<table>
<thead>
<tr>
<th>Chapter 3: Standardized Assessment</th>
</tr>
</thead>
</table>

- Standardized and Norm-Referenced Tests
- Test Construction
- Reliability
- Validity
- Questionnaires and Developmental Inventories
- Strengths of Standardized Tests
- Limitations of Standardized Tests
- Prudent Use of Standardized Tests
- Assessment in a Multicultural Society
Standardized testing of human behavior or skills has a long history in psychology. The term “mental tests” dates back to 1890, when the American psychologist James McKeen Cattell used it for the first time (Anastasi & Urbina, 1997). The testing movement is the result of a need to determine the intellectual, sensory, and behavioral (“personality”) characteristics in individuals. The educational need to classify children with limited intellectual abilities has been an important force in the development of psychological tests. The European experimental psychologists of the 19th century had discovered that in measuring sensory discrimination and reaction time of human participants, it is essential to standardize the conditions and instructions; such standardization was quickly incorporated into measuring any kind of human performance, including psychological tests. By the end of the 19th century, people attending scientific or industrial expositions were taking various tests that assessed their sensory and motor skills, the result of which were compared against norms (Anastasi & Urbina, 1997).

Formal and systematic measurement of intelligence, begun with the French psychologists Binet and Simon at the beginning of the 20th century, heralded the modern era of psychological testing. In subsequent years, tests to measure aptitude, personality, and educational achievement were developed. The need to assess various abilities of a large number of army recruits at the beginning of World War I in 1917 gave a significant boost to psychological testing.

Speech-language pathology has borrowed and accepted the tradition of assessing communication skills through standardized tests. Clinicians now commonly use them in diagnosing disorders of communication and qualifying children for intervention in public schools. In this chapter, we describe the essential features of standardized tests as a means of assessing communication skills and diagnosing communication disorders. We take note of the strengths and limitations of standardized tests. In the next chapter, we discuss assessment of children of ethnocultural minority groups for whom many available standardized tests are inappropriate. In Chapter 5, we describe alternatives to standardized assessment, along with an integrated and comprehensive approach that may be appropriate to all children.

**Standardized and Norm-Referenced Tests**

_Standardization_ is a research process that includes careful selection of test items, administration of the items to a representative sample drawn from a defined population, statistical analysis of results, establishment of age-based norms, and development of instructions and response scoring procedures. A standardized test should be self-contained in that all needed stimulus materials and recording forms should be included in it. The instrument should allow for uniform administration and scoring of responses by different examiners in varied settings to minimize any influence different examiners and settings may have on the skills being measured.

Most standardized tests also are _norm-referenced_. Norm-referenced tests are always standardized. Norm-referenced tests help compare a client’s performance to the performance of another group of individuals called the _normative group_. A _norm_ is the performance measure of a normative group on a tested skill. The normative group is usually a representative sample of individuals of the same age and possibly the same sex as the client.
A norm-referenced test should provide detailed information on the normative sample, including:

- The size of the sample; the American Psychological Association (1999) states that it should not be less than 100; meeting only this minimum requirement of a sample size, however, will result in a test of extremely limited applicability.
- The ethnocultural and socioeconomic levels of the individuals selected for standardization; the greater the variety of these variables and the more adequate the different subsamples selected, the higher the applicability of the test.
- The geographic distribution of selected participants for standardization; once again, the greater the geographic distribution of the samples, the higher the applicability of the test.
- Other relevant variables including the IQ or medical status of the sample; when children are the participants, the manual should also specify the range of education and occupation of the parents.
- Descriptive statistics; the manual should include means and standard deviations for all groups on whom the test items have been administered; additional statistical transformation of the raw scores (e.g., percentiles) also may be specified.

The test developers carefully select test items that they believe represent the skill the test seeks to measure. The test items are then administered to the normative group. The results are analyzed to establish the normal distribution of scores. This normal distribution is often represented by a bell-shaped curve with the range of scores or values measured on the horizontal access and the number of participants receiving a particular score on the vertical access. The bell-shape results because there will be a concentration of people in the middle (median or mean score) with a rapid decrease in the number of people as one moves away from the middle in either direction (Figure 3–1). In other words, most people’s scores will center around the mean, and progressively fewer people’s scores will deviate from the mean to the progressively greater extent. The peak of the bell-shaped curve will represent the mean or median score, the average performance on the test. This would also be the 50th percentile.

Because the mean represents a statistical average of individual performance, the score of an individual may not exactly correspond to the mean; it may deviate, or be different from, the mean. The extent to which an individual’s score deviates from the mean is called the standard deviation. In a normal curve:

- 68.26% of all scores will fall within 1 standard deviation of the mean (34.13% above and 34.13% below the mean)
- 95.44% of all scores will fall within 2 standard deviations of the mean (47.5% above and 47.5% below the mean)
- 99.72% of all scores will fall within 3 standard deviations of the mean (49.85% above and 49.85% below the mean)

A given individual’s score on a test may be understood in terms of standard deviation from the mean. The distribution within which a score falls is indicative of the extent to which the score deviates from the average score for the standardization sample.
The raw score of an individual is the initial scores given based on his or her correct responses to test items. Typically, the raw scores of different individuals on most standardized tests are not meaningfully compared with each other. Therefore, most standardized test scores are statistically transformed into other kinds of scores, collectively known as derived scores or scales. Such derived (transformed) scores of different individuals may be compared to determine their relative strengths and limitations. Percentiles, standard scores ($z$), and stanines are among the more frequently encountered transformed scores in standardized test manuals.

**Percentiles**

Percentiles, also called percentile scales, represent the percentage of individuals in the standardization sample scoring at or below a given raw score. The 50th percentile would represent the mean or median score. Therefore, those who score above the 50th percentile...
are “above average” and those who score below the 50th percentile are “below average” in the tested skill. Percentiles are a ranking system except that the lower the percentile, the lower the skill level and the higher the percentile, the higher the skill level.

A person whose score represents the 30th percentile is thought to have performed lower than 70% of those in the normative sample—a significantly below average performance. On the other hand, a person whose score represents the 95th percentile will have scored better than 94% of individuals on whom the test was standardized—a very high score because only 5% of the sample did better than that person. The highest raw score that no one else has obtained in the standardized sample would have a percentile rank of 100, although it does not mean a perfect raw score on the test. The lowest raw score of an individual that no one else has scored would have a percentile rank of zero, although once again, it does not mean a zero raw score.

Percentages of correct responses and percentiles refer to different measures. Percentage of correct responses is calculated on the basis of the number of correct and incorrect responses out of the total number of test items. The percentile, on the other hand, is based on the number of individuals who score below or above a particular raw score.

Percentile ranks are commonly used in comparing individuals on a skill being assessed. They are available for most standardized tests. A limitation of percentile ranks is that they can tell that two individuals are apart on a skill (one with a higher level than the other), but the scores cannot tell how much apart. Generally, scores around the mean are clustered and, hence, are close to each other. Scores that are at the extremes are fewer and also farther apart than those that are clustered around the mean. The distance between a score of 90 and 99, away from the mean, may be much bigger (larger differences between the individuals) than the distance between the scores 40, 50, and 60 (around the mean, with smaller differences in skills measured).

**Standard Scores (Z- and T-Scores)**

The standard scores represent the degree to which a child’s score deviates from the mean. Deviation from the mean is expressed in terms of a standard deviation (SD); it is a measure of the distance between the group mean and an individual score. There are two common types of standard scores: the Z score and the T score.

The Z scores are linearly derived standard scores. Z scores are based on the assumption that the mean is zero and the SD is 1. Z scores retain the original relationship between the raw scores. To calculate a Z score, the clinician needs to know the score of an individual \((X_i)\), the mean for the standardization group \((M)\), and the SD of the normative sample. The formula for calculating the Z score is as follows:

\[
z = \frac{X - M}{SD}
\]

If a child’s score is 70, the mean for the standardization sample is 60, and the SD for that sample is 5, then the child’s Z score is 2 (70 minus 60, divided by 5). This means that the child’s score fell 2 SD above the mean. If another child’s score is 60, but the mean is 70, and the SD is the same 5, then the child’s Z score will be −2 (60 minus 70, divided by 5). This means that the child’s score fell 2 SD below the mean. As these examples show, if a child’s score deviates to the negative side of the mean, then the Z score value will be...
negative; if it deviates to the positive side of the mean, the z score value will be positive. The z score for a child who scores the same as the sample mean will be zero (no deviation from the mean). Note that it is clinically much more meaningful to know that the child performed at, below, or above the mean for the standardization group than to say that the child scored 60 or 70 on a test.

The T scores are technically normalized standard scores because the distribution of scores of the standardization sample will have been transformed to fit a normal probability (bell-shaped) curve (Anastasi & Urbina, 1997). The T scores are based on the assumption that the mean is 50 and the SD is 10. Consequently, on a T score scale, a child who scores 60 is judged to be 1 SD above the mean and a child who scores 30 is 2 SD below the mean.

Most test manuals provide a table of z scores or T scores to readily interpret a child’s raw score on the test. Clinicians then can make diagnostic and intervention decisions based on the interpretation.

**Stanine**

Developed by the U.S. Air Force, a stanine is a statistical product of World War II. It is based on a nine-unit scale with a mean of 5 (representing the average performance) and an SD that approximates 2. The term stanine is created out of standard and nine. Because the scores range only from 1 to 9, all raw scores are converted to one of the single digits within that range. Percentages under the normal curve are used to assign stanine to a distribution of test scores. The same percentages are then assigned to each stanine, as shown in Table 3–1.

According to the table, the lowest scoring 4% of the individuals tested will get a stanine of 4 whereas the highest scoring 4% of individuals tested will get a stanine of 9. Stanine scores of 2, 3, and 4, respectively, represent 7, 12, and 17% of the tested individuals. Note that the mean is 5 with the largest percentage of individuals; the other percentages are roughly evenly distributed on either side of the mean. Scores 6, 7, 8, and 9 represent the corresponding percentages of individuals with above-average performance.

**Equivalent Scores**

Unlike the scales described so far which compare an individual’s scores to the normal distribution, equivalent scores suggest that a given score is representative of a particular age group or educational grade. For example, a 7-year-old could obtain a raw score of 70 on a standardized test. The test manual then may suggest that a raw score of 70 on the test was typically obtained by 6-year-old children in the standardization sample. This

| Table 3–1. Percentages Under the Normal Curve and Stanine Score Assignment |
|-----------------------------|--------|--------|--------|--------|--------|--------|--------|--------|
| Percentage                  | 4      | 7      | 12     | 17     | 20     | 17     | 12     | 7      | 4      |
| Stanine                     | 1      | 2      | 3      | 4      | 5      | 6      | 7      | 8      | 9      |
is often interpreted to mean that the 7-year-old child’s tested skills match that of a 6-year-old and that the tested child’s skills are more like those of a child who is 1 year younger. These types of age- or grade-equivalent scores are not standard psychometric measures, although they are popular in educational assessment. Parents and teachers alike find age- or grade-equivalent scores easy to understand; tested children may be classified as performing at, below, or above their age or grade levels. Many tests provide such equivalency scores and most clinicians, especially in educational settings, use them in reporting test results.

Clinicians, however, should be aware of serious limitations of equivalent scores. Such scores do not take into account variations in performance that might be common or expected in a particular age group. For example, we know that the 7-year-old child in the previous example scored similar to a typical 6-year-old, but we do not know how unusual this is. That score could be within the 40th percentile indicating that it is not that unusual for a child of that age, or it could be more than 2 standard deviations below the mean, indicating that it is unusual or below normal limits for that age. This example shows that age- or grade-equivalency scores are not accurate indicators of impairment (McCauley & Swisher, 1984a, 1984b). In fact, age- or grade-equivalent scores may be “the least useful or most dangerous scores to be obtained from standardized tests because they lead to gross misinterpretations of a client’s performance” (Haynes & Pindzola, 2004, p. 63). Clinicians should prefer one of the standard scores to equivalent scores.

Test manuals should provide instructions for interpreting raw scores using the various scales. It is important for the clinician to review this information prior to interpreting test results. In general, however, any score that falls in the bottom 5% of the normal distribution, or approximately −1.5 to −2.0 standard deviations below the mean is cause for concern (Haynes & Pindzola, 2004).

Basal and Ceiling Levels

Many standardized tests, particularly language and intelligence tests, require the examiner to establish a basal level and a ceiling level. Tests will define both the levels in terms of correct and incorrect scores. A basal score is the entry level and the ceiling score is the test-terminating score.

In many relatively long tests, testing may begin at some arbitrary level judged appropriate for the child, instead of at the very beginning of the test. This will help save test administration time. If the judged entry level proves wrong, the clinician may move down to find the basal level. In a given test, for example, a basal score might be defined as 5 consecutively correct responses. If at the initial level of testing, a child scored only 2 out of 5 responses correctly, the judged entry level would have been too high; the examiner then moves down the test items to find the first 5 consecutively correct responses. Once a basal is established, all test items prior to it are considered correct and the testing continues forward.

A ceiling in a given test might be defined as 6 out of 8 consecutively incorrect responses. In this example, the examiner continues to present the test items until the child gives 2 incorrect responses on 8 consecutive test items. The ceiling represents the highest number or level of test items administered or the point at which the test administration is stopped because the remaining higher level items are all considered failed. A test may require that a basal and ceiling be established to calculate a raw score.
Test Construction

Clinicians who select and administer standardized tests to assess and diagnose disorders of communication should have a basic knowledge of how they are constructed. Such knowledge is helpful in carefully evaluating available tests and selecting the most appropriate ones to assess a given set of skills.

Test developers generally consider the following principles in constructing and standardizing assessment tests. However, some test developers may be more successful than others in adhering to the principles of test construction.

**Test Purpose.** Tests are designed for different purposes, some limited and others more ambitious. Therefore, the experts designing a test must clearly state the purpose for which a test was created so that the practitioners do not misuse them. The authors should also present evidence that the test measures what it purports to measure. This is a matter of validity, which is addressed in a later section.

**Stimulus Items.** Stimulus items should be carefully selected to meet the purposes of the test. Simply put, stimulus items should help assess the targeted skills. These stimulus items may include verbal instructions, photographs, drawings, objects, and written information. These test items should also be assessed for their relevance, validity, and reliability. The selected photographs and drawings should be attractive, clear, and unambiguous. All stimuli should be culturally appropriate for the intended population of children.

**Test Administration and Scoring Procedures.** These procedures are predetermined and specified in the manual so that different administrators stimulate and score the responses in a uniform manner, thus minimizing the influence of individual examiners. To maximize consistency in test administration and scoring, written directions must outline exactly what the examiner does and says when providing the test instructions and administering the test. Specific rules for scoring client responses must be developed, along with accompanying resources (e.g., conversion tables for raw scores, percentiles, standard scores, stanines, means, and standard deviations).

**Normative Sample.** Norm-referenced tests are expected to assess skills that are representative of the children in the relevant population (e.g., children in the age range of 2 to 7 years or children belonging to certain ethnocultural groups). Because the test cannot be standardized on the entire relevant population—which is often very large—the test developers select a much smaller normative sample that is expected to represent that population.

Although the normative sample is smaller than the population it represents, it cannot be too small to be biased. Test developers select a sample of participants they believe represents the relevant population. According to the Standards for Educational and Psychological Tests compiled by the American Psychological Association (1999), the number of subjects in each group should not be less than 100. For most nationally or even regionally used tests, one probably needs a much larger sample than 100 participants. The sample should
adequately represent the various educational, socioeconomic, and occupational levels of the families from which the participants are sampled. Proportionate sampling of children who belong to different racial and ethnic groups and those who live in different geographic regions (e.g., rural, urban, and suburban) is also important. Standardized tests that are considered inadequate often will not have sampled the participants adequately to achieve the stated test purpose.

**Statistical Analysis.** The test developers should provide the range of raw scores as well as derived scores. Various statistical analyses of the normative data may be made to give derived scores of the kind described previously. The analysis should include means and standard deviations for all groups in the normative sample. Confidence intervals and information on the standard error of measurement should also be included.

**Validity and Reliability.** It is important to establish that the test measures what it purports to measures and that it does it consistently. These are matters of validity and reliability, respectively. The test manual should include information on the kinds of validity the authors established. Minimally, the manual should report construct validity. Test-retest, interjudge, and internal consistency types of reliability should also be reported (Hutchinson, 1996). Other kinds of reliability and validity, if reported, will enhance the usefulness of a test. Because of their technical nature and importance to test selection, the concepts and procedures of reliability and validity are discussed in some detail in the next section.

**Test Manual.** A comprehensive test manual is essential to properly administer a test and to make correct interpretation of the results. Hutchinson (1996) states that the manual should include detailed information on the following: the rationale for the test; a description of how the test was developed; the purpose of the test; examiner qualifications; detailed instructions for the administration, scoring, and interpretation of the test; reliability and validity indexes; information on the normative sample; and descriptive statistics. The manual should be written in a clear and concise language while giving all essential information. Informative tables and attractive graphs will help understand the procedures of analysis and facilitate interpretation of results.

## Reliability

**Reliability** is consistency across repeated measures of the same phenomenon with the same instrument. In other words, test reliability means that the results obtained for the same individuals are replicable and that different examiners report similar scores for the same individuals. Consistency of measures is critical for both standardized and naturalistic measures. A test will not be trustworthy if it cannot provide consistent results when administered to the same individuals by different examiners or when the same examiners administer it to the same individuals on different occasions.

Typically, reliability, much like its twin concept of validity to be described in the next section, is expressed as a correlation coefficient (Hegde, 2001, 2003; Salvia & Ysseldyke,
1981). Standardized tests often report such reliability correlation coefficients. It is preferable for tests to have correlation coefficients of .90 or above. The closer a coefficient of reliability is to 1.0, the more reliable the test.

Reliable measurement of communicative behaviors is a difficult and often neglected issue in speech-language pathology. The difficulty is common to both standardized tests and naturalistic observations (e.g., speech and language samples). It is not always evident that clinicians can reliably measure the frequency of stuttering or make reliable judgments about various voice qualities.

Some factors facilitate acceptable reliability of measures whereas others impede it. To achieve improved reliability, it is necessary to:

- Define the behavior being assessed in operational (measurable) terms; poorly defined or broadly conceptualized skills are hard to measure in the first place, let alone reliably.
- Describe the administration and scoring procedures in clear and precise language; standardized tests that do not give clear instructions on test administration and response scoring cannot lead to good reliability of scores; similarly, clear and precise description of speech-language sampling procedures and scoring methods will facilitate reliable measures.
- Train test administrators or observers of naturally occurring behaviors (as in obtaining a speech-language sample or judging voice deviations); poorly trained test administrators or observers cannot be expected to get consistent measures on the same individuals even with the same tests or methods of observation.
- Simplify the behavior to be assessed or observed; although oversimplification of a complex skill will not serve any purpose, the more complex or abstract the behavior being assessed, the more difficult it is to achieve reliable measures; breaking a complex behavior into more easily observable components might be helpful.
- Record the observational or testing session; although not commonly done, beginning clinicians may audio-record a testing session until they become proficient in test administration; naturalistic observations (e.g., language sampling) is typically audio-recorded for later analysis.
- Take note of the natural variability of the behavior or skill being measured; some behaviors, no matter how well the observers are trained and how carefully the measurement procedures are developed, may fluctuate across time and situations; stuttering is known to vary in this manner; some voice characteristics also may naturally vary; the only available solution to this natural variability of a phenomenon being measured is to control the conditions under which it is observed and repeat observations until the behavior stabilizes.

A concept central to reliability is error of measurement (Hegde, 2003). When a phenomenon is measured repeatedly, it is unlikely that the measured values will be identical, even if the phenomenon is stable; very few phenomena, if any, are stable over time. Furthermore, some behaviors are more stable than others. Speech-language production and comprehension skills may be relatively more stable than frequency of stuttering or the “mood” of a person. Consequently, some fluctuations in scores across repeated mea-
measurements is unavoidable because the behavior being measured may change over time, combined with errors in measurement. Therefore, variability in repeated scores obtained for the same individuals may have two sources: a true difference (variability) in the behavior being measured, and an error variance which is due to common and unavoidable errors in measurement. A reliable test minimizes the error variance.

Reliability is established in different ways, giving rise to different kinds of reliability. Each type captures a different aspect of reliability, and addresses different kinds of measurement concerns.

**Interobserver Reliability**

*Interobserver reliability*, also known as *interjudge reliability*, is consistency of test scores recorded by two or more examiners administering the same test to the same individuals. In the context of language samples or other naturalistic (nonstandardized) measures, interobserver reliability refers to consistency of measured values across observers who measure the same phenomenon in the same individuals. Whether it is a standardized or more naturalistic measure, reliability is a basic requirement that all observations and measurements must fulfill.

Interobserver reliability is higher if two or more observers are trained strictly according to the manual of a test or trained the same way to make naturalistic observations. But the critical issue here is whether the test itself allows for acceptable interobserver reliability. If two well-trained test administrators cannot get similar scores on the same individual, for the same and relatively stable behavior, then perhaps there is a reliability problem with the test itself.

**Intraobserver or Test-Retest Reliability**

Also known as *intrajudge reliability*, *intraobserver reliability* is the consistency of scores the same individuals obtain when the same examiner readministers the test or repeats a naturalistic observation. In the context of standardized tests, intraobserver reliability is typically described as *test-retest reliability* because the main method of establishing this type of reliability is to repeat the test on the same individuals. In the context of nonstandardized naturalistic measurement of skills (e.g., speech-language sampling, base rating, probing during or after treatment), the term *intraobserver reliability* is more appropriate than *test-retest* reliability. Obviously, if repeated administrations of a test or naturalistic observations of a behavior yield divergent scores, the clinician cannot make diagnostic or treatment decisions.

The test-retest method is probably one of the easiest methods of establishing reliability of standardized tests. The examiner then correlates the two sets of scores of the same persons. An obvious question for the test constructor is the optimum time duration between the two administrations of the same test to the same individuals. If the duration is too short, the individuals tested may simply recall their responses from the first administration; this will inflate the reliability correlation coefficient. If the duration between the two administrations is too long, the behavior being measured may change, thus yielding a low reliability correlation coefficient. Generally, the same individuals are retested within a month, perhaps in two weeks. Tests standardized for children usually have shorter interval durations than those developed for adults because of more rapid changes.
in skills in the former, as against the latter, populations. Even for adults, however, the interval should not exceed 6 months unless there is some special reason to extend it (Anastasi & Urbina, 1997).

Another question that needs to be answered about the usefulness of retest reliability is whether the skill being measured will change—often improve—because of the testing itself. Motor dexterity test scores may improve because of practice the individuals gain during the test. Articulation or morphologic production may improve with repeated practice. The problem, however, is more acute in measuring opinions and attitudes than relatively stable speech-language skills. We return to this problem in a later section on inventories and questionnaires.

To facilitate an understanding and evaluation of a test, its manual should specify not only the test-retest correlation coefficient, but also the interval between the two administrations. The manual should further specify any changes that may have occurred in the life of the individuals who were retested (e.g., illness, counseling, or treatment during the interval). If the retest reliability of a test is good, but an individual clinician does not get consistent scores on the same individuals, it is possible that the clinician needs training in the administration and scoring of the test.

**Alternate- or Parallel-Form Reliability**

*Alternate-form reliability* refers to the consistency of measures when two forms of the same test are administered to the same person or group of people. Obviously, it requires two versions of the same test (Hegde, 2001). This type of reliability minimizes some of the problems inherent to retest reliability. For instance, alternate forms of the same test may reduce the chances of recall of responses from the previous administration because the test items that sample the same behaviors may be different on the two forms. When there is an appropriate temporal interval between the two administrations, alternate form reliability indicates both the degree of temporal stability of responses and the consistency of test items.

The main difficulty in establishing parallel-form reliability is the construction of two versions of the same test that include items comparable in difficulty. If one form is easier than the other, or if one form contains items more familiar to certain individuals than to others, then the scores on the two forms may not reflect its true reliability. If the two forms are administered about the same time in quick succession, the resulting acceptable reliability coefficient will only indicate consistency across forms, not temporal stability. Such a coefficient may also partly reflect some degree of practice effect. The test manual should, therefore, specify the time interval between the two parallel-form administrations. Furthermore, because of the need to construct two parallel forms of the same test, a two-fold effort is needed to standardize a test and establish this type of reliability.

**Split-Half Reliability**

*Split-half reliability* is a measure of internal consistency of a test. To establish this type of reliability, an examiner correlates the scores from one-half of the test with those from the other half of the test. Another method of calculating split-half reliability is to correlate responses given to odd-numbered items with those given to even-numbered items. The correlation calculated with the split-half method is sometimes described as coefficient of
internal consistency. Such a coefficient does not suggest stability of scores across time because there is only a single administration of the test.

To establish split-half reliability, both halves of a test should be equivalent in the content and difficulty level; the entire test should measure the same skill. Split-half reliability is desired for some tests, but may be inappropriate for a test whose items are in a hierarchical order. For example, some tests begin with easier items and progress toward increasingly difficult items. Splitting such a test into equivalent halves will yield low reliability coefficients. A solution to this problem is to correlate the responses on odd-even items; this method samples items throughout the test to create two halves. Therefore, responses given to both the easier and more difficult items are sampled for correlation.

A test that reliably measures something may not measure something specific. In other words, what a test measures reliably may not be what the examiner wants to measure in given individuals. Therefore, having decided that a test reliably measures something, the clinician needs to evaluate whether what is measured is what needs to be measured. This is the matter of validity.

Validity

Validity is the extent to which a test measures what it is constructed to measure. It reflects the truthfulness of the measure. For example, if a test claims to measure a child’s expressive vocabulary, it should adequately measure that skill. If it measures something else, it is invalid for the purpose of measuring expressive vocabulary, although it may measure some other skill quite validly (as well as reliably). Therefore, note that validity of a test is always judged against what it is purported to measure.

Validity of a test is also judged against an examiner’s intended use. Many tests may be used for different purposes, some not suggested by the original test developers. For example, a test of speech sound production in longer sentences may later be used to assess the maximum speed with which syllables are produced without losing intelligibility. In such contexts, it is the test user who should judge whether the test is valid for the new purpose for which it will be administered. In a general sense, validity is “the extent to which interpretations of the test scores are meaningful and acceptable” (Hegde & Maul, 2006, p. 96). Communication is a complex process that involves many kinds of skills, so it is important that the test actually measure what the test developer intended it to measure and what an examiner uses it to measure.

There are different kinds of validity, but the kinds are simply a result of different ways in which a test’s validity can be established. Some types of validity are based on expert’s careful judgment whereas other kinds require statistical analyses of additional data gathered to demonstrate a new test’s validity. It is a common misconception that a test is fully valid if it contains any one of the several types. A truly valid test, however, should demonstrate all major kinds of validity, a description of which follows.

Content Validity

Content validity is demonstrated by expert judgment that the test indeed includes items that are relevant to assessing the purported skill. To validate a test through this method,
one should have a thorough knowledge of the skill being measured. If the experts find that a test of English morphologic features contains items to sample the production of English morphologic features, then the test is validated from the standpoint of its content.

Three major criteria guide the content-validation of a test under construction (Salvia & Ysseldyke, 1981):

- ** Appropriateness of the items included:** The test should contain only those items that help sample the purported skill and it should not contain items that are judged irrelevant to the skill being measured; most carefully designed tests meet this criterion.

- **Completeness of the items sampled:** The number of items included in the test should be sufficient to make a comprehensive assessment of the skill being measured; unfortunately, many standardized tests fail to meet this criterion because they assess a skill with a single or few items.

- **Way in which the test items assess the content:** The manner of assessment should fit the skill; for example, speech sounds need to be evoked in words and in three word-positions (initial, medial, and final); morphologic features should be evoked with specific kinds of questions, whereas certain conversational skills (e.g., topic maintenance) should be assessed through narrative speech.

Content-validation of a test requires that the entire skill range to be tested is described beforehand, not after the fact. The test items should be appropriately balanced across subskills; they should not be overloaded on skills for which it is easier to write objective test items. Content validity is not achieved by a superficial examination of resemblance of the test items to the skill being measured; such an examination may suggest what is known as *face validity*, which is unsatisfactory. Content validation requires an expert’s critical examination of each item’s relevance and ability to sample the behavior under observation.

Mere content validity, though reported to be satisfactory, is not a sufficient justification to use a test. The test must have other kinds of validity as well, although all other kinds begin with content validity. A test that includes irrelevant items, an inadequate number of relevant items, or empirically invalid evoking procedures for assessing a skill is unlikely to have other kinds of validity.

**Construct Validity**

*Construct validity* is the degree to which a test measures a predetermined theoretical construct. A *construct* is a theoretical statement supported by prior empirical observations or experimental data on a behavior or skill being measured. In essence, a test, to have construct validity, should produce test scores that reflect a theoretical statement known to be valid about the behavior being measured. For example, observational studies on language acquisition in children have supported the theoretical statement that expressive vocabulary improves with age. Therefore, a test of expressive vocabulary in children should yield scores that show progressive improvement with age. Similarly, tests of speech sound production should reflect positive changes across the age levels, a theoretical
construct supported by empirical evidence. Another construct a speech or language test may use in demonstrating its validity is the difference in scores of children within the normal range of intelligence and those with significantly below-normal IQs.

Construct validity is expressed in terms of a correlation coefficient: according to the examples given, age levels and the test scores should be positively correlated. Most speech and language tests have the potential to meet the requirements of construct validity because of known developmental changes in communication skills across progressively higher age groups. A test that does not reflect such changes may be judged invalid.

Construct validity may be demonstrated not only by positive correlation with some variables, but also by low or negative correlation with other variables. For instance, a test of oral reading skills is expected to correlate poorly with arithmetic reasoning skills because they are theoretically unrelated skills. The validity of a test of oral reading skills that correlates highly with a test of arithmetic reasoning skills may be highly suspect. When low or negative correlation with other variables is used as the theoretical construct, the resulting validity is sometimes described as discriminant validation, still a variety of construct validation (Anastasi & Urbina, 1997).

Another kind of construct validity, especially relevant for clinical disciplines, is based on experimental findings (Anastasi & Urbina, 1997). For instance, if prior experimental research has confirmed that a particular treatment procedure effects significant changes in certain language skills, a relevant language test given before and after that treatment should yield different scores. If the pre- and posttest scores of that language tests do not show improvement in children who have received that treatment, then perhaps the test is not valid. In such instances, the clinician, of course, will have measured language skills through other means (e.g., language samples or systematic probes) and will have documented improvement in those skills.

Construct validity, especially when based on well-established empirical and experimental findings, can be a strong indicator that the new test measures what it purports to measure. When combined with good content validity, acceptable construct validity may inspire confidence in the use of a standardized test.

Construct validity is as good as the theoretical construct on which it is based. Clinicians cannot always assume that the constructs themselves have strong empirical support. Even the widely held assumption that language skills improve across age levels may be more true for younger children than for older students and adults. Some generally believed theoretical constructs may be wrong. Moreover, just because the test scores correlate well with a theoretical construct does not mean that it is valid. For instance, a child language test may correlate well with height in children (both increase), but the relation is spurious. The theoretical construct and the skill being measured should be meaningfully related.

**Criterion Validity**

Criterion validity is the assurance that the test measures what it purports to measure because it is correlated with another meaningful variable. Unlike both content and construct validity that are inherent to the test itself, criterion validity requires some external variable to validate a new test.
Several types of external variables may serve as criteria for establishing the validity of a new test. We consider the two most common forms relevant for speech-language tests: concurrent validity and predictive validity.

**Concurrent Validity.** Concurrent validity refers to the degree to which a new test correlates with an established test that is already regarded as valid for measuring the same skill or behavior. For example, the scores on a new test of articulation may be correlated with those of an established test whose acceptable validity has been demonstrated. The new test is judged valid if the scores on it positively correlate with those of the older, reliable test designed to assess the same skill. On a given individual, the new and the old test should lead to similar decisions: the same or similar skill level or the same or similar deficiency level, suggesting the same diagnosis.

It is important to remember that the two tests should lead to same decisions on a given individual, but the two test scores should not correlate too high. Only a moderate correlation between the scores of the two tests is desirable; too high a correlation means the two tests are very similar. Perhaps the new test is an unnecessary duplication. A moderate correlation, although ensuring concurrent validity, will suggest that the new test helps assess aspects of the same skill untouched by the older test.

**Predictive Validity.** The degree to which a test predicts future performance on a related task is called predictive validity. Unlike content, construct, and concurrent validity, predictive validity takes time to establish; it may not be reported in the first edition of the test. As the test gains acceptance, the scores made by individuals may be checked against some other criterion of performance. For example, the scores of children on a language competence test, accumulated over a period of time, may be correlated with grade-point average or other indexes of academic performance (Hegde, 2001). If children who scored relatively high on the language test also scored relatively high grade-point averages (and vice versa), then the test accurately predicts the future academic performance, which is the criterion predicted.

Several other criteria may be used in establishing predictive validity of new tests. A single test may be validated against multiple criteria—provided that the test is designed for multiple uses. Each use of a test may be validated against its own predictive criterion. For example, a phonological skills test may be validated against (correlated with) spelling skills or oral reading skills—two related but distinctly different domains—measured years later. Scores on a morphologic language test designed for preschool children may be correlated in the future when the same children’s scores on writing skills become available. Similarly, the scores on a test of oral narrative skills may later be correlated with written language skills.

Most college students know that the admission officers believe their admission tests (including the Graduate Record Examination—GRE) are valid because the scores predict performance in future college courses. A real estate salesmanship test may be eventually validated against the number of homes the agents sell.

Like other kinds of validity, the criterion validity has its limitations. As noted, it may not be available at the time a test is first published. Unless other kinds of acceptable
validity are reported for it, an invalid test may become entrenched before it is shown to be incapable of predicting relevant future performance. In some cases, there may be what is known as criterion contamination (Anastasi & Urbina, 1997). For instance, classroom teachers who know their students’ scores on a language test may give better grades to those who scored better on the (language) test, and lower grades to those who scored lower on the test. The resulting high criterion validity for the language test may be specious.

As Anastasi and Urbina (1997) point out, different types of validities may be interrelated, in that the construct validity is a more inclusive term. For instance, predictive validity—that the scores of a test accurately predict some future performance—may be considered a theoretical construct the test meets. Similarly, the assumption that one valid test should correlate moderately but not too highly with another valid test, also may be considered a theoretical construct similar to the well-recognized theoretical construct that speech-language skills in children increase with chronologic age. It is also evident from the previous discussion that validity of a test may evolve over time. Evidence of predictive validity of a test is especially likely to accumulate in the years after its initial publication.

A broader understanding of validity and reliability of tests and limitations of different kinds of validation and reliability procedures will help clinicians select the most suitable tests in assessing their clients. Clinicians also need to appreciate the basic strengths and limitations of all standardized tests, including those that report acceptable reliability and validity. We address these issues in a later section in this chapter; we describe commonly used standardized tests in Chapters 6 (speech production tests) and 8 (language tests).

**Questionnaires and Developmental Inventories**

In addition to standardized tests, speech-language pathologists often rely on questionnaires and developmental inventories to seek information about the children’s speech and language development from their parents or caregivers. This information may be particularly helpful when trying to evaluate young or difficult-to-test children, or as a supplement to other information obtained during the assessment. Parents may be asked to respond to questionnaires and inventories that itemize and organizes various behaviors into a developmental or observational list.

There are several advantages to using inventories or developmental questionnaires to obtain information from parents regarding their child’s speech, language, and general development. For instance, parent-reported information may:

1. Provide a way to tap into a parents’ unique knowledge of how their child communicates on a day-to-day basis in a variety of settings and situations.
2. Offer information on changes in the child’s skill over an extended period of time; this kind of information can be obtained only through parental input.
3. Give insight into the parental actions and reactions to any limitations or deficiencies they note in their child’s speech and language skills.
4. Help assess young children’s (under age 3) communication skills without the use of costly and structured tests that require time, training, and child cooperation.
5. Represent an infant’s or toddler’s communication skills better than a language sample collected in a clinical setting; the results of the former source may be less influenced by factors such as setting, content, and examiner unfamiliarity.

6. Assist in planning for the assessment session because it can be obtained prior to actually seeing the child; such information may be helpful in selecting child-specific tools and procedures needed to complete the evaluation.

7. Offer an adjunct to supplement the clinician’s observations.

8. Help establish rapport and create a “therapeutic alliance” between the parents and the clinician because of the clinician’s request for taking part in assessment.

9. Eliminate the need for expensive stimulus items.

10. Afford a cost-effective and time-efficient means for a general evaluation or screening of a child’s language.

11. Provide a valid and efficient means of assessing child's naturalistic communication skills.

12. Provide a means of checking the reliability of information obtained in the clinic.

The reliability and validity of parent/caregiver-reported information is a concern to professionals. Parents or caregivers may not provide completely accurate information. Some parents or caregivers may miss subtle language behaviors that a trained clinician would note and others might overestimate their child’s speech and language skills. Generally, any recalled information is subject to memory problems. Also, the older the child on whom the parents recall early developmental information, the greater the chances of memory distortions. Despite these concerns, several studies have shown that parents and primary caregivers can be excellent informants and that there is a strong relationship between parent-reported information and clinician-measured data (Boudreau, 2005; Conti-Ramsden, Simkin, & Pickles, 2006; Dale, 1991; Dale, Bates, Reznick, & Morisset, 1989; Fenson et al., 1993; Meisels, 1991; Miller, Manhal, & Lee, 1991; Squires & Bricker, 1991).

**Strengths of Standardized Tests**

Standardized tests are commonly used mostly for historic, regulatory, and practical reasons. A variety of historical trends, actual strengths, educational policies, and commonly offered arguments justify the use of standardized tests:

- The history of standardized assessment in the United States is longer than that of alternative approaches; many standardized tests are readily available.

- Standardized tests have been historically promoted as “objective” in the sense that the examiner’s biases would not influence the results.

- Clinicians are routinely and traditionally trained in administering standardized tests due to the historic belief that standardized assessment is better because they are more formal and objective than other kinds of assessment, which are often dubbed as “informal,” implying “less objective.”
• Many clinicians tend to receive little or no systematic training in using alternative procedures; therefore, they continue to use what they are trained in.

• Some school districts may have a policy of qualifying children for services based only on the results of standardized tests, although the federal special education laws do not require that the decisions be based on the results of standardized tests.

• The concept that deviation from a statistical norm suggests a disorder in need of remediation is entrenched in the U.S. educational and clinical establishment. Most people seem to more easily grasp that concept than such alternative views as skill deficiency or mastery levels that do not match the social and academic demands the child faces.

• The tests are convenient and easier to administer. Once learned to administer, standardized test administration is more routine and easier than designing a child- and family-specific speech-language assessment procedure.

• It is much easier to analyze and interpret the results of standardized tests than the results of alternative assessment procedures.

• Standardized tests are conveniently packaged. All stimulus materials, acceptable and unacceptable response characteristics, response scoring procedures, and interpretations are ready-made for the clinician. Most alternative procedures require extensive preparation, often special preparation for individual clients.

• Clinician believe that standardized tests are efficient even if they inadequately sample communicative behaviors. A variety of speech and language skills may be sampled in a relatively short duration, regardless of the problem that each behavior is only sampled superficially.

• Tests permit comparison of results across various test administrations and examiners. More flexible alternative assessment results do not easily lend themselves to such comparisons.

• In spite of a number of alternative approaches available to assess communication, none have replaced the standardized tests.

• Clinicians may believe that many standardized tests, although inappropriate to assess children who belong to ethnocultural minority groups, are quite appropriate to assess mainstream children on whom the tests may have been standardized. The strong possibility that the tests may provide only limited information even on children who belong to the standardization group is often unappreciated.

• Some who favor standardized testing believe that tests are indeed more capable of avoiding bias in assessing all people, including minority cultural groups, than clinicians who make judgments without objective test scores. Used appropriately, test results may “provide a safeguard against favoritism and arbitrary and capricious decisions” (Anastasi & Urbina, 1997, p. 549). This may be especially true when the tests are appropriately standardized on specific target groups (various cultural minorities) with test items that are unique to the specific groups.

• At least certain kinds of biases are not a property of the test itself, but a matter of faulty interpretations of results. Tests do not say why a child behaved in a
certain manner during testing; the clinician will have to find out the answer. For instance, a language test might reveal that a child belonging to a certain cultural minority does not maintain eye contact during conversation. The test cannot be faulted for discovering this fact. No test makes inappropriate cultural or clinical decisions because “tests couldn’t see whether the youngster is in rags or tweeds, and they couldn’t hear the accents of the slum” (Gardner, 1961, 48–49). If anything goes wrong with the finding in our example, it is the clinician’s interpretation that the child who did not maintain eye contact has a pragmatic language disorder.

- An additional argument in favor of prudent use of standardized tests in assessing cultural minorities is that it is indeed both fair and essential to know what negative effects unfavorable social conditions have on the skills of people so that appropriate remedial actions can be taken. Using procedures that are insensitive to the negative consequences of unfavorable socioeconomic conditions does a disservice to people who experience such consequences (Anastasi & Urbina, 1997).

- Test constructors now follow widely accepted ethical guidelines in standardizing tests. Authors of tests have been making and continue to make systematic efforts to eliminate biased stimuli, culturally loaded instructions, stereotypic characterization of diverse people, and so forth. Test developers are making significant efforts to sample ethnoculturally diverse populations.

- Standardized tests, like any kind of assessment instruments, may be misused. It is the clinician’s responsibility to use them properly. This argument correctly puts the responsibility on those who train clinicians in assessment procedures and their fair and objective interpretations that lead to beneficial consequences to those who take tests.

Reasons and arguments such as these have made the administration of standardized tests the main task of assessment in speech-language pathology, psychology, and education. However, a prudent use of standardized tests requires that the clinicians understand their limitations as well.

**Limitations of Standardized Tests**

Limitations of standardized tests and objections to their widespread use began to surface in late 1950s and early 1960s. Various legal challenges to the fairness of intelligence testing and passage of civil rights laws in the 1960s forced attention to the limitations or inappropriateness of standardized tests in assessing people from cultural minority groups. Critics contend that some practical advantages of standardized tests have over-ridden their serious and multiple disadvantages, leading to their widespread use. Indeed, even the strongest advocates of standardized testing are fully aware of its limitations and offer thoughtful discussions (e.g., see Anastasi & Urbina, 1997, a classic, still a current, and highly respected, book in psychological testing). Over the decades, the validity of standardized tests as means of assessing communication and its disorders also became a matter of serious concern. Due to extensive discussions in psychological, educational, and communication disorders literature, the limitations of standardized tests are now

- The highly structured and standardized formality of most standardized tests do not allow for a sampling of natural social interactions inherent to communication.
- Most standardized tests sample behaviors may not represent the client’s functional communication skills in social situations.
- Most standardized tests provide limited opportunity for the child to initiate interaction. Test items evoke responses, but do not allow for much spontaneous communication.
- Many tests artificially isolate individual aspects of communication (e.g., expressive syntactic skills or receptive vocabulary). Test items do not help assess language skills that are integrative of different structural aspects.
- Speech and language tests generally are less efficient in assessing more global conversational skills and abstract language production than they are in assessing the production of discrete words and grammatic morphemes.
- Most standardized tests sample verbal language and overlook preverbal or nonverbal communication.
- The standardized formats and protocols allow for little individual variation in communication across children.
- Most standardized tests limit the role of family members in the assessment process and intervention planning.
- Few standardized tests of early language skills have good predictability of language and communication skills expressed at a later age.
- Norm-referenced standardized tests of speech and language are based on the assumption that all children learn communication skills in a certain invariable sequence. This assumption, however, is highly questionable and individual differences in learning language are well-documented.
- Comparing a child’s performance with the norms is one of the most troublesome, even if cherished, aspects of standardized tests. A child that does not demonstrate a particular speech or language skill by a certain age may be labeled disordered; this may be valid only if individual differences in language acquisition were insignificant.
- Standardized tests do not typically include children with language disorders or those with other clinical conditions that may affect language skills in their normative sample. Therefore, the norms do not accurately reflect the general population, which includes children with limited language skills.
- The interaction style of many young children may not match the type of interaction required on a given test. In fact, there is little interaction except for giving responses to preselected test items.
- Standardized tests do not sample behaviors adequately. Although most test constructors pay attention to the sample size (the number of children tested),
they ignore the problem of inadequate sampling of skills (Hegde, 1998). Most tests contain only one or a few stimulus items for any particular skill; the results may not be an accurate representation of the child’s behavior in social settings.

- Standardized tests are designed only to determine whether or not a problem exists compared with a sample of the general population. Therefore, they may not identify a child's overall strengths and weaknesses.

- Standardized test items can be subjectively scored. Even though they are constructed to be objective, it is nearly impossible for the examiner to completely eliminate some subjective scoring. This may skew the scores a client receives on a given test.

- Standardized tests assume that the average performance for an age level on a certain task sets the age norm for that behavior. This is not always the case because of individual differences that exist within the range of normalcy (Hegde, 1998). Therefore, clinicians should be wary of using age-equivalency scores to describe a child’s level of functioning.

- Results from standardized tests do not always translate into specific treatment goals. Additional measurements, including language samples and base rates, are often necessary to specify treatment goals.

- The stimulus items or procedures of some standardized tests may be discriminatory or inappropriate for children of diverse cultural or socioeconomic backgrounds. Lack of responses to such items may not suggest skill deficiency.

- The participant sampling may be biased against children from low income, minority, or rural families. Although sampling of minority children in test standardization is on the increase, many historically standardized tests have included, for the most part, children from the mainstream social strata.

- Most tests are standardized on local, not national, samples. Although efforts are being made to draw larger and nationally more representative samples, many tests are not designed to be used nationally; nonetheless, they are used nationally.

- Children of some ethnocultural minority groups may react differently to unfamiliar examiners. Children from different cultural backgrounds may communicate better in their own social and cultural situations than in formal and unfamiliar testing situations.

It is unlikely that the use of standardized tests, regardless of their increasingly appreciated limitations, will decline significantly because of some of the historic views, current training of clinicians, and educational policies. Therefore, it is important to consider ways of limiting the negative consequences of tests and using them in a prudent as well as limited manner.

### Prudent Use of Standardized Tests

Probably the best way of limiting the negative consequences of standardized tests is to develop and use a new approach to assessment that integrates the strengths of various alternative assessment procedures as well as those of standardized assessment. In
Prudent use of standardized tests that help minimize their negative consequences requires that the clinicians consider (a) the ethical issues in standardized assessment, (b) criteria for selection of standardized tests, (c) fair and valid interpretation of test results, and (d) assigning a valid role to standardized tests in assessment. Among several sources the clinicians should regularly consult are the Standards for Educational and Psychological Testing, which is a set of standards jointly developed and published by the American Educational Research Association (www.aera.org), American Psychological Association (www.apa.org), and the National Council on the Measurement in Education (www.ncme.org). The current standards were published in 1999; the standards are being revised in 2007 (to check on the progress, visit www.ncme.org). In addition, speech-language pathologists should consult the relevant position papers and other publications of the American Speech-Language-Hearing Association.

**Ethical Issues in Standardized Assessment**

Ethically responsible use of standardized tests is based mostly on the codes of conduct that professionals adhere to. For instance, before administering a test, the clinician should get herself or himself trained in its administration and scoring. This is part of the ethical code of most professionals, who are required to provide only those services they are qualified to offer (American Speech-Language-Hearing Association, 2003; Anastasi & Urbina, 1997).

An ethical principle that authors should adhere to is to avoid publishing or releasing in other ways, tests that are not adequately standardized. A tests should be made available for general use only when it has been standardized with acceptable reliability and validity. Diagnostic tests of various kinds should not be made available in popular media.

All clinicians are expected to protect the privacy of their clients, and this applies to the results of standardized tests as well. Clinicians should always seek to test skills that are relevant to making a valid diagnosis, designing an intervention plan, and providing other kinds of services. All test takers should be informed of the purpose of testing, the general testing format, and how the results will be used. The results of the tests and other assessment procedures should be held with strict confidentiality.

Test publishers, too, have certain ethical responsibilities in selling standardized tests. Publishers limit the sale of tests to qualified (usually a master’s degree in a relevant discipline), and perhaps also licensed, professional. Tests should not be sold to those who do not meet such qualifications, nor should such persons seek to purchase standardized tests.

**Criteria for Selection of Standardized Tests**

Selecting the most appropriate test for a given child or situation can be a daunting task. What follows is a brief description of major criteria that can be helpful in selecting a test that is appropriate for the child being assessed.

*Select a test that has a detailed and comprehensive test manual.* The test manual should contain specific administration procedures and detailed information on the size and make-up of the normative sample, test reliability, and validity.
Select a test that is based on a large and diverse normative sample. Tests that have sampled too few participants or have not included sufficient number of participants from diverse cultural, linguistic, and socioeconomic groups are generally inappropriate. If a test had sampled subgroups, the number of participants in each should be 100 or more.

Select a test that samples the skills adequately. Tests that give multiple opportunities to produce a skill (e.g., a given phoneme or grammatic morpheme) should be preferred to those that give only one or to opportunities.

Select a test that has current normative data. Tests should be updated on a regular basis to keep the normative data current. Racial and cultural composition of societies and educational and economic status of people change over time and the test norms should reflect such changes.

Select a test that has strong reliability and validity. Tests that offer different kinds of reliability and validity should be preferred over those that report only one kind or a weaker kind. Tests that offer strong and extensive evidence of reliability and validity data should be preferred over those that offer only limited or weak data.

Select a test that has clear instructions for the administration and scoring of the test. Tests that are difficult to administer or score because of limited or unclear description of administration and scoring procedures should be avoided.

Select a test that has appropriate stimulus items. Tests that contain dated stimuli and stimuli that are socially and culturally inappropriate to the child being assessed should be avoided.

Select a test that has been revised recently. Some tests are more frequently revised than other tests. More recently revised tests are preferable to older editions of the same tests or other tests that have not been lately revised.

Select a test that you are well-trained to administer. A clinician who is not trained in the administration and scoring of a new or better test that becomes available should first get the training.

Select a test that will yield useful diagnostic information and help design treatment goals. Some tests may provide solid assessment results that are not easily translated into treatment goals. Tests that generate both solid diagnostic information and treatment target suggestions should be preferred.

The clinician would not base selection of a test only on one or two criteria just summarized. To be selected, a test should meet multiple criteria. The greater the number of criteria a given test meets, the more useful it is in making reliable and valid clinical judgments.

Fair and Valid Interpretation of Test Results

It is generally believed that test results are strictly interpreted according to the test manual. Although all clinicians strive to make justified interpretations of test results, it is not uncommon for them to face dilemmas in finding the meaning of the results. The manual itself may be unhelpful, ambiguous, or unclear about interpretations.
Fair and valid interpretation of test results has been a particular concern in assessing children and adults who belong to ethnocultural minority groups on whom the tests administered have not been standardized. Potential misinterpretation of results in cases of such adults and children is now well-recognized. The best solution is not to administer tests that are standardized only on mainstream children to children of ethnocultural diversity. Unfortunately, in certain situations (as in a school district that requires standard test scores to qualify a child for services), clinicians may be forced to use an available test. In such cases, it is the clinician's responsibility to interpret the test results in light of the child's ethnocultural and linguistic background. Rigid adherence to the manual may lead to faulty interpretations of results. This issue is further addressed in the next section.

The need for fair and valid interpretation may require some flexibility even in the case of children of the kind who were included in the standardization sample. For instance, positive or negative results of an articulation test that gives only one opportunity to test a speech sound may have to be checked against the results of a speech sample before concluding that the child does or does not produce the sound. If the clinician observes that the child was somewhat puzzled by a speech or language stimulus item and gave a wrong response may hold judgment until other kinds of evidence emerges to make a better judgment.

Although the results of a test should be generally interpreted according to the manual, the clinician should not ignore signs that the test, for whatever the reason, may not have sampled the child's behavior adequately or fairly. Doing so with the assumption that the manual is inviolable may do more harm than good to the child tested.

The Role of Standardized Tests in Assessment

Finally, prudent use of standardized tests requires that the clinician assign a proper role to them in the larger context of assessment. Clinicians who believe that the standardized tests are the only or the main means of assessment of communication skills will give them an exaggerated status, and consequently, run the risk of drawing invalid conclusions. Such clinicians fail to assign a proper role to standardized tests in assessment.

Attention to all ethical issues, careful selection of tests, and fair and valid interpretations will be productive strategies if the tests are seen as one of several sources of information on the child's communication skills. Tests need not be a primary source; it should never be the sole source. We have noted earlier that a detailed case history, a carefully done interview, an adequate language sample taken in the clinic and home, and reports from other professionals provide information that tests cannot provide. Furthermore, we will see in Chapter 5 that various alternative approaches, including client-specific and criterion-referenced, authentic, and portfolio assessment offer valuable information while at the same time help to mitigate the negative consequences of standardized tests.

A proper role for standardized tests is that they are one potentially useful source of information. No strong advocate of standardized tests has considered their results infallible. It is often said and written that case history, interview, and language samples supplement standardized tests. Assigning a proper role to tests, one might say that standardized tests supplement other procedures that adequately sample social communication in more naturalistic contexts.
Assessment in a Multicultural Society

As noted in earlier sections of this chapter, a significant issue with the use of standardized tests is their appropriateness in assessing children who belong to various ethnocultural minority groups. In fact, the limitations of standardized tests are often discussed in the context of assessment in an ethnoculturally diverse society.

Tests standardized on mainstream American children are especially not useful in assessing varied ethnocultural groups. Children who speak a dialect other than Standard American English either because of African American culture or because of bilingual status of the family are a case in point. Because of special concerns in assessment of children belonging to various ethnocultural groups, we have devoted Chapter 4 to this topic.