularly relevant to consider in the arena of diagnostic assessment, when a misdiagnosis of learning needs may lead to an inappropriate learning intervention.

---

<sup>1</sup> A 'cognitive model' is a theory about the progression of understanding and skill necessary to make progress in a content area such as reading.

An important prerequisite step in adapting assessment tools for ELL students is the explicit specification of which skills and abilities are representative of the construct of interest, and which skills may be embedded in the item or task that are not directly relevant to what is being measured. For example, if the ability to apply mathematics to real world situations is the targeted mathematics skill, then context is a critical component of the test. However, ELL students may struggle with reading and be unable to demonstrate their ability to solve problems. If the problems are translated or presented orally, this change in presentation may allow them to demonstrate their knowledge and skills. Oral presentations and translations, in such cases, are accommodations. They change the mode of presentation but do not change the content being measured. In contrast, if an English language skill is not related to the content being assessed, then an accommodation is unlikely to ameliorate the impact of this skill on performance. For example, if on the test of mathematics problem-solving students are required to use written language to describe their problem-solving process when their process would be demonstrated more accurately using numeric, symbolic, or graphic means, then having an accommodation such as a scribe could result in an invalid score on the test.

Research by Abedi et al. (2004) demonstrates that a key issue in the design and use of assessment tools for ELL students is the need to make sure that, on tests of content other than the language arts, the reading and language requirements of the assessments are made as simple and accessible as possible. The use of “simplified language,” “modified language,” or “plain language” is intended to reduce the reading level and to increase the accessibility of an assessment tool to a nonnative English speaker. Research has shown that this accommodation helps both English language learners and native English speakers (Abedi, Lord, Hofstetter, & Baker, 2000). Abedi et al. also say that the most promising accommodations for ELL students include provision and use of customized dictionaries and glossaries and using modified English. Modified English revises the test item language to reduce language complexity without changing the fundamental content of the test item.

Additional training of those who score or rate assessments may also be needed to ensure valid assessment of ELL students. Shaw (1997) found that while most responses were reliably scored, ELL spelling and syntax on certain responses were significant sources of error. Shaw recommended using raters who are knowledgeable about typical patterns in written English used by ELL students. Another recommendation was that, as new assessments are developed, the use of performance items be exploratory, pending evidence for their reliability and validity with ELL students (Shaw, 1997).

## **ASSESSMENT ACCOMMODATIONS**

Anytime an assessment is administered, some test-takers may have cognitive, sensory, physical, or language issues that interfere with interpretation of the assessment results. Physical disabilities may influence a student’s ability to demonstrate his or her knowledge and skills on the test. As

such, test scores may not accurately reflect the student's understanding (or misunderstanding) in the content area. For diagnostic assessment processes, incorrect interpretation could lead to inappropriate assignment of instructional interventions or remediation strategies. To more precisely determine misconceptions and skill deficits for students with challenging personal attributes, accommodations can be applied to the test administration. The *AERA/APA/NCME Standards for Educational and Psychological Testing* (American Educational Research Association et al., 1999) defines an accommodation as "...the general term for any action taken in response to a determination that an individual's disability requires a departure from established testing protocol" (AERA, et al., 1999, p. 101).

Accommodations are designed to maintain the integrity of the tested construct so that interpretations of test results do not differ for students taking the accommodated test, as compared to the non-accommodated test. Effective accommodations should not materially alter the nature of the task or the required response, and they should yield scores that are valid indicators of the construct being assessed.

Possible accommodations include changes to the presentation of material, student's response method, timing or schedule of administration, or setting of test administration. Presentation accommodations include changes to the format in which test items are delivered to students, such as presenting material in Braille or magnifying text. Response mode accommodations include changes in the manner in which students respond to test items, such as providing assistive technology devices or allowing students to dictate their responses. Timing accommodations change the amount of time or distribution of time allowed to complete the test. Changes in the schedule for an assessment might include testing at times during the day in which the student is most productive. For example, students might be provided with additional time to take a test or the testing session might be broken into multiple shorter sessions. Setting accommodations require changes in the physical setting in which students take tests. These accommodations include providing a testing environment that is free from distractions such as noise or other students.

An Individualized Education Program (IEP) team typically assigns accommodations by considering the student's personal characteristics in light of the targeted construct. IEP teams must understand the construct, so as to avoid providing accommodations that detract from the valid interpretation of results.

In applying accommodations during formative assessment events, it is important to match accommodation decisions to the intended purpose of the assessment tool. For example, if the assessment results will be used to predict and track student progress toward achieving a state standard (an interim assessment purpose), the accommodations used should closely match those used for the state test. On the other hand, if the purpose of the formative assessment is more directly focused on learning improvements in the classroom, then greater flexibility in the choice

and application of accommodations may be warranted. However, even in the classroom accommodations must be designed to minimize the influence of disabilities or language demands, rather than leading to inaccurate assessment results. Only then can the assessment results help teachers and students determine whether or not students are learning the targeted knowledge and skills.

### **TEST DESIGN FOR STUDENTS WITH SPECIAL NEEDS**

When choosing or developing a formative or diagnostic assessment, a number of considerations will aid the applicability and interpretability of the assessment results for students with special needs. The principle of universal design can be applied to assessments used for students in special education program or ELL students. Universal design asserts that assessments should be designed so that the greatest number of people can use them without the need for modification. To achieve this goal, unnecessary obstacles must be eliminated.

To maximize universal design, developers of diagnostic and formative assessments should consider the needs of students with disabilities and English language learners when designing their assessments and making decisions about such issues as time limits, wording of test items, and response formats. One of the most common accommodations, extra time, has been shown to improve performance for general education students as well as students with disabilities and English language learners (Abedi, Hofstetter, & Lord, 2004; Elliott, Kratochwill, & McKeivitt, 2001; Zuriff, 2000). Careful consideration of the amount of time required to complete a test (or whether or not time limits are needed at all) may reduce the need for extended time accommodations. Research has also shown that language simplification helps both English language learners and native English speakers (Abedi, Lord, Hofstetter, & Baker, 2000), suggesting that greater attention should be expended on ensuring that assessments use language that is maximally accessible.

Test developers should also include special education students and English language learners during the field testing of assessment tasks. Field testing provides critical information about the performance of the assessment, and inclusion of students from these groups will help identify problems during the earliest stages of test development. In tests using normative samples for comparisons it also may be important to ensure that students from these subgroups are represented in the normative sample proportionately.

## ISSUES IN THE USE AND INTERPRETATION OF DIAGNOSTIC ASSESSMENTS

In this section of the *Guide* we discuss a number of issues in the choice, evaluation, use, and interpretation of formative assessment instruments. There are a number of excellent resources that can provide further information on these topics (see Appendix).

### CRITERIA FOR CHOOSING AN ASSESSMENT

The *Standards for Educational and Psychological Testing* (AERA, et al, 1999) provides extensive guidelines for the effective and responsible use of assessments. The *Standards* contain detailed information on best practices in test planning, test design and development, administration, security, and test use and interpretation. An important component of the *Standards* is focus on the technical aspects of test development, use, and interpretation. Users are encouraged to consider a range of criteria in deciding which assessment(s) to use. One of the most important criteria is the match of an assessment tool to the assessment purpose. Test developers and publishers can sometimes be overly optimistic in describing the breadth of applications of their assessment tools. However, assessment tools seldom work well for all purposes. Assessment tools must be designed and developed in one way for one purpose, and in a different way for another purpose.

Another critical consideration in choosing an assessment tool is the alignment of the content and skills on the assessment tool to curricular content and standards. The purpose of the Washington State diagnostic assessment legislation is to support assessment processes that lead to improved student learning of the Essential Academic Learning Requirements (EALRs) and the associated Grade Level Expectations (GLEs). As a result, it is important that the assessment tool shows alignment to the Washington State curriculum standards. Users need to seek additional information or conduct their own evaluation, to determine if the content alignment of an instrument is sufficient for the desired assessment purpose.

Another important feature of instrument design that users should consider, is the relative specificity and detail provided in score reporting. Sometimes an assessment tool only presents information at a global or overall level (e.g., “mathematics concepts” and “mathematics computation”). While more global score reporting at this level may be sufficient for summative purposes, lack of specificity undermines the utility of formative assessment, diagnosis, and feedback. Greater specificity provides the basis for more targeted intervention and more focused feedback to the student or teacher. Therefore, users should critically examine the kinds of information and score reports that will be provided by an assessment tool to determine whether or not it will meet user needs.

Critical review of the technical properties of an assessment is very important before adopting an assessment (see below). Users should review information on the stated purpose and development of an instrument to determine whether or not it matches user needs. Users should critically examine evidence that the test developer or publisher has obtained independent reviews of the

instrument to ensure it is sensitive to all test takers, and that it is not biased against protected groups of students.

## **TECHNICAL QUALITY**

Examination of evidence for reliability and validity of the use and interpretation of assessment results should be a paramount concern for all those who use assessments. While many people do not like to deal with technical issues involving formulas, psychometrics, and statistics, how well an assessment works, and therefore how effectively it supports and enhances student learning, depends on the technical adequacy of the assessment tool and the assessment results. We briefly discuss here major aspects of reliability and validity, as well as the need for technical quality in test construction, reporting, and the review of bias and sensitivity in test use.

### ***EVIDENCE FOR RELIABILITY***

Reliability in assessment refers to the consistency of results across different evaluators, occasions, tasks, or forms of the assessment. If no learning changes have occurred, assessment results should not vary substantially, regardless of the evaluator, the day of testing, or the test form (in the case of multiple forms of a test). If results from an assessment tool are not reliable, then the results cannot be trusted; they are likely to vary depending on changes in the conditions of the assessment event rather than differences in the student's skills or abilities. So for example, if a student's scores depend on who gives the assessment, which day the assessment is given, or which test form is administered, the resulting scores are unreliable.

Several distinct sources of unreliability are usually defined, and it is the responsibility of the test developer to minimize the effects of these sources. Evidence for reliability should be provided for each use of an assessment instrument. One way of estimating the reliability of results is called *internal consistency*. Internal consistency refers to the consistency with which examinees respond to the different items on the assessment. If responses to items measuring the same knowledge or skill are highly inconsistent, then a measure of internal consistency would be diminished. This kind of reliability can be maximized by careful analysis of the items when assessment tools are developed, to ensure that items function well together.

A second method for gathering evidence for reliability is to determine whether or not examinees would get the same results if they took two different forms of a test (i.e., *alternate forms reliability*). If two test forms differ in difficulty or content, they are not comparable, and reliability will be diminished. This measure of reliability can be optimized during test development, if careful steps are taken to ensure that all forms of the test are developed using the same test blueprint, and selecting items for each sub-skill that are about the same level of difficulty.

A third method for estimating reliability is to examine the consistency of those who assign scores

to students' responses (i.e., *inter-rater reliability* or *inter-judge agreement*). If one rater or teacher assigns a different score to a student performance than a second rater or teacher, then part of the score is associated with who did the scoring rather than how well the student performed. There also may be inconsistencies that occur when only a single rater or scorer is used, due to fatigue or gradual changes in way the rater uses the scoring criteria. Careful specification of assessment goals and criteria, using clear and specific scoring keys or rubrics, training and practice with previously scored sample or model papers, and occasionally having two scorers rate the same student responses, are procedures for enhancing intra and inter-rater reliability.

Another method for gathering evidence of reliability is commonly referred to as *test-retest* reliability. For this method of reliability the issue of concern is whether or not the same assessment results would be obtained if the assessment tool were administered to the same students at more than one point in time. Over short periods of time, before learning or development has occurred, different administrations of an assessment tool should produce the same or similar results.

Reliability of assessment results is most often evaluated with statistical analyses that produce a correlation or similar index of the degree of consistency of measurement. Such indices typically range from 0, completely unreliable, to 1.00, perfectly reliable. There is no strict cutoff value for reliability estimates. The degree of reliability expected should be matched with the importance of the use of the assessment results—the more important the usage, the higher the expected measure of reliability. Rules of thumb should never be interpreted strictly, but estimates of .85 or higher are considered good, and reliability estimates of .90 or higher are recommended for important, high-stakes uses of assessment results (e.g., placement or classification decisions; Henson, 2001; Nunnally & Bernstein, 1994).

There are some important relationships between reliability and validity. If reliability is the consistency of measurement, validity is the accuracy of measurement. Reliability is prerequisite to validity. If measurement is inconsistent, it is difficult to be accurate. It is also possible to have high reliability but little or no validity. For example, a ruler can give perfectly consistent measurement, but if it is an inch short, it is never accurate. Finally, reliability puts an upper limit on validity. Assessment accuracy depends on a certain level of dependability in the assessment results.

What aspects of reliability are important in formative and diagnostic assessment? Because most formative and diagnostic assessments require repeated measurement over time to monitor and evaluate student progress, two of the more important measures of reliability are test-retest and alternate-forms reliability. If multiple raters or scorers are used in evaluating results, evidence for inter-rater reliability is important as well.

## ***EVIDENCE FOR VALIDITY***

Validity refers to how accurately an assessment tool measures the specific skill or conceptual understanding it is designed to measure, and whether or not the results, conclusions, and inferences derived from the assessment tool are accurate (Messick, 1989; 1994; 1995). Evaluation of validity includes consideration of how well the assessment results serve their intended purpose, and whether or not the assessment results are used and interpreted correctly. A number of different types of evidence can be gathered to support the validity of an assessment.

***Content-related evidence*** for validity is gathered by determining whether or not the content of an assessment tool is appropriate for its stated purpose. The sample of items, tasks, or performances in an assessment tool should represent the important content, skills, or behaviors of the domain of interest. Content-related evidence for validity is usually obtained by having a panel of experts judge whether or not items on the assessment tool are relevant and fully representative of the content domain. For example, to gather content-related evidence for validity of a 3<sup>rd</sup> grade mathematics test, experts would be selected (e.g., elementary math teachers) and asked to provide ratings on how well each item matched the mathematics curriculum for 3<sup>rd</sup> graders. Using alignment studies to evaluate whether or not state's standards-based tests match the state's content standards is another method of obtaining content-related evidence for validity.

***Criterion-related evidence*** for validity refers to evidence that a test can predict performance on some current or future standard or criterion of performance. When the prediction between the test and the criterion is measured at a single point in time it is called ***concurrent evidence*** for validity and when the test is used to predict performance at a later point in time it is called ***predictive evidence*** for validity. Typically this criterion-related evidence for validity is evaluated using correlational statistics; the higher the correlation, the stronger the evidence that the test can predict the criterion performance of interest. For example, students' scores from a 4<sup>th</sup> grade standardized reading test could be correlated with the students' classroom grades in reading (concurrent evidence); SAT/ACT scores during high school could be correlated with first year college grade point averages (predictive evidence).

The most general and overarching type of evidence for validity is ***construct-related evidence***, which refers to how well the construct of interest is being measured. There are many ways to gather construct related evidence for validity. ***Convergent evidence*** for validity demonstrates that test scores are related to behaviors and other assessments that are indicators of the same construct. Criterion-related evidence and content-related evidence are both types of convergent evidence for validity. ***Discriminant evidence*** for validity shows that test scores are *unrelated* to behaviors and test scores that are indicators of different constructs.

For example, construct-related evidence can be obtained by showing that student scores on a reading test correlate highly with the students' scores on another reading test (convergent evidence) and correlate much lower with their scores on a mathematics test (discriminant

evidence). Another way to gather convergent evidence for validity is to show that there are differences in test scores between groups of students who should differ in their performance on the test. For example, there should be substantial differences in performance for students who have completed an instructional unit when compared to students who are just starting the unit.

What types of evidence for validity are important to gather when using formative and diagnostic assessment? Different types of evidence may be more or less important, depending on the purpose and use of the assessment results. For example, if the primary purpose of a formative assessment is to predict how well the students are likely to do on the state test, predictive evidence would be one of the most important kinds of evidence to gather. In many applications of formative assessment, content-related evidence for validity is important – especially in demonstrating that assessment tasks and items are closely tied to local curricula and are specific and extensive enough to support detailed assessment of student strengths and weaknesses.

Ultimately, the most important validity issue is whether or not the use and interpretation of assessment information leads to accurate decisions about how to support student learning, adapt instruction to learning needs and properly intervene, to allow students to reach their full potential. Evidence that the assessment results will support these uses is the most important evidence needed.

Gathering evidence for validity of assessment results is not solely the responsibility of the assessment developer or test publisher. Any user of an assessment tool should gather evidence to determine whether or not the assessment results support the planned interpretation and use. Studies are needed to determine whether or not scores are valid across individuals, groups, instructional interventions, and contexts. In that sense, validity studies are an ongoing responsibility of assessment developers and users. This is done by monitoring and evaluating the success of individual students, as well as the performance of the assessment system overall, to determine whether or not the consequences of interpretation and use of assessment results are those that are intended.

### ***TEST FORMS, SCORES, AND REPORTS***

Another indication of the utility and appropriateness of an assessment tool is the match between the design and features of the instrument and its intended use and purpose. The number of assessment forms available should match plans for the frequency of administration. Most formative assessment processes require repeated assessment events and, in the case of progress monitoring, many parallel forms of the assessment are needed. Some assessment tools have only one or a few forms, and are not usable in formative assessment applications where reuse of the form can lead to over-familiarity, memorization, or teaching to the test. When choosing a formative assessment tool, it is important to verify that a sufficient number of forms are available and to ascertain that technical work has been completed to ensure comparability of the forms. Each form should represent curricular content appropriately and the forms should be equated for

difficulty to ensure that differences from one form to another are due to true proficiency differences and not differences in the test. On the other hand, screening and diagnostic assessment purposes may not require multiple assessment forms. One or two forms of an assessment tool may be sufficient for these purposes.

Scores resulting from the assessment, and the design of score reports, should also match assessment purpose. For example, if detailed diagnostic information on student strengths and weaknesses is needed, an assessment tool that only provides general reading skill score (e.g., literal comprehension) will not serve the users' purpose well. For diagnostic assessment purposes, a substantial degree of specificity is needed to provide feedback that is detailed enough to guide instructional decisions, make instructional adjustments, and provide clear direction to students for improvement.

The specificity needed for instructional decision-making also suggests that certain kinds of score information, like percentile ranks or grade equivalent scores, have little utility in formative and diagnostic assessment. Knowing the relative standing of a student in relation to a norm group does not help users identify learning needs or progress toward learning goals. Assessment tools should be chosen that provide results in a metric that is understandable to students and teachers and that can easily be related to progress on a continuum toward learning targets. Similarly, assessment reports should be designed to clearly communicate the progress of learning and the relation of performance to learning goals and targets.

### ***BIAS AND SENSITIVITY REVIEWS***

A basic principle of assessment development and score interpretation and use is a commitment to fairness and accuracy (see *Code of Fair Testing Practices*, 2004). Assessment developers and users must ensure that all students have an equal opportunity to demonstrate their knowledge and skills and that construct-irrelevant test design, characteristics, or procedures do not result in the differential performance of test-takers with the same ability. In reviewing and choosing an assessment tool, it is very important to determine whether or not the test developer has conducted thorough reviews of the assessment for test bias and for sensitivity.

Bias is the presence of some characteristic of an assessment, a test item, or task in the assessment, that results in different performance for two individuals who have the same knowledge and skill, but who are from different student subgroups. Test bias can be minimized or prevented through careful test development processes including clear specification of the content to be measured and the training of item writers. However, no matter how careful the test development, field-testing and item analysis (e.g., Differential Item Functioning or DIF) must be conducted to gather evidence for potential sources of bias. Items identified as showing systematic differences between groups of test-takers are usually reviewed by panels of diverse, independent stakeholders, who provide advice and recommendations on item appropriateness.

Sensitivity refers to the appropriateness of test language, content, and design for all test-takers. The goal of sensitivity review is to ensure that the assessment is accessible and respectful of all people and does not unfairly disadvantage or disturb the test-taker. Sensitivity review is intended to eliminate language or topics that are inflammatory, controversial, insulting, and/or slanted. Sensitivity review is usually incorporated into the test development process but should also be augmented by a sensitivity review panel. The review panel should be composed of independent reviewers who broadly represent a variety of community groups. The goal of the review is to ensure sensitivity to different gender, cultural, religious, ethnic, socio-economic, and disability groups, as well as to avoid items, text, or topics that may elicit strong or negative reactions, or emotions from students that impede or interfere with their performance (Zeiky, 2006). Test users should review technical documents for published diagnostic and formative assessments to determine whether or not bias and sensitivity reviews and DIF studies have been conducted, to ensure the validity of assessments for all students.

## **IMPLEMENTATION, USE AND INTERPRETATION**

Implementation of new assessment systems or tools by a teacher, school or district can be challenging. We briefly discuss here a number of difficulties, problems, and pitfalls that are common in current assessment practice, or that may occur in the implementation of a new assessment system. We then focus on several suggestions from the literature for effective implementation of formative assessments.

### ***DIFFICULTIES, PROBLEMS, AND PITFALLS***

The assessment literature (e.g., Amrein & Berliner, 2002; Barton, 1999; Black & Wiliam, 1998b; Cizek et al., 1995; Dorn, 1998; Heubert & Hauser, 1999; Popham, 1999; Stevens, et al. 2000) describes a number of difficulties associated with current use and implementation of tests and other assessment tools. These difficulties include issues in assessment design (e.g., wrong test type for stated purpose; technical adequacy at a different level than the inferences made; tests that measure construct irrelevant skills; confusion of NRTs with CRTs), assessment implementation (e.g., lack of time; delayed access to results; use of tests that do not support the assessment purpose; teaching to the test), interpretation and use of assessment results (e.g., misinterpretation of test results; drawing conclusions not supported by the results), resources for assessment use and interpretation (e.g., need for greater assessment literacy of participants, professional development; funding for test development and implementation; funding to ensure technical adequacy), and consequences of implementation (e.g., narrowing the curriculum; teaching to the test).

Several authors describe weaknesses in current assessment practices that directly undermine learning and instructional effectiveness including: a) tests that emphasize superficial learning and recall, b) teachers who appear to be unaware of the assessment work of colleagues and do not trust or use other teachers' assessment results, and c) an emphasis on quantity and presentation of

work rather than on quality of work in relation to learning (Black & Wiliam, 1998b). Both in questioning and written work, research shows that teachers' assessment practices focus too much on low-level knowledge and skills, mainly memorization and recall (Cizek et al., 1995). Cizek, et al. (1995) also say that current assessment practices overemphasize grading functions and underemphasize feedback and advice for learning, focus on competition rather than personal improvement, and use comparative assessment interpretations, ensuring that some students will be labeled as “low performers” or “low ability students.”

Another difficulty that may be embedded in current assessment practices is the inability to use and apply assessment results in support of learning (Cizek, et al., 1995). Teachers' feedback often serves social and managerial functions instead of learning functions. Teachers may be able to predict student performance but know too little about student learning needs or strategies to apply that information for student improvement. Teachers often start “new” every year in assessing students and may not use or may have no information on student performance from previous teachers. Finally, grading is often given higher priority and importance than analysis of student work for learning intervention.

There are some common implementation pitfalls that schools encounter when trying to improve their use of assessment processes for decision making, and providing timely and ongoing feedback to students about their progress, strengths, and areas for improvement. Some of the barriers that teachers face can include lack of time and limited assessment literacy skills. Even if commercially produced assessment tools are used, teachers may not know how to interpret results, communicate results to stakeholders (i.e., students and parents), provide the kinds of descriptive feedback necessary for student improvement, diagnose needs for particular intervention strategies, or implement those strategies.

### ***SUGGESTIONS FOR SUCCESSFUL IMPLEMENTATION***

We review here some suggestions for implementation that have been made specifically for the use of formative and diagnostic assessments (see in particular Black & Wiliam, 1998b; Stiggins, 2007b). ***Assessment design and choice of assessment tool*** is an important prerequisite step to successful implementation. Black and Wiliam (1998b) discuss the importance of refining and clarifying assessment purpose to guide design and use of assessment tools and the use of high quality assessment tools that match learning targets. The *Washington State Comparative Guide* is designed to provide support in making such choices. For progress monitoring assessment tools, it is important to select a tool that provides representative sampling of the content domain, is closely aligned to the delivered curriculum, has enough equated forms to allow for the intended frequency of assessment, provides score reporting that gives detailed feedback, has strong evidence of technical adequacy, and has been screened for bias and sensitivity. For diagnostic assessment tools, it is important that there are sufficient items for each concept or skills area to have reliable sub-skill scores. It is also important that answer choices for multiple-choice items or scoring

protocols for constructed-response items provide information about the sources of learning difficulties. A diagnostic assessment tool may measure fewer skills in a more focused way than a more general classroom assessment tool or a progress-monitoring tool.

Another critical feature of a successfully implemented formative assessment process is the clear linkage of assessment with curriculum and instruction. Teachers should explicitly design feedback strategies that connect assessment results with instructional decision-making and planning for intervention. It is also important to clearly identify and communicate learning targets to students and communicate assessment results and expectations to students during the learning process. A commonly overlooked issue is the need to explicitly design assessments and activities that focus on transfer and generalization of knowledge and skills. This helps to ensure that learning is focused on attainment of the skills and conceptual understanding of interest and not on details of a particular assignment or assessment. Finally, test users should make sure that analysis and reporting of assessment results are at a level of specificity that allows clear and direct linkage of results to instructional intervention.

As described earlier, student involvement is a key component of a formative assessment process. Increased involvement enhances student engagement with content and can strengthen student motivation and self esteem. To ensure involvement, teachers should design methods to regularly use assessment results to provide detailed descriptive feedback to students. Feedback should be clearly linked to expectations for learning. Teachers should also plan ways to use student self assessment and self monitoring as additional interventions for instructional improvement.

It is important to develop a formative assessment process that supports effective use of results. This can be done by providing clear guidelines on the appropriate interpretation and uses of assessment results including explicit discussion of the ways in which assessment results should not be used. Reports should be designed so that they are useful for instructional purposes, provide sufficient detail to inform instruction, provide enough descriptive information for effective feedback to students, and display assessment results in ways that are easy to communicate and understand (e.g., graphs of learning curves). For diagnostic assessment tools, reports should indicate causes of learning difficulties (misconceptions and skill deficits) that interfere with students' progress.

Another suggestion for effective use of a formative assessment system is to design systems for the more ***integrated involvement of teachers*** and ***professional development opportunities*** to aid teachers in using assessment information in appropriate ways. Properly applied, a formative assessment process requires a greater emphasis on feedback useful for learning. This may require changes in classroom practice. In particular, for the full benefits of a formative assessment process to be realized, teachers need to know how to interpret and use assessment results to adjust instruction and to provide detailed descriptive feedback to students. Teachers may not know how to use assessment information in these ways, which necessitates additional teacher support and

professional development for effective implementation. Professional development opportunities must be provided, including pre- and post-assessment training on the use of the system and analysis of reports, data interpretation, and the use of data to inform instruction and specific interventions.

Stiggins (2007) also suggests several school or district level practices to support the effectiveness of implementation of a formative assessment process. First he suggests the establishment of policy that communicates clear achievement expectations for students. He also recommends coordination of assessment systems across the district, and the communication of results in a timely and understandable way. To ensure assessment accuracy, investment must be made in fostering assessment literacy among the participants and in evaluating implementation of the assessment system.

## CONCLUSION

This *Guide* has presented a wide range of information including clear definitions of assessment purposes, research on the use of diagnostic and formative assessment processes, accounting for students with special needs in assessment administration, and the technical issues associated with assessment development and the interpretation and use of results. More information about the issues and ideas presented in this *Guide* can be found in the resources listed in the references and in the appendix that follows.

## REFERENCES

- Abedi, J. (2002). Standardized Achievement Tests and English Language Learners: Psychometrics Issues, *Educational Assessment*, 8(3), 231–257.
- Abedi, J., Hofstetter, C. H., & Lord, C. (2004). Assessment Accommodations for English Language Learners: Implications for Policy-Based Empirical Research, *Review of Educational Research*, 74(1), 1–28.
- Abedi, J., Lord, C., Hofstetter, C., & Baker, E. (2000). Impact of accommodation strategies on English language learners' test performance. *Educational Measurement: Issues and Practice*, 19(3), 16–26.
- American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- Amrein, A. L. & Berliner, D. C. (2002). High-stakes testing, uncertainty, and student learning. *Education Policy Analysis Archives*, 10(18). <http://epaa.asu.edu/epaa/v10n18/>
- Bangert-Drowns, R. L., Kulick, J. A. & Morgan, M. T. (1991a). The instructional effect of feedback in test-like events), *Review of Educational Research*, 61, 213-238.
- Bangert-Drowns, R. L., Kulik, J. A. & Kulik, C.-L. C. (1991b). Effects of frequent classroom testing, *Journal of Educational Research*, 85, 89-99.
- Barton, P. (1999). *Too much testing of the wrong kind: Too little of the right kind in K-12 education*. Princeton, NJ: Educational Testing Service.
- Biggs, J. (1998). Assessment and classroom learning: A role for summative assessment? *Assessment in Education*, 5, 103-110.
- Black, P. J., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom, *Phi Delta Kappan*, 86(1), 9-21.
- Black, P. J. & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, 5(1), 7-73.
- Black, P. J. & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan*, 80, 139-144.
- Bloom, B. S., Hastings, J. T., & Madeus, G. F. (1971). *Handbook on formative and summative evaluation of pupil learning*. New York: McGraw-Hill.
- Boston, Carol (2002). The concept of formative assessment. *Practical Assessment, Research & Evaluation*, 8(9). <http://PAREonline.net/getvn.asp?v=8&n=9>

- Brookhart, S. M. (2001). Successful pupils' formative and summative uses of assessment information. *Assessment in Education*, 8, 153-169.
- Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In J. H. McMillan (Ed.), *Formative Classroom Assessment: Theory Into Practice* (pp. 43-62). New York: Teachers College Press.
- Brown, S., Rust, C., & Gibbs, G. (1994). *Strategies for Diversifying Assessment in Higher Education*. Oxford: Oxford Center for Staff Development.
- Cizek, G. J., Fitzgerald, S. M. & Rachor, R. E. (1995). Teachers' assessment practices: preparation, isolation and the kitchen sink, *Educational Assessment*, 3, 159-179.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests*. Thousand Oaks, CA: Sage.
- Code of Fair Testing Practices in Education*. (2004). Washington, DC: Joint Committee on Testing Practices. <http://www.apa.org/science/fairtestcode.html>
- Crooks, T. J. (1988). The impact of classroom evaluation practices on students, *Review of Educational Research*, 58, 438-481.
- Dorn, S. (1998). The political legacy of school accountability systems. *Education Policy Analysis Archives*, 6(1), 1-34.
- Elliott, S. N., Kratochwill, T. R., & McKeivitt, B. C. (2001). Experimental analysis of the effects of testing accommodations on the scores of students with and without disabilities, *Journal of School Psychology*, 39(1), 3-24.
- Fuchs, L. S. (1993). Enhancing instructional programming and student achievement with curriculum-based measurement. In J. J. Kramer (Ed.), *Curriculum-Based Measurement* (pp. 65-1030). Lincoln, NE: Buros Institute of Mental Measurements.
- Fuchs, L. S. & Fuchs, D. (1986). Effects of systematic formative evaluation: a meta-analysis, *Exceptional Children*, 53, 199-208.
- Fuchs, L. S., Fuchs, D., Hosp, M. K., & Hamlett, C. L. (2003). The potential for diagnostic analysis within curriculum-based measurement. *Assessment for Effective Intervention*, 28, 13-22.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes. *American Psychologist*, 18, 519-521.
- Harlen, H., & Crick, R. D. (2003). Testing and motivation for learning, *Assessment in Education*, 10(2), 169-207.

- Henson, R. K. (2001). Understanding internal consistency reliability estimates: A conceptual primer on coefficient alpha. *Measurement and Evaluation in Counseling and Development, 34*, 177-189.
- Heubert, J.P., & Hauser, R.M. (Eds.) (1999). *High Stakes: Testing for tracking, promotion, and graduation*. Washington, DC: National Academy Press.
- Johnson, D.W., & Johnson, R.T. (1990). Co-operative learning and achievement. In: S. Sharan (Ed.), *Co-operative Learning: theory and research* (pp. 23-27). New York: Praeger.
- Kellough, R. D., & Kellough, N. G. (1999). *Middle School Teaching: A Guide to Methods and Resources* (3<sup>rd</sup> Ed.). Upper Saddle River, NJ: Merrill.
- Kulik, C.-L. C., Kulik, J. A. & Bangert-Drowns, R. L. (1990). Effectiveness of mastery-learning programs: a meta-analysis, *Review of Educational Research, 60*, 265-299.
- Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership, 63*(3), 19-24.
- Lewis, A. (2006). *Celebrating 20 years of research on educational assessment: Proceedings of the 2005 CRESST Conference* (CSE Technical Report 698). Los Angeles, CA: CRESST.
- McManus, S. (2008). *Attributes of Effective Formative Assessment*. Washington, DC: Council of Chief State School Officers.
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice, 20*(1), 20–32.
- McMillan, (2007). Formative classroom assessment: The key to improving student achievement. In J. H. McMillan (Ed.), *Formative Classroom Assessment: Theory Into Practice* (pp 1-7). New York: Teachers College Press.
- Messick, S. (1975) The standard problem: meaning and values in measurement and evaluation, *American Psychologist, 30*, 955-966.
- Messick, S. (1989). Validity. In R.L. Linn (Ed.), *Educational measurement* (3<sup>rd</sup> ed., pp. 13-103). New York: American Council on Education.
- Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13-23.
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

- Natriello, G. (1987). The impact of evaluation processes on students, *Educational Psychologist*, 22, 155-175.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.
- Popham, J. (1999). Why standardized tests don't measure educational quality. *Educational Leadership*, 56(6), 8-15.
- Popham, W. J. (2001). *The truth about testing: An educator's call to action*. Alexandria, VA: Association for Supervision and Curriculum Development.
- Ramaprasad, A. (1983). On the definition of feedback, *Behavioural Science*, 28, 4-13.
- Rosenshine, B., Meister, C. & Chapman, S. (1996). Teaching students to generate questions: a review of the intervention studies, *Review of Educational Research*, 66, 181-221.
- Sadler (1983). Evaluation and the improvement of academic learning. *Journal of Higher Education*, 54, 60-79.
- Sadler, R. (1989). Formative assessment and the design of instructional systems, *Instructional Science*, 18, 119-144.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagne, & M. Scriven (Eds.), *Perspectives of Curriculum Evaluation* (pp. 39-83). Chicago: Rand McNally.
- Shaw, J. M. (1997). Threats to the validity of science performance assessments for English language learners, *Journal of Research in Science Teaching*, 34(7), 721-743.
- Stevens, J. J. (1995). Confirmatory factor analysis of the Iowa Tests of Basic Skills. *Structural Equation Modeling: A Multidisciplinary Journal*, 2(3), 214-231.
- Stevens, J. J., Estrada, S., & Parkes, J. (2000). *Measurement issues in the design of state accountability systems*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.
- Stiggins, R. L. (1999). Assessment, pupil confidence, and school success. *Phi Delta Kappan*, 81, 191-198.
- Stiggins, R., Arter, J. A., Chappuis, J., & Chappuis, S. (2007). *Classroom Assessment for Student Learning: Doing It Right--Using It Well*. Prentice-Hall.
- Ravitz, Jason (2002). CILT2000: Using Technology to Support Ongoing Formative Assessment in the Classroom, *Journal of Science Education and Technology*, 11(3), 293-296.

- Taylor, C. S. & Nolen, S. B. (2007). *Classroom Assessment: Supporting Teaching and Learning in Real Classrooms*. Columbus, OH: Merrill-Prentice Hall.
- Zeiky, M. (2006). Fairness review in assessment. In S. M. Downing and T. M. Haladyna (Eds.), *Handbook of test development* (pp. 359-376). Mahwah, NJ: Erlbaum.
- Zuriff, G.E. (2000). Extra examination time for students with learning disabilities: An examination of the maximum potential thesis. *Applied Measurement in Education*, 3, 99-117.

## ***APPENDIX: RESOURCES FOR EDUCATORS INTERESTED IN FORMATIVE ASSESSMENT***

In this appendix, we list a number of resources and links to internet sites that may be useful to educators interested in formative assessment and related topics.

### **INFORMATION ON LOCATING ASSESSMENT INSTRUMENTS:**

American Educational Research Association (AERA) FAQ/Finding Information About Psychological Tests: <http://www.apa.org/science/faq-findtests.html>

Buros Institute of Mental Measurements website on testing:  
<http://www.unl.edu/buros/bimm/index.html>

To determine if there is a Buros review for a particular test, go to the following web address:  
<http://buros.unl.edu/buros/jsp/search.jsp>

The ERIC/AE Test Locator can be found at <http://www.ericae.net/testcol.htm>.

The *ETS Test Collection* is an extensive library of more than 25,000 tests and assessments:  
<http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=e462d3631df4010VgnVCM10000022f95190RCRD&vgnnextchannel=85af197a484f4010VgnVCM10000022f95190RCRD>

The University of Chicago Library also has a useful test collection at  
<http://www.lib.uchicago.edu/e/su/tests/>.

### **OTHER ASSESSMENT AND MEASUREMENT RESOURCES:**

**The ABC's of School Testing** (<http://www.apa.org/science/jctpweb.html>)

A videotape developed by the Joint Committee on Testing Practices (JCTP) and a collaboration of several other testing organizations. Designed to help parents understand the many uses of testing in schools today. In addition to the videotape, two publications are also included: *Leader's Guide* and the *Code of Fair Testing Practices*.

AERA Position Statement on High-Stakes Testing in Pre-K – 12 Education:  
<http://www.aera.net/policyandprograms/?id=378>

The Assessment Training Institute provides newsletter articles and other publications about classroom and formative assessment as well as videos and training sessions for a fee.  
<http://www.assessmentinst.com/>

The Center for Research on Evaluation, Standards, and Student Testing (CRESST) has many useful resources and publications:

CRESST products and resources: <http://www.cse.ucla.edu/products.html>

CRESST newsletters (<http://www.cse.ucla.edu/products/newsletters.asp>) offer full texts of the organization's activities and policy views since Fall 1991

CRESST policy briefs provide guidance to educators and policy makers:  
<http://www.cse.ucla.edu/products/policy.html>

CRESST technical reports: <http://www.cse.ucla.edu/products/reports.asp>

Ericae.net contains valuable information about assessment, evaluation, and research:  
<http://ericae.net/nintbod.htm>

FAST is a part of the CCSSO Formative Assessment Initiative from the Council of Chief State School Officers. They have several reports available:  
<http://www.ccsso.org/projects/scass/Projects/Formative%5FAssessment%5Ffor%5FStudents%5Fand%5FTeachers/>

National Center on student progress monitoring: <http://www.studentprogress.org/default.asp>

National Council on Measurement in Education (NCME) has a series called ITEMS: The Instructional Topics in Educational Measurement Series. The goal of ITEMS is to improve the understanding of educational measurement principles by providing brief instructional units on timely topics in the field, modules developed for use by college faculty and students as well as by workshop leaders and participants. <http://www.ncme.org/pubs/items.cfm>

The National Education Association (NEA) website has a number of publications and resources on assessment:

NEA Teacher Toolkit is a suite of Web-based classroom tools designed by NEA members for teachers: <http://www.nea.org/marketplace/ttk.html>

#### [Peer Assessment Teaches Students How to Think](#)

A veteran teachers reflects on the value of students' self-evaluations and peer assessment.  
<http://www.nea.org/teachexperience/ifc070501.html>

#### [Accountability and Testing - Balanced Assessment Report](#)

More and more, state and federal legislators and education policy makers are relying on multiple, large-scale standardized testing programs to measure student ...  
<http://www.nea.org/accountability/balanced.html>

#### [Accountability and Testing - Assessment](#)

NEA has long supported comprehensive assessment of students' learning. In fact, NEA policy states that "a student's level of performance is best assessed with ...  
<http://www.nea.org/accountability/assessment.html>

The National Research Council (2001) has produced a book on classroom assessment in science, *Classroom Assessment and the National Science Education Standards*, that includes information on and examples of applications of formative assessment: <http://www.nap.edu/catalog/9847.html>.

Northwest Regional Educational Laboratory provides an extensive professional development toolkit on assessment: <http://www.nwrel.org/assessment/toolkit98.php>

Rights and Responsibilities of Test Takers: Guidelines and Expectations  
<http://www.apa.org/science/ttrr.html>

The Standards for Educational and Psychological Testing  
<http://www.apa.org/science/standards.html>

### **LISTSERVS RELATED TO ASSESSMENT AND MEASUREMENT:**

Subscribe to: [AERA-D](#) - Sponsored by the AERA division that studies educational measurement and research methodology.

[Send e-mail to: [LISTSERV@ASUACAD.BITNET](mailto:LISTSERV@ASUACAD.BITNET) with message: Subscribe AERA-D yourfirstname yourlastname (omit signature)]

Subscribe to: [ASSESS](#) - Discussion on assessment in higher education.

[send e-mail to: [LISTSERV@LSV.UKY.EDU](mailto:LISTSERV@LSV.UKY.EDU) with message: Subscribe ASSESS yourfirstname yourlastname (omit signature)]

Subscribe to: [ASSESS-P](#) - Sponsored by the Psychological Assessment/Psychometrics Forum at St. John's University. Topics include clinical and research settings, psychometric theory and application.

[Send e-mail to: [LISTSERV@SJUVM.STJOHNS.EDU](mailto:LISTSERV@SJUVM.STJOHNS.EDU) with message: Subscribe ASSESS-P yourfirstname yourlastname (omit signature)]

Subscribe to [ARN-L](#) - Assessment Reform Network - Sponsored by FairTest and ERIC/AE

[Send e-mail to [listserv@cua.edu](mailto:listserv@cua.edu) with message: Subscribe ARN-L yourfirstname yourlastname (omit signature)]

Subscribe to: [EVALINFO](#) - General listserv of the American Evaluation Association. Circulates updated job bank information, AEA membership form, AEA meeting info., and a list of AEA SIG's.

[Send e-mail to: [listserv@BAMA.UA.EDU](mailto:listserv@BAMA.UA.EDU) with message: Subscribe EVALINFO yourfirstname yourlastname (omit signature)]

Subscribe to: [K12ASSESS-L](#) - The goal of K12ASSESS-L is to provide educators with a fast, convenient, and topical electronic discussion forum focusing on issues related to educational assessment in grades K-12. K12ASSESS-L is a place for local assessment personnel to share and obtain resources, ideas, and support. Visit the [K12ASSESS-L Home Page](#).

[Send e-mail to: [mailserv@lists.cua.edu](mailto:mailserv@lists.cua.edu) with message: Subscribe K12ASSESS-L yourfirstname yourlastname (omit signature)]

Subscribe to: [PSYCHOEDUCATIONAL ASSESS](#) - For those interested in psychoeducational assessment, especially special education related assessment. Most list participants are school

psychologists. [Send e-mail to: [LISTSERV@LISTSERV.ARIZONA.EDU](mailto:LISTSERV@LISTSERV.ARIZONA.EDU) with message: Subscribe PSYCHOEDUCATIONAL\_ASSESS yourfirstname yourlastname (omit signature)]

**LINKS TO TESTING-RELATED ASSOCIATIONS AND ORGANIZATIONS:**

[American Counseling Association \(ACA\)](#)

[American Educational Research Association \(AERA\)](#)

[American Speech-Language-Hearing Association \(ASHA\)](#)

[Association of Test Publishers](#)

[International Personnel Management Association \(IPMAAC\)](#)

[The Joint Committee on Standards for Educational Evaluation \(JCSEE\)](#)

[National Association of School Psychologists \(NASP\)](#)

[National Council on Measurement in Education \(NCME\)](#)

[Personnel Testing Council of Metropolitan Washington, DC \(PTC\)](#)

[Society for Industrial and Organizational Psychology \(SIOP\)](#)

[Society for Personality Assessment \(SPA\)](#)

**ADDITIONAL SELECTED ARTICLES AND READINGS ON ASSESSMENT:**

Atkin, J.M., Black, P., & Coffey, J. (2001). *Classroom Assessment and the National Science Education Standards*. Washington, DC: National Academy Press.

Angelo and Cross, 1993). *Classroom Assessment Techniques: A Handbook for College Teachers*

Black, P. (1998). *Education Assessment: Designing Assessments to Inform and Improve Student Performance*. San Francisco: Jossey-Bass.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Research Council.

Chappuis, J., (2005) Helping students understand assessment. *Educational Leadership*, 63(3), 39-43.

Chappuis, S., Stiggins, R.J., Arter, J., Chappuis, J. (2005). *Assessment for Learning: An Action Guide for School Leaders*. Assessment Training Institute, Portland, OR.

An article from the National Center for Fair & Open Testing Journal, *Fair Test Examiner* on the value of formative assessment: <http://www.fairtest.org/facts/FormativeAssessment.pdf>

Gardner, John (ed.) (2006). [Assessment and Learning](#). London, England: Sage Publications.

Gregory, K., Cameron, C., & Davies, A. (2000). Self-assessment and Goal-setting. Merville, British Columbia, Canada: Connections Publishing.

Herman, Aschbacher, and Winters, (1992). *A Practical Guide to Alternative Assessment* ()

Leahy, S., Lyon, C., Thompson, M., & Wiliam, D. (2005). Classroom assessment: Minute by minute, day by day. *Educational Leadership*, 63(3), 19-24.

Marzano, R. (1992). *A different kind of classroom: Teaching with dimensions of learning*. Alexandria, VA: Association for Supervision and Curriculum Development.

Marzano, R. J. (2006). *Classroom Assessment & Grading that Work*. Alexandria: VA: Association for Supervision and Curriculum Development.

Meisels, S., Atkins-Burnett, S., Xue, Y., Bickel, D. (2003). Creating a System of Accountability: The Impact of Instructional Assessment on Elementary Children's Achievement Scores, *Educational Policy Archives*, 11(9), pp. 19.

Rodriquez, M. (2004). [The Role of Classroom Assessment in Student Performance on TIMSS](#). *Applied Measurement in Education*, 17(1), 1-24.

Stiggins, R. (2004). New Assessment Beliefs for a New School Mission. *Phi Delta Kappan*. September: pp. 22-27

Stiggins, R., (2005). From formative assessment to assessment for learning: a path to success in standards-based schools. *Phi Delta Kappan*, 87(4), 324-328.

Stiggins, R., (2007). Assessment through the student's eye. *Educational Leadership*, 64(8), 22-26.

Stiggins, R., (2007) Five assessment myths and their consequences. *Education Week*, 27(8), 28-29.

Stiggins, R., & Chappuis, J., (2006) What a difference a word makes: assessment "for" learning rather than assessment "of" learning helps students succeed. *Journal of Staff Development*, 27(1), 10-14.

Tunstall, P. & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British Educational Research Journal*.

Wiggins, G. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Wiliam, D. (2006). [Does Assessment Hinder Learning? ETS Europe Breakfast Seminar: Professor Dylan Wiliam's Speech.](#)

Wiliam, D. The impact of educational research on mathematics education. In A. Bishop, M.A. Clements, C. Keitel, J. Kilpatrick & F.K.S. Leung (Eds.), *Second International Handbook of Mathematics Education* (pp. 469-488). Dordrecht, Netherlands: Kluwer Academic Publishers.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education*, 45 (4), p. 477-501.