Assessment for Identification of Reading Problems

Shutterstock

Most informed reading professionals agree that there is a place in the assessment regimen for both formal and informal measures.

From Chapter 3 of *Instructing Students Who Have Literacy Problems*, 6/e. Sandra McCormick. Jerry Zutell.

LEARNING FROM TEXT

Using Textual Features. In informational text, the author often uses certain textual features to highlight information. In this and many texts, words printed in **boldface type** signal that a term is defined right there in the passage for you. Understanding the meanings of technical terms is important when reading informational material. An often-used study strategy is that of underlining significant details; research shows this to be constructive. If you intend to keep this book, underline the boldfaced technical terms and their meanings or, if not, engage in some note taking related to these terms and definitions, paraphrasing the points in your own words.

Assessment is the total process of collecting information to make instructional decisions. Testing is one part of assessment.

Formal assessment uses standardized tests. A common type of standardized test is called **norm-referenced.** These are published tests for which norms based on the performances of large numbers of students have been developed. Norms allow comparisons of student performances with those of a typical group of the same grade or age.

Informal assessment can employ any number of nonstandardized measures, such as teacher-prepared tests; daily, ongoing observations; published informal inventories; checklists; interest inventories; interviews; and others. Currently many of these measures are assembled into a student portfolio, allowing assessment of change over time.

RTI

In most programs a variety of assessment types are used. As just one example, Response to Intervention (RTI) programs require teachers to employ four types: (a) screening, (b) progress monitoring, (c) diagnosis, and (d) outcome evaluation (Fuchs & Fuchs, 2008).

Examples of formal and informal assessment procedures are interspersed throughout all chapters in Part 2. The examples are placed to reflect the pattern and sequence that teachers in real school or clinic settings generally use. The organization in this book, therefore, provides both a scope and a sequence for diagnostic procedures. Part 2 has the following organization:

- This chapter presents information about assessment techniques that are used before students enter a special reading program. These procedures are employed to determine eligibility for placement in a Title I or other remedial reading class, an LD program, or a reading clinic.
- Chapter 4 discusses the first type of assessment usually conducted once students are enrolled in a program—assessment to determine or confirm reading level.
- Chapters 5 and 6 present a variety of tests that are often used next in the assessment process. These are employed to determine specific reading and

Informal assessment employs many types of nonstandardized measures, such as teacher-prepared tests and tasks, daily observation, published informal inventories, and others.

Anthony Magnacca/Merrill Education

writing strengths and weaknesses. Chapters 5 and 6 also discuss measures of interest and attitude so teachers can structure environments that facilitate learning.

SOME GENERAL ISSUES RELATED TO ASSESSMENT

Formal Testing versus Informal Testing

Periodically there are tensions between educators who advocate formal testing and those who prefer informal measures. Most authorities, however, desire a “reasonable and appropriate balance” between the two, a recognition of the weaknesses—and strengths—inherent in each type, and selection of those tests that are best in their category, whether that category be formal or informal. The most useful assessments of literacy, regardless of category, reflect our present understandings of reading and writing processes, resemble authentic literacy tasks, and reflect the complexity of literacy learning.

In addition, there must be an understanding of the specific purpose for which each category of test is best suited. For example, Figure 3–1 presents Farr’s (1992) description of assessment audiences—that is, what different groups or individuals legitimately need to know from tests and what types of tests best fit that aim.

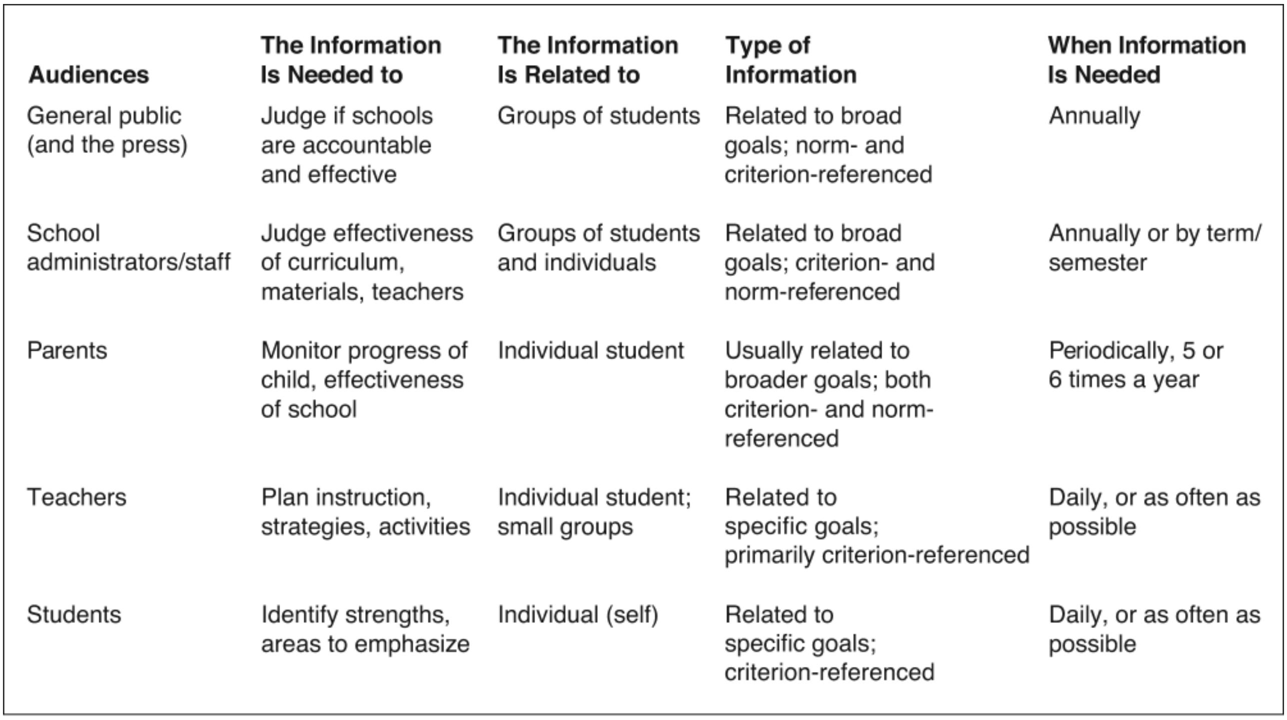
LEARNING FROM TEXT

Using Illustrative Aids. What is the main idea of Figure 3–1?

High-Stakes Testing versus Low-Stakes Testing

Although most informed individuals agree that there is a place in the assessment regimen for both formal and informal measures (see Figure 3–1), there remains some controversy about high-stakes testing versus low-stakes testing. State-mandated

FIGURE 3–1 ***Assessment Audiences***



Audiences	The Information Is Needed to	The Information Is Related to	Type of Information	When Information Is Needed
General public (and the press)	Judge if schools are accountable and effective	Groups of students	Related to broad goals; norm- and criterion-referenced	Annually
School administrators/staff	Judge effectiveness of curriculum, materials, teachers	Groups of students and individuals	Related to broad goals; criterion- and norm-referenced	Annually or by term/ semester
Parents	Monitor progress of child, effectiveness of school	Individual student	Usually related to broader goals; both criterion- and norm-referenced	Periodically, 5 or 6 times a year
Teachers	Plan instruction, strategies, activities	Individual student; small groups	Related to specific goals; primarily criterion-referenced	Daily, or as often as possible
Students	Identify strengths, areas to emphasize	Individual (self)	Related to specific goals; criterion-referenced	Daily, or as often as possible

Source: Figure from Farr, R. (1992, September). Putting it all together: Solving the reading assessment puzzle. *The Reading Teacher, 46*(1), 26–37. Reprinted with permission of the International Reading Association. www.reading.org

testing has been termed **high stakes** when serious "high-stake" consequences are tied to students' performances. These consequences can affect students as well as the educators who work with them and may be positive or negative.

In some states, school improvement plans have been linked to students' reading, writing, and math scores on state-developed tests. In states where legislative ruling ties grade-level assignment to test results, test performance can affect students' promotions or retention and even high school graduation. In certain districts, whether their students make a good showing or a poor showing on these tests may influence educators' salaries, or even a school's accreditation. As a result, in some circumstances, teachers and principals say they have felt undue pressures to teach not to children's needs but "to the test."

High-stakes testing emerged in its present form in the 1980s as a result of educational reform movements and grew by leaps and bounds. Begun with the good intent of improving schooling, policymakers and lawmakers called for standards, and assessments based on these standards, to be developed in order to spur accountability. Most U.S. states responded to that call, some employing commercially produced standardized tests for annual evaluations and others developing their own statewide assessments.

Some individuals have been satisfied with the results of high-stakes testing, with the media pointing out the rising test scores in certain states with strict accountability procedures. In other situations, these results have been contested, with educators contending that improved state test scores do not jibe with test scores of the same students on national standardized measures such as the National Assessment of Educational Progress (NAEP) (Hoffman, Assaf, & Paris, 2001).

A few states have used low-stakes testing to comply with compulsory state directives for yearly assessments. A hallmark of **low-stakes testing** is that it is *primarily* designed to *plan* for higher-quality instruction and is not used to reward or penalize learning or instruction after the fact. Low-stakes assessment often employs informal, rather than standardized, measures and school districts may select from several informal procedures (e.g., using a published, informal reading inventory [IRI] or having students read from a set of graded books).

Appropriate Interpretation of Test Scores

When using any assessment, teachers must realize that the scores provided are approximations. An important concept when using formal assessment procedures is that of **standard error of measurement.** This term refers to the principle that scores provided by tests are only estimations of an individual's "true" score and that a student's true score lies within a range of scores. Consider this hypothetical case: Jerry's computed score on a standardized reading test is 3.5. However, since the standard error of measurement for this particular test is 0.7, his "true" score could lie anywhere between 2.8 (seven points below 3.5) and 4.2 (seven points above 3.5). Test manuals report (or should report) the standard error of measurement for their test. Table 3–1 defines other terms commonly used in association with standardized test scores.

TABLE 3–1 ***Types of Scores Provided by Common Standardized Tests***

Raw Score
The number of questions a pupil has answered correctly on each subtest or on the total test. Raw scores mean little, but provide the basis for determining more helpful scores.

Grade Equlvalent
The score expected of the average student at the grade level designated. For example, a score of 4.2 indicates the student scored at the same level as the average student in the group used for norming who was in the second month of fourth grade.

Percentile Rank
The percentage of students in the norming group who had scores lower or higher than this student's score. A percentile rank of 55, for example, means that 55% of the group on which the test was normed scored lower. Percentile rank should not be used to determine growth.

Stanine
A statistical interpretation of percentile rank useful in examining an individual's score.

Normal Curve Equivalent (NCE)
A statistical interpretation of percentile score useful in examining group performance. Unlike percentile scores, these scores have been transformed into equal units of achievement. NCE is often used in Title I programs.

Extended Scale Score
Scores that can be used to follow a student's achievement over an extended period, even for several years. These scores are not provided by all standardized test manuals.

Remembering that test scores are estimates also is important when using informal measures. Assigning numerical scores to human abilities is not an exact science by any means. A score derived from an assessment instrument represents a good ballpark figure and is helpful because it gives us a place to begin when making instructional or placement decisions. However, such scores should never be interpreted as invariably definitive.

Especially in regard to informal assessment, MacGinitie (1993, pp. 556–558) highlighted several common biases that come into play in appraisals of human performance.

1. *Assimilation bias*—tendency to base judgments on early evidence and ignore evidence obtained later.
2. *Category bias*—tendency to assign all attributes ascribed to a category to a person we believe fits that category.
3. *Confirmation bias*—tendency to hold to beliefs, failing to look for other possibilities.
4. *Contrast bias*—exaggeration of differences between earlier and later findings.
5. *Negativity bias*—tendency to allow negative statements or information to take a disproportionate influence over positive.

Test scores should be interpreted in light of other available evidence, especially teacher observation.

In addition, sometimes scores obtained from a single test may simply be wrong. Teachers at times are heard to say something like, "Juan's standardized test score indicated he is reading at fourth-grade level, and I don't understand this because he is having no difficulty handling fifth-grade material." They seem reluctant to rely on their own observations if these do not agree with results of formal testing. Test scores should be interpreted in light of other available evidence, especially teacher observation.

Because of limitations of tests and other assessment procedures, teachers need to be tentative in their decisions. Margolis (2001) stated this point well when he reminds us that, in particular, "most reading textbooks recommend that group test scores be considered hypotheses to be validated through diagnostic teaching and observation" (p. 377). Decisions should be reappraised periodically, recognizing the biases inherent in both formal and informal evaluations. Furthermore, achievement should not be confused with ability; in remedial students in particular, the two often are not synonymous.

Reading assessment should be conducted in various settings and undertaken while students are reading for various purposes. Interpretations often are more accurate when this is done and frequently are in contrast to interpretations that rely merely on a single measure. It is especially important that students' behaviors be assessed while they are engaged in real reading in authentic texts and not just when they are taking tests.

ISSUES RELATED TO FORMAL ASSESSMENT

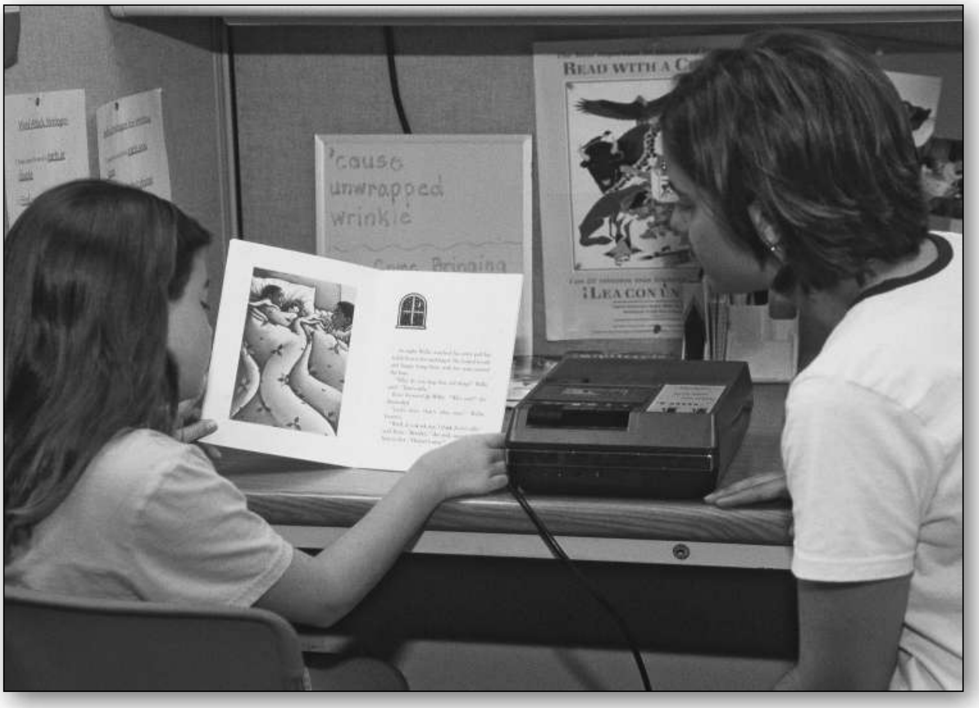
The first assessment task of a reading teacher is to identify those students who warrant remedial services. This is called **assessment for identification.** Since a good deal of assessment for identification involves use of formal measures, teachers should be aware of advantages and limitations of these tests and should be knowledgeable about their proper selection, administration, and interpretation.

Judging the Merits of Test Quality

Two categories of standardized tests may be administered to students with reading problems.

1. *Survey tests,* which are designed to determine students' general reading levels.
2. *Diagnostic tests,* which are used to analyze a student's specific strengths and weaknesses in reading strategies, knowledge, and skills.

Assessment should be carried out over time, in various settings and social contexts, and while students are reading for different purposes.



Anthony Magnacca/Merrill Education

Issues related to the technical acceptability of tests are of crucial importance.

Teachers need to consider many factors when they choose a standardized survey or diagnostic test. Issues related to the technical acceptability of tests are of crucial importance. The technical acceptability of a test is built on three factors: norms, validity, and reliability.

Norms. **Norms** are scores that represent an average and are used for comparing one student with other students. Test makers develop norms by administering their test to a large sample of individuals. To develop adequate norms, they must use a sample of students who are similar in age, IQ range, and general characteristics to the group with whom the published test is to be used. Most test publishers also try to select their sample from urban, suburban, and rural areas and, if they are attempting to develop **national norms**—that is, norms based on a nationwide sample—they select their sample from many regions of the country. (**Local norms,** based on data from certain schools or certain areas, are occasionally used, but most often school districts use national norms.) Based on performance of students in the sample, **grade norms,** that is, the average score of students from a given grade, are determined.

Test manuals should report the characteristics of the sample on which the test was normed so teachers can determine if the test is appropriate for their students. In addition, norms must be revised at least every 15 years to remain current; check the manual of the test you are considering to see when norms were last revised.

Validity. The **validity** of a test is the degree to which it measures what it claims to measure. **Content validity** is the extent to which a test assesses all aspects of the subject matter about which conclusions will be made. An example of a test sometimes used in reading assessment that lacks content validity is one consisting simply of a list of isolated words that students read orally. These tests purport to specify a student's instructional level based on this performance and claim to measure general reading ability. However, they obviously do not measure all factors involved in real reading.

Some other types of validity are **construct validity** (the degree to which performance on a test actually measures the extent to which an individual possesses a trait), **concurrent validity** (the degree to which performance on a test predicts performance on a criterion external to that test), and **predictive validity** (the extent to which a test predicts future performance in an area). Test manuals should report evidence of validity.

Reliability. The **reliability** of a test relates to the degree of consistency of its scores. In other words, if a student took the same test more than once, would he or she make approximately the same score every time? Or, is it likely that a score obtained on this test might just be a chance hit? In the latter case, administering such an unreliable test would be a waste of time because of the good possibility of its providing erroneous information.

Test makers can determine a reliability coefficient for a test by computing a coefficient of correlation between two alternate forms of the test or between scores obtained from repeated administration of the same test. Adequate reliability coefficients for a test used to compare groups should be above 0.60, but should be above 0.90 if used for diagnostic purposes with individual students (Salvia & Ysseldyke, 1982). Reliability coefficients should be reported in test manuals. If they are not, this often means the test developer has not checked the reliability of the instrument. Examine the manual of the test you are considering to see if reliability coefficients are reported—and adequate.

Buros, who edited the *Mental Measurements Yearbook* for 40 years, stated that one of his goals was "to make test users aware of the importance of being suspicious of all tests—even those produced by well-known authors and publishers—which are not accompanied by detailed data on their construction, validation, uses, and limitations" (Mitchell, 1985, p. xiv).

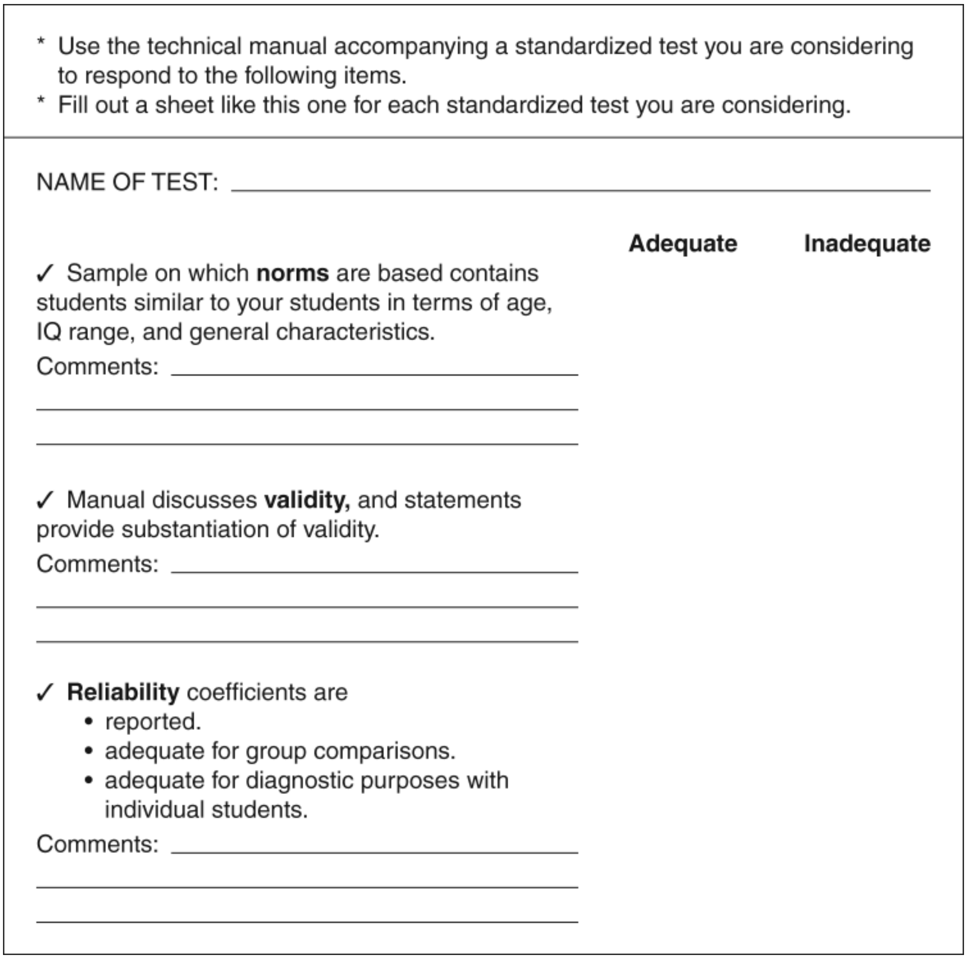
Figure 3–2 provides a checklist for evaluating tests *you* may be considering in order to help you determine if a test(s) is advisable to use.

Advantages of Standardized Tests

Generally, standardized tests save time since they can be administered to many students simultaneously. Group tests may also be used with individual students. In addition, if a standardized test has been properly devised, the test passages and questions have been checked out with numerous students. Some items are discarded in this process, and new items are tested until a final, suitable group of passages and questions is chosen. Most teachers do not have the time to prepare tests with such thoroughness. In addition, many test makers now monitor **passage dependency**—that is, they take care to ensure that a student must actually read a passage to answer the questions, rather than being able to answer based merely on previous knowledge. Finally, standardized tests are usually available in two or more equivalent forms so that students can be retested to measure growth.

Survey Tests. Group standardized survey tests can be employed to select students who require remedial programs by comparing their performances with the performances of others. In fact, administration of a standardized test is usually

FIGURE 3–2 ***Checklist for Judging the Technical Acceptability of Tests***



* Use the technical manual accompanying a standardized test you are considering to respond to the following items.
* Fill out a sheet like this one for each standardized test you are considering.

NAME OF TEST: ______________________________

Adequate **Inadequate**

✓ Sample on which **norms** are based contains students similar to your students in terms of age, IQ range, and general characteristics.

Comments: ______________________________

____

____

✓ Manual discusses **validity,** and statements provide substantiation of validity.

Comments: ______________________________

____

____

✓ **Reliability** coefficients are

- reported.
- adequate for group comparisons.
- adequate for diagnostic purposes with individual students.

Comments: ______________________________

____

____

required by federal, state, or local mandate for determining eligibility for most LD classes and U.S. Title I remedial reading programs. Figure 3–3 illustrates a typical example found on a group standardized survey test used to determine students' approximate reading levels.

Diagnostic Tests. Although *grade scores* on standardized diagnostic tests are not very reliable, if these grade-level scores are ignored and replaced with an analysis of student performance on specific reading tasks, some helpful diagnostic information may be obtained. With careful reflection about each error and its possible causes, a teacher may find these tests to be useful.

Disadvantages of Standardized Tests

There has been an increase in the use of standardized tests every decade since the 1950s. Although standardized tests provide helpful information if they are applied to the appropriate purpose, there also are disadvantages in using them.