Chapter 5

Norms and Scores

Sharon Cermak

There are three kinds of lies: lies, damned lies, and statistics.

–Disraeli

INTRODUCTION

Suppose that on the *Bayley Scales of Infant Development*,[1] Kendra, age 14 months, received a score of 97 on the Mental Scale and 44 on the Motor Scale. Does this mean that her mental abilities are in the average range? Does it mean that she is motorically retarded? Does it mean that her mental abilities are twice as good as her motor abilities?

The numbers reported are raw scores and reflect the number of items that Kendra passed. Since there are many more items on the Mental Scale than on the Motor Scale, it is expected that her raw score on this scale would be higher. These numbers are called raw scores and cannot be interpreted. It is not known what their relationship is to each other, or what they signify relative to the average child of comparable age. Raw scores can be interpreted only in terms of a clearly defined and uniform frame of reference. The frame of reference in interpreting a child's scores is based upon the scoring system which is used on the test.

Scores on psychological, educational, developmental, and per-

Sharon Cermak, EdD, OTR, is Associate Professor of Occupational Therapy at Sargent College of Allied Health Professions, Boston University, University Road, Boston, MA 02215.

ceptual tests are generally interpreted by reference to *norms*. Norms represent the test performance of the standardization sample. They are established by determining what a representative group of persons do on a test. (This process is described in detail in Chapter 4.) In Kendra's example, her performance would be compared to the performance of the babies who were part of the normative sample when the *Bayley Scales of Infant Development*[1] was standardized. Thus, the raw score is converted to a derived standard score in order to: (1) indicate Kendra's standing relative to the normative sample, and (2) provide a way to compare Kendra's Mental Scale to her Motor Scale.

This chapter will address how and why various scoring systems for norm-referenced tests are established. Norm-referenced tests show how well a person does in comparison to an established set of performance scores. The statistical concepts related to the description of test performance will be addressed including measures of central tendency, variability, and types of distributions. Then a description of the various types of scoring systems applicable to standardized test development will be presented, including age equivalent scores, grade equivalent scores, percentiles, standard scores, deviation IQ scores, and stanines. Following this, the value and limitations of norm-referenced scoring systems will be discussed.

STATISTICAL CONCEPTS

Before proceeding to a discussion of the applied aspects of testing, it is helpful to examine some statistical concepts that underlie the development and use of norms.

Measures of Central Tendency

One way to describe a group of test performance scores is by examining *measures of central tendency*, that is using one number to represent the performance of the group. For example, to establish how long the average six-year-old can stand on one foot, 20 six-year-olds could be tested (see Table 1) and one number could be calculated to represent the average length of time.

There are three possible ways to describe this performance, the

TABLE 1. Scores of 20 Six-Year Olds on a Test of Standing Balance

Name	Score	Name	Score	Name	Score
Jolene	37	Daphne	19	Dae-Won	12
Anne	30	Megan	18	Joey	10
Randi	27	Ellen	17	Justin	9
Nicole	26	Nathan	15	Ani	7
Jose	24	Drew	15	Kara	6
Judith	23	Amanda	15	Marcia	1
Simon	19	Linda	14		

Mean = 17.2	Sum of Scores ÷ Number of Children
Mode = 15	Score which occurs most frequently
Median = 16	The score which is ranked in the middle

mean, *median*, and *mode*. The *mean* is the arithmetic average which is computed by adding the scores of all the children and dividing it by the number of children tested, in this case, 20. In the example given in Table 1, the mean is 17.2 seconds.

The mean is the most common, and generally the most useful, measure of central tendency. However, a possible disadvantage of the mean is that if one score is extremely different from the others, this one score will distort or skew the mean. For example, if 19 of the six-year olds stand on one foot between 5 and 35 seconds, but one child stands for 180 seconds, the mean would be substantially higher than the 17 seconds and not represent the typical performance of the group. In general, when a test is standardized on a large number of children, one extreme score does not influence the result as dramatically as in this example because the other scores offset it.

Another measure of central tendency is the *mode*. This is the score that occurs more often than any other single score. For example, the modal age of first graders is six. In Table 1, the mode is 15 because this score occurs three times, whereas all the other scores occur only once or twice. The mode is often used with *nominal* or

categorical data and when scores are so highly *skewed* that one value predominates.

A third measure of central tendency is the *median*. The median is the score most in the middle of all the scores. It is the point or score that divides the distribution of scores in half. In Table 1, the data is ranked from highest to lowest and the score in the middle (between the 10th and 11th scores of 17 and 15) is 16, thus the median is 16. The median is used when data are in *ranks*, or when the presence of a few extreme scores distorts the arithmetic average (the mean). It is the value of the median that is also its limitation. The median does not reflect the magnitude of the impact of every score in the distribution, even when certain of these scores are very high or very low. For example in Table 1, if one of the children, Daphne, had stood on one foot for 115 seconds instead of 19 seconds, the median of the distribution would not have changed and would continue to be reflective of the scores in the distribution. On the other hand, the mean would have changed from 17.2 to 22.0.

The mean, median, and mode are differentially related depending on the symmetry or skew of a distribution.[2] In some distributions (i.e., the normal curve, where the distribution is symmetric and unimodal), all three measures of central tendency are equal, but in many distributions they are different. The choice of which measure is best will differ from situation to situation. The mean is used most often because it includes information from all of the scores. However, when a distribution has a small number of very extreme scores, the median may better describe central tendency.[3]

Variability

Another method of describing a set of scores is by their *variability*, also known as *measures of dispersion*. For example, suppose it is known that the average six-year-old can stand on one foot for 17 seconds. Amanda however was only able to stand on one foot for 15 seconds. How can Amanda's performance be interpreted? What is normal for her age group? Knowing the average score helps but is not enough. The range of scores that are considered appropriate for a six-year-old needs to be considered. That is why variability is

examined, to determine whether different groups of scores have different dispersions or distributions.

For example, in the following 2 sets of scores both have a mean of 3.

Set A: 1,2,2,3,3,3,4,4,5

Set B: 1,1,1,2,3,4,5,5,5

However, graphically it is obvious that they are very different (see Figure 1). If only indicators of central tendency were provided, the two sets of data would not be adequately described. An indicator of the variability or dispersion of the scores is necessary.

Two ways of describing the variability of a set of scores are in common use. The *range* is the difference between the largest and the smallest scores in a distribution. It is calculated by subtracting the smallest score from the largest score, and adding one. This measure is crude and unstable. As can be seen in the preceding example, the range [(5-1) + (1) = 5] is the same for both sets of scores. Also, one high or low score would have a major effect on the range.

A more frequently used measure is the *variance*. This is a measure of the total amount of variability in a set of test scores. The variance is based on the difference between each individual's score and the mean of the group. The variance measures how widely the scores in a distribution are spread about the mean. It is calculated by subtracting the difference between each score and the mean, squar-

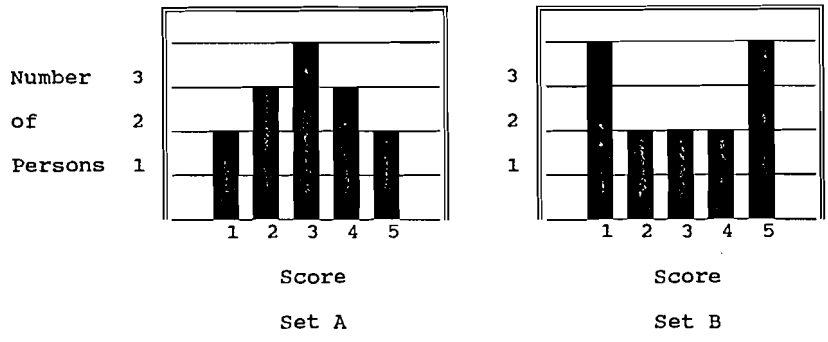


FIGURE 1. Distributions from Two Sets of Scores, Each with a Mean of 3

ing each of these numbers, adding them, and dividing the total sum of squares by the number of scores.[4]

Since the variance is calculated using squared deviations, the result is in terms of *squared units*. Because squared units are unwieldy to use in other calculations, generally the *square root of the variance* is computed. The square root of the variance is known as the *standard deviation*. It is the most commonly used measure of variability, and indicates the dispersion of scores around a given score, usually the mean. Generally, the larger the standard deviation, the more widely scattered are the scores. In the preceding example (Figure 1), the standard deviation of the scores in set A is 1.15 ($\sqrt{12/9}$) and in set B is 1.69 ($\sqrt{26/9}$). The importance and application of the standard deviation is discussed in the sections on normal curve and standard score.

The Normal Curve and Other Types of Distributions

The *normal curve* is a statistically derived distribution and is particularly helpful for comparing a child's score to that of other children. The *baseline* of the normal curve is divided into *whole and fractional standard deviation units* and serves as an interpretive yardstick for contrasting different examinee's test performances.

The normal curve is a *symmetrical bell-shaped curve*, in which the mean, median, and mode are identical. Because of the properties of the normal curve, the distribution can always be divided into predictable proportions, and there is an exact relationship between the area bounded by given standard deviation units and the proportion of cases found within that area under the curve (see Figure 2).

The total area under this curve equals 100%. Thirty-four percent of the cases (representing 34% of the area under the curve) are expected to fall between the mean and plus one standard deviation. Similarly, the area between the mean and minus one standard deviation represents 34% of the total area. It can be seen by examining Figure 2 that the area under the normal curve between plus and minus one standard deviation of the mean is 68%, thus if the sample is normally distributed, 68% of the cases would receive scores in this range. Furthermore, 95% of the area under the normal curve is

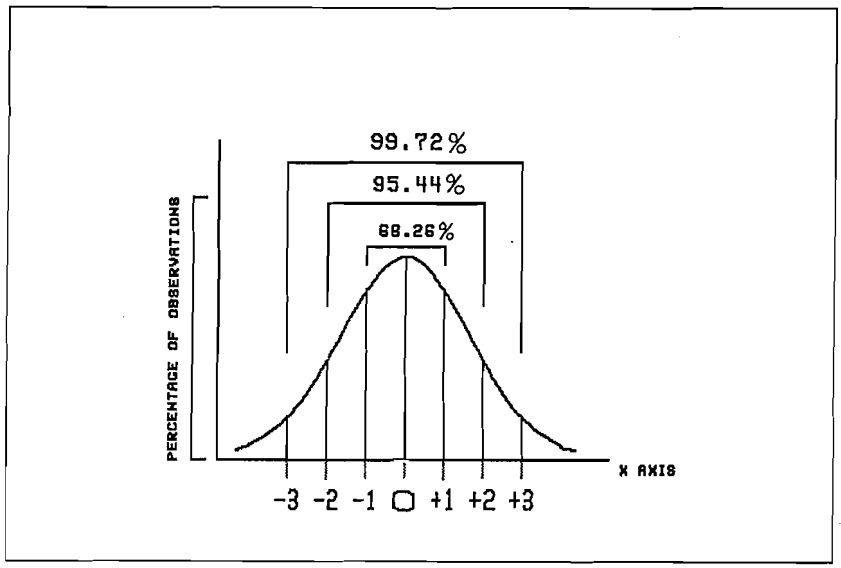


FIGURE 2. The Normal Curve

between plus and minus two standard deviations, and 99.7% of the area is between plus and minus three standard deviations.

Since the majority of scores are between -1.0 and $+1.0$ standard deviations from the mean (that is, 68% of the scores fall in this range), scores in this range are generally considered within normal limits. A standard score below -1.0 is considered to reflect possible dysfunction and a standard score above $+1.0$ is considered to represent a strength. (As discussed in Chapter 7, when interpreting an individual score, it is important to consider the standard error of measurement.)

Although many educational and behavioral variables are distributed in an approximately normal manner, not all scores fall into a normal distribution. On some measures, scores cluster at one end of the curve in what is known as a *skewed distribution*. For example, on the Space Visualization Contralateral Use (SVCU) score (derived from hand usage on the Space Visualization Test of the *Southern California Sensory Integration Tests*),[5] scores of seven-year-olds tended to cluster in the 26-29 range, although a few children

obtained scores in the 0-25 range.[6] In this instance, when describing the average or typical performance of seven-year-olds, use of the group mean of 26 (and standard deviation of 3.8) would not be appropriate. The few low scores lowered the mean so that it did not reflect typical performance. Since the maximum score on the SVCU is 29, the few low scores could not be offset. In fact, a better measure of central tendency with this distribution is the mode (28), because it occurred more than twice as frequently as any other score, and thus is a better representation than the mean of what a "typical" seven-year-old would score.

Figure 3 shows the distribution of test scores for the 30 seven-year-olds in the study. This is a *negatively skewed distribution*.

Skew refers to the symmetry of a distribution. The distribution of scores on a test that is easy and on which most students earn high scores is known as a *negatively skewed distribution*. Conversely, on a very difficult test in which most children earn low scores, the distribution tails off to the higher end of the continuum and is called a *positively skewed distribution*. Figure 4 shows an example of a positively and negatively skewed distribution.

NORMS AND SCORES

One of the distinguishing characteristics of a standardized test is the provision of norms to aid in the interpretation of individual scores. Norm-referenced test interpretation involves some method of examining how an individual's test score compares to the scores of others in some known group. An individual's test performance is typically interpreted by comparing it to the performance of a group of subjects of known demographic characteristics (age, sex, race, etc.). This known group is called the *normative sample* or *norm group*. Norms are usually in the form of a table of equivalents between raw scores (i.e., number of correct responses) and one of several *derived scores*.

Derived scores are based on a transformation of the raw score to some other unit of measurement which enables comparison to the norm. One method of conceptualizing the kinds of *transformation* scores can undergo is by categorizing them as either *Developmental Scores* or *Within Group Measures*.[4] In developmental norms, the

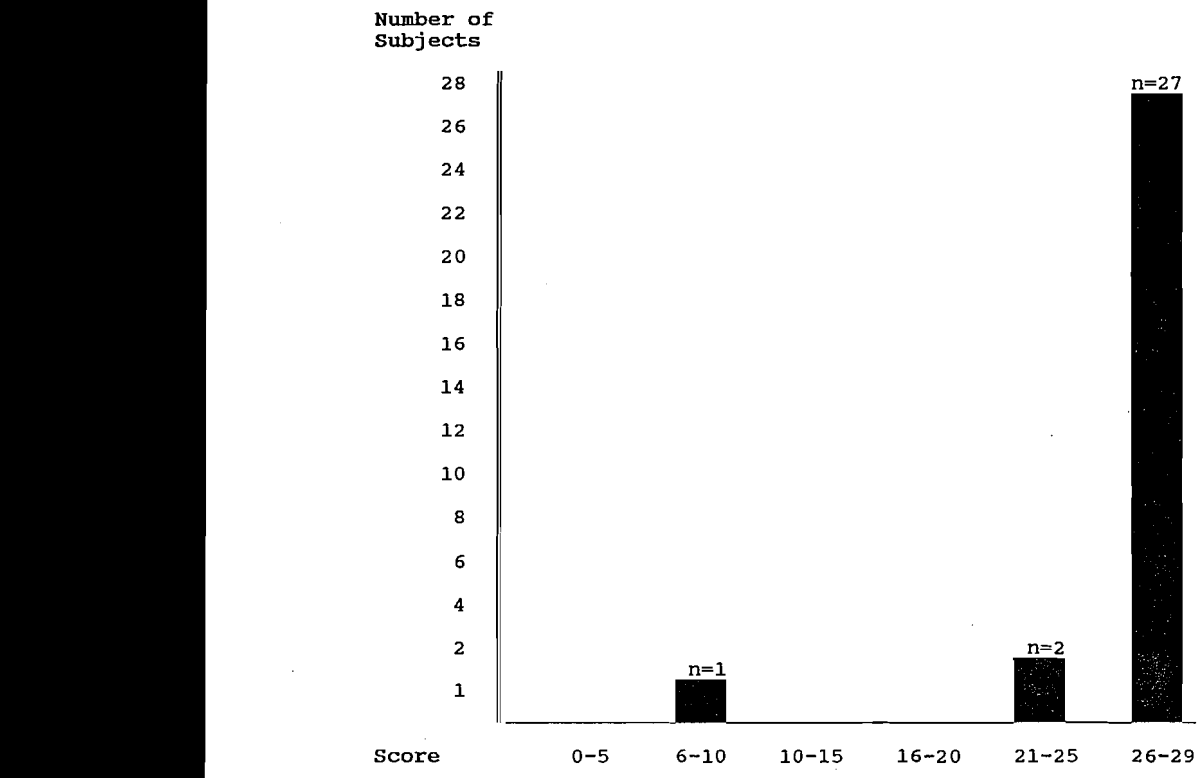


FIGURE 3. Negative Skew of the Distribution of Scores for Seven-Year-Olds on the SVCU

individual's performance is compared to the performance of children at many ages, and the score gives an indication of "how far along the normal developmental path the individual has progressed."[4,p77] In within group norms, the individual's score is compared to performance of a standardization group, usually of comparable age or grade. Whereas within group scores have a uniform and clearly defined quantitative meaning, developmental norms are psychometrically less sophisticated.[4] How test results are organized and presented depends to a large extent on the type of interpretation to be made. The various types of transformed scores are listed below,

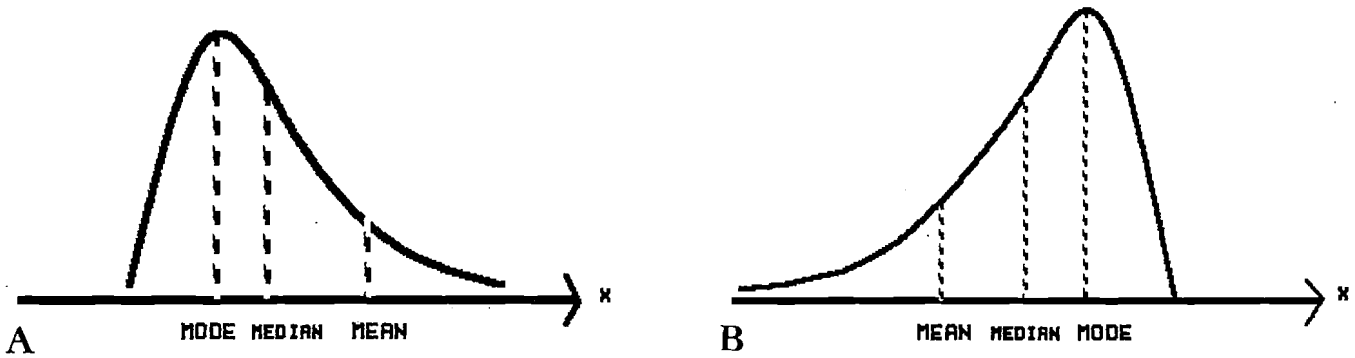


FIGURE 4. **A** Represents a Positively Skewed Distribution and **B** Represents a Negatively Skewed Distribution

grouped in the appropriate category. The following discussion examines each type of score within the two categories.

Developmental Scores	Within Group Scores
age equivalent	percentile rank
grade equivalent	standard scores
ordinal scale	stanines
	deviation IQs

Age Equivalent Scores

The *age equivalent* of a particular raw score is the chronological age of those children whose mean raw score is the same as the raw score in question; it is the raw score that a child at the 50th percentile in a particular age group would receive. The *Developmental Test of Visual-Motor Integration* (VMI)[7] is a test that uses age equivalents to report scores (although the more current revision of the test includes both age equivalents and a number of other methods to convert raw scores).[8]

For example, suppose that the average performance of 10-year-olds on the VMI was 17 correct drawings and that Michael drew 17 of the forms correctly. He would then have earned an age equivalent score of 10 years. Generally, age equivalents are expressed in years and months, using a hyphen.[2] For example, a score of 10 years 4 months is commonly represented by 10-4.

The primary advantage of age norms is that they are easily understood. However, many variables cannot be expressed meaningfully using age norms. For example, acuity of vision does not change during childhood. If a 20-year-old has 20/20 vision (which is normal for a 10-year-old also) it is not meaningful to say that the 20-year-old has an age equivalency score of 10-0. In addition, for many factors the age norms are only appropriate within a certain period of growth. For example, on a test of tactile perception, a 16-year-old might receive an age equivalent score of 8-0 because certain tactile perceptual abilities mature in the 6- to 8-year range (e.g., performance of a 6- to 8-year-old would be comparable to performance of an adult). If the score of the 16-year-old were writ-

ten out in age equivalents as 8-0, it would appear as dysfunction, when in fact it represents normal abilities.

An additional problem with age equivalents is that a year's difference at one time in life is frequently different than a year's growth at another time. Take the example of a child who is delayed one year. If the child's chronological age is four and the child is delayed one year, this is a 25% delay because the age equivalency score (3) divided by the chronological age (4) is 75%. However, if the child is 10, and is delayed one year, his delay is only 10%, because the age equivalency score (9) divided by the chronological age (10) is 90%.

An additional problem with age equivalency scores is that they represent what a child in a particular age group at the 50th percentile would receive for a score. However, what is actually "normal" includes scores lower than the 50th percentile (in fact, performance as low as the 16th percentile which is one standard deviation below the mean is generally considered to be within normal limits). If there is a skewed distribution in the norm sample, the interpretation of age scores becomes even more difficult.

Grade Equivalent Scores

Grade norms or *grade equivalents* are often used in educational and academic achievement tests. A grade equivalency score means that a child's raw score is the average performance for that grade (grade equivalents may be based on the median or mean).[9] Thus, a grade equivalent of 4.6 is read as fourth grade, sixth month level. (The summer months are assumed to represent an increment of one month on the grade equivalent scale.)[9] Generally, a decimal point is used in the representation of grade scores.

The advantage of grade equivalency scores is that they are easy to understand. The disadvantages are similar to those presented for age equivalency scores in that they are easily misinterpreted. For example, Niki, a sixth grade child achieved a reading grade of 9.5 on the *Wide Range Achievement Test* (WRAT).[10] This does not mean that she reads at the ninth grade level. It means that what she knows as a sixth grader, she knows well. She did very well on a sixth grade reading test but did not take a ninth grade reading test.

Since grade equivalency scores depend on the particular items placed on a test and the particular norm group used, they are not interchangeable between tests or for different forms which are administered to different grades. It is a misinterpretation to say that a grade equivalent of 3.2 on the WRAT means the same thing as a grade equivalent of 3.2 on the *Peabody Individual Achievement Test*.[11] Also, grade equivalency scores between subtests, even on the same test, are not necessarily comparable.

Another problem with grade equivalency scores is that test publishers often do not have the financial resources to do a nationwide stratified sampling of children of all ages, grades K through 12, on a month-by-month basis. The publishers usually test at only a few grades, establish a relationship between test scores and grades, and then use the *relationship line* (a statistical manipulation) to estimate the various grade-month points.[9,12] These estimates are made by *interpolation* and *extrapolation*, and are often based on the assumption that what is tested is consistent from year to year. Nitko[9] provides further elaboration of the methodology in computing grade equivalents.

Another problem with the use of grade equivalency scores alone is that they do not provide information about an individual's percentile standing. Thus, an individual might get a higher grade equivalent on a reading test than on a mathematics test, yet have a substantially lower percentile rank on the reading test than on the math test.

Table 2 displays a hypothetical third grade pupil's test results. In Melissa's case, two identical grade equivalents have the same percentile rank. With Deborah, one grade equivalent can be higher than another yet associated with a lower percentile rank than the lower grade equivalent. This is because the scores for one subject area are more variable than for another.

Ordinal Scales

Ordinal scales, another type of developmental score, are based on developmental sequences, and are used to identify stages reached by a child. Qualitative descriptions are often provided. In ordinal scales, successful performance at one level implies successful performance at other preceding levels. The scales developed

TABLE 2. The Relationship Between Grade Equivalents and Percentile Ranks for Three Students

Child	Grade Equivalent	Percentile Rank	Grade Equivalent	Percentile Rank
Melissa	3.9	80	3.9	80
Sara	3.9	68	3.3	68
Deborah	3.9	68	3.6	74

within this framework are based on the sequential patterning and uniformity of developmental sequences. For example, in grasp, use of the entire hand in palmar prehension precedes thumb in opposition to the palm which precedes thumb-finger opposition.

An early example of the application of ordinal scales is the work of Gesell and associates[13] in which the sequential patterning of early behavior development is emphasized. The most well known ordinal scales are Piaget's stages of development.[14] These stages, which span the period from infancy through adolescence, are known as the sensorimotor, preoperational, concrete operational, and formal operational stages. Examples of ordinal scales based on the work of Piaget are the *Ordinal Scales of Psychological Development* designed for children ages two weeks to two years,[15] and the *Concept Assessment Kit-Conservation*, designed for ages four to seven years.[16]

Rank ordering of tasks is performed first in designing an ordinal scale, then age may be considered. Since these scales generally provide information about what the child is actually able to do, they share important features with criterion-referenced tests. A major problem encountered in ordinal scales is inconsistency in the anticipated sequences. "There is a growing body of data that casts doubt on the implied continuities and regularities of intellectual data."[17,p276] Moreover, when dealing with special populations, the developmen-

tal sequence may not be the same for the handicapped child as for the non-handicapped child.

Percentiles

A *percentile rank* indicates an individual child's position relative to the standardization sample. It represents the percentage of the standardization group who scored at or below a given raw score.[2,3,4,9] For example, if a raw score of 33 indicates a percentile rank of 80, it means that 80% of the group members had raw scores of 33 or less. Conversely, a student with a score of 33 scored as well as or better than 80% of the normative sample on the test.

The middle score in a distribution is the one that equals 50% of the scores. This score, at the 50th percentile, is the median and describes the average performance in a percentile distribution.[3]

A percentile-equivalency table typically provides raw scores and their percentile equivalents. The *Miller Assessment for Preschoolers* is an example of a test that utilizes percentiles.[18] The scoring system for the *Test of Motor Impairment* is also based on percentiles.[19] This test yields an index of dysfunction which is based on the percentage of subjects in the standardization sample scoring in a comparable manner.

Advantages of percentiles are that they are easy to understand, easy to compute, and suitable to any type of test. Therefore, they are widely adaptable and applicable. In addition, the table of norms can always be interpreted in the same way, regardless of the nature of the distribution of raw scores from which they are derived.[20] In other words, even when the distribution of scores is not normal, the interpretation of percentile norms does not change.

However, when using percentile ranks, it is important to remember that they refer to the percentage of persons earning equal or lower scores and not the percentage of items answered correctly. Also, percentile ranks do not increment equally with raw score intervals. In other words, if a change from the 50th to the 60th percentile represented an improvement of 5 raw score points, the change from the 85th to the 95th percentile would not also represent 5 raw score points if the scores are normally distributed. Since scores tend to be clustered near the middle in a normal distribution,

a small raw score change near the center would result in a larger percentile change. A larger raw score difference at the extreme ends of the curve is needed to yield comparable percentile changes. This point is illustrated in Table 3 in performance on the *Stanford-Binet Intelligence Scale*[21] which has a mean of 100 and a standard deviation of 16.

As can be seen, an increase in 16 IQ points from 100 (the average score) to 116 represents a shift from the 50th to the 84th percentile, a change of 34 percentile units. In contrast, an increase in IQ of the same 16 points, but from a score of 132 to 148 (which is near the high end of the distribution), represents a percentile shift of less than two percent. Because of this characteristic of percentiles, these norms are ordinal not interval scales. As such, it is inappropriate to compute arithmetic means of these values or correlate them with other measures using a Pearson product-moment coefficient.[20]

Standard Scores

Standard scores express an individual's distance from the mean in terms of standard deviation or variability of the distribution. As previously described, a percentile only indicates how a specific individual's test score compares to other examinees from the standardization sample. However, a standard score represents in standard deviation units where an examinee's score is *with reference to the mean* of the distribution of the standardization sample.

TABLE 3. Relationship Between IQ and Percentile Scores

IQ Score	Percentile	% change
100	50	34
116	84	
132	98	14
148	99.9	2

Standard scores are derived scores that transform raw scores in such a way that the set of scores always has the same mean and the same standard deviation. They are used appropriately only with *equal-interval* or *ratio scores*. The advantages of standard scores are that they have uniform meaning from test to test and scores can be compared between tests. Moreover, unlike the percentile, the standard score unit has the same meaning throughout the range as illustrated in Table 4. As can be seen, an increase in IQ of 16 points (one standard deviation) from either 100 to 116, or from 116 to 132, or from 132 to 148 each represents a standard score increment of 1. In contrast, these score changes represent percentile rank changes of 34%, 14%, and 2% respectively. A disadvantage to standard scores is that they are not as familiar to the layperson.

There are several types of standard scores. Some of the most commonly used types are *z-scores*, *deviation IQ*, and *stanines*.

A z-score is defined as a standard score with a mean of 0 and a standard deviation of 1. A raw score is converted to a z-score using the equation:

$$z = \frac{X - \overline{X}}{SD}$$

In the equation, X = the raw score, $\overline{X}$ = the mean, and SD = the standard deviation.

Z-scores are interpreted in standard deviation units. Thus, a z-

TABLE 4. Relationship Between IQ, Percentile and Standard Scores

IQ Score	Standard score	Percentile	% change
100	0.0	50	34
116	1.0	84	
132	2.0	98	14
148	3.0	99.9	2

score of +1.5 means that the score is 1.5 standard deviations above the mean of the standardization sample. A z-score of −0.8 means that the score is .8 standard deviations below the mean.

Using the example presented previously of six-year-old Amanda who stood on one foot for 15 seconds will further explain this point. The average for the norm group was 17.3 seconds (see Table 1) and the standard deviation was 3.8 seconds. Amanda's standard score would be:

$$z = \frac{15 - 17.3}{3.8} = \frac{-2.3}{3.8} = -.6$$

This can be interpreted that Amanda's score is six-tenths of a standard deviation from the mean. The negative sign shows that the score is *below* the mean. A score is negative when the raw score is below the mean and positive when the raw score is at or above the mean.

Examples of tests which use this unit of measurement are the *Southern California Sensory Integration Tests*[22] and the *Sensory Integration and Praxis Tests*.[23] By using this form of scoring, it is possible to compare a score on one test to a score on another test, and to compare different subtests within a single test. In addition, when scores are distributed in a normal manner, z-scores can be converted to percentile scores. Table 5 provides a conversion of z-score to normal curve percentile rank correspondences. Using this table, Amanda's standard score (z-score) of −.6 would be equivalent to a percentile rank of 27.

A possible disadvantage to z-scores is that they contain decimals and minus values, which are sometimes confusing to interpret. Negative scores seem to imply a problem, however, any z-score between −1.0 standard deviations below the mean and +1.0 standard deviations above the mean is considered to be within normal limits.

In order to circumvent this possible misunderstanding, measurement specialists have developed a variety of scoring systems to transform z-scores, so that all scores within normal limits will be positive. The general procedure involves selecting a desired new mean and standard deviation. All z-scores are then multiplied by the

TABLE 5. Normalized z-Scores and Percentile Equivalents

z SCORE	PERCENTILE	z SCORE	PERCENTILE
3.0	99.9	-0.1	46.0
2.9	99.8	-0.2	42.1
2.8	99.7	-0.3	38.2
2.7	99.6	-0.4	34.5
2.6	99.5	-0.5	30.9
2.5	99.4	-0.6	27.4
2.4	99.2	-0.7	24.2
2.3	98.9	-0.8	21.2
2.2	98.6	-0.9	18.4
2.1	98.2	-1.0	15.9
2.0	97.7	-1.1	13.6
1.9	97.1	-1.2	11.5
1.8	96.4	-1.3	9.7
1.7	95.5	-1.4	8.2
1.6	94.5	-1.5	6.7
1.5	93.3	-1.6	5.5
1.4	91.9	-1.7	4.5
1.3	90.3	-1.8	3.6
1.2	88.5	-1.9	2.9
1.1	86.4	-2.0	2.3
1.0	84.1	-2.1	1.8
0.9	81.6	-2.2	1.4
0.8	78.8	-2.3	1.1
0.7	75.8	-2.4	0.8
0.6	72.6	-2.5	0.6
0.5	69.1	-2.6	0.5
0.4	65.5	-2.7	0.4
0.3	61.8	-2.8	0.3
0.2	57.9	-2.9	0.2
0.1	54.0	-3.0	0.1
0.0	50.0		

W.J. Popham, MODERN EDUCATIONAL MEASUREMENT, (c) 1981, p. 167. Reprinted by permission of Prentice-Hall, Inc., Englewood Cliffs, NJ.

new standard deviation followed by addition of the new mean. One example is known as the *T-score*. A T-score is simply a z-score multiplied by 10 (to eliminate the decimal) and with 50 added (to eliminate the minus value).[12] T-scores have a mean of 50 and a standard deviation of 10. An example of a test that uses T-scores is the *Walker Problem Behavior Checklist*,[24] a checklist of child be-

haviors and characteristics which is completed by the child's teacher and/or parent. Other commonly used derived scores have a mean of 100 and a standard deviation of 15 or 16.

Normalized Standard Scores

When scores fall in a normal distribution, it is possible to state precisely what proportion of the distribution's scores are exceeded by a score at a particular point along the baseline of the curve. For example, a raw score which is +1 standard deviations above the mean in a normal distribution of scores equals or exceeds 84% of all the scores.

However, not all distributions are normal. When it is believed that the attribute being measured is normally distributed in the real world, but the data are distributed in a non-normal fashion, for ease of interpretation, the raw scores can be converted to *normalized standard scores*. A normalized standard score is a standard score (z or T) that would be equivalent to a raw score if the distribution had been normal.[12] The Standing Balance test of the *Southern California Sensory Integration Tests*[22] is an example where failure of the normative sample raw scores to assume a *bell-shaped curve* resulted in an alternate method of scoring standing balance. Non-linear transformations used to calculate normalized standard scores are further described in Popham.[12]

Deviation IQ Scores

One type of normalized standard score is the *deviation IQ score* used with certain tests of mental abilities such as *Wechsler Intelligence Scale for Children-Revised*.[25] The deviation IQ is actually a standard score with a mean of 100 and a standard deviation of 15. (Some IQ tests, such as the *Stanford-Binet Intelligence Scale*,[21] use a standard deviation of 16.) For example, if Bethany has an IQ of 115, this means she has scored one standard deviation above the mean for her age group (and has a percentile rank of 84).

The deviation IQ is valuable to control for variability caused by raw score distributions that have different standard deviations at different ages. The deviation IQ is only comparable from age to age

or from test to test when using the same standard deviations. Wechsler was the first to use the deviation IQ in the *Wechsler Adult Intelligence Scale*.[26] The deviation IQ is also used in the revision of the *Stanford-Binet Intelligence Scale*.[21]

Stanines

A variation of the standard score is the *stanine scale* which is a system of derived scores that divides the distribution of raw scores into nine parts (the term stanine was derived from standard nines).[27] The highest stanine score is 9, the lowest is 1, and stanine 5 is located precisely in the center of the distribution. In normal distributions, stanines have a mean of 5 and a standard deviation equal to 2. Thus, a score between 3 to 7 stanines is considered within normal limits. The percentage of a group that falls within each stanine in a normal distribution is as follows:

Stanine	1	2	3	4	5	6	7	8	9
Percentage	4	7	12	17	20	17	12	7	4

One of the greatest advantages of stanines is that they can be applied to any type of data that approximate a normal distribution and that can be ranked from high to low. The top four percent of the students are assigned to a stanine of 9, the next seven percent to a stanine of 8, etc. Since an individual's stanine is determined by identifying the percentile to which a person's raw score would be equivalent, and then using the percentages to locate the proper stanine, stanines are a form of normalized scores.[12] For example, if Julio's score were equivalent to the 21st percentile, then his score would be in the third stanine.

Disadvantages to stanines are that they reflect coarse groupings of scores. However, this is seen as an advantage by some educators who, because of the imprecision of measurement "prefer to use gross descriptors in communicating test results and thus not misrepresent the precision of data-gathering devices."[12,p169] An example of a test that reports scores in stanines is the *Bruininks-Oseretsky Test of Motor Proficiency*.[28]

Relationship of Percentile, Standard Score, Normalized Standard Score, Deviation IQ (Types of Within Group Scores)

The relationship among stanines, percentiles, as well as other standard scores is shown in Figure 5.

If the scores are based on a normal distribution and when certain statistical conditions are met, then the different types of scales can be translated into any of the others.[4] However, it is important to use caution with between-test comparisons since variables such as the standardization samples may be different and an individual's relative standing on the tests may vary as a function of the standardization sample. For example, the normative sample for the *Screening Test for Auditory Comprehension of Language*[29] was composed of children from Tennessee, whereas the standardization sample for the *Southern California Sensory Integration Tests*[22] was from Los Angeles.

Interpreting Norm Scores

Table 6 provides a summary of various norm-referenced scores. Each type of score describes an individual's performance in reference to his/her location in a norm group. As discussed previously, each type of score has certain advantages and disadvantages. For example, grade-equivalents are easy to understand but are frequently misinterpreted. Standard scores are technically more accurate than grade-equivalents but are more difficult for the layperson to understand.

In considering norm scores, it is important to recognize that they provide relative rather than absolute information. Norms reflect how a particular group (the norm group) performed on a particular test at a particular point in time. Norms as such should not be considered as performance "standards." By nature of the type of scores, 50% of scores are below the mean, and 50% are above the mean. It does not make sense to consider bringing all subjects "up to normal." Similarly, norm scores do not provide any information about the mastery of a skill, although it is often assumed that if a

score is greater than -1.0 standard deviations below the mean, then performance is within normal limits and therefore acceptable.

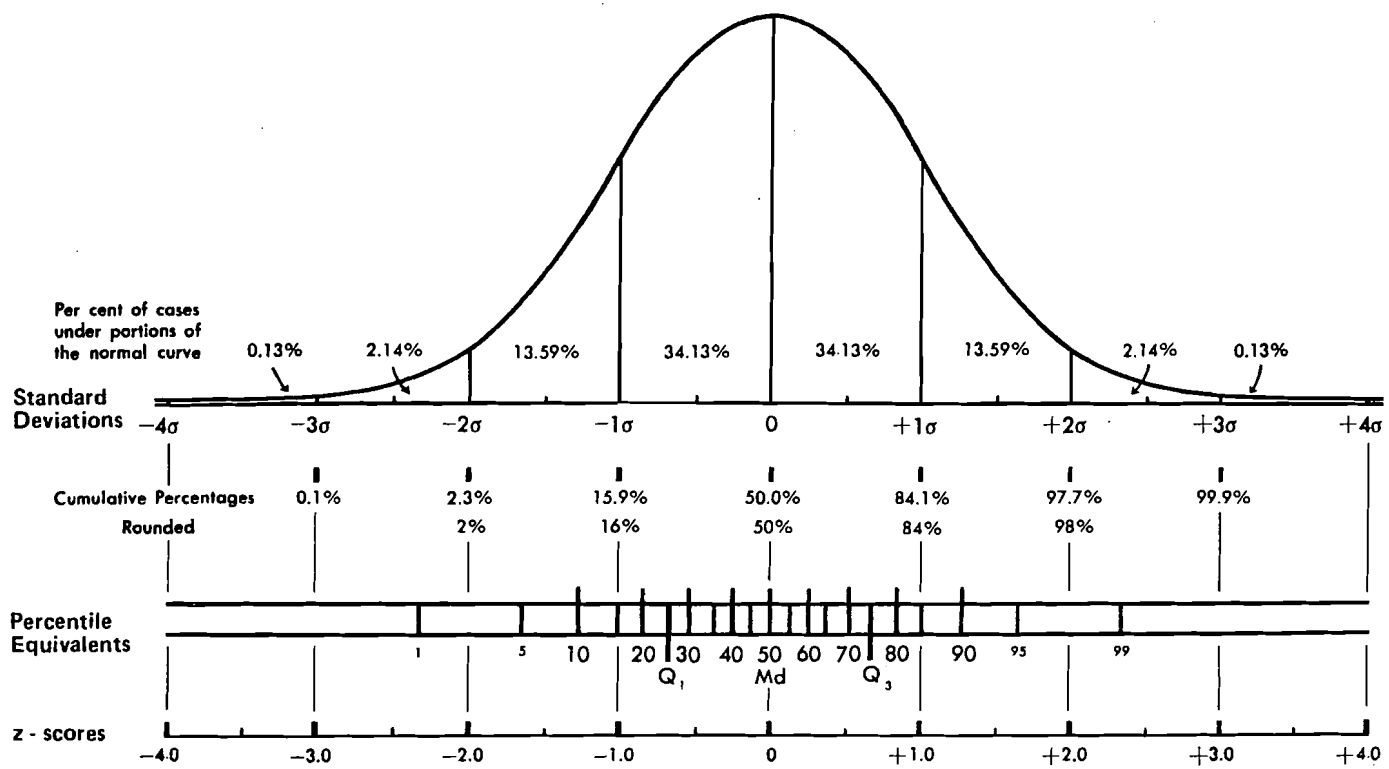
SUMMARY AND RECOMMENDATIONS

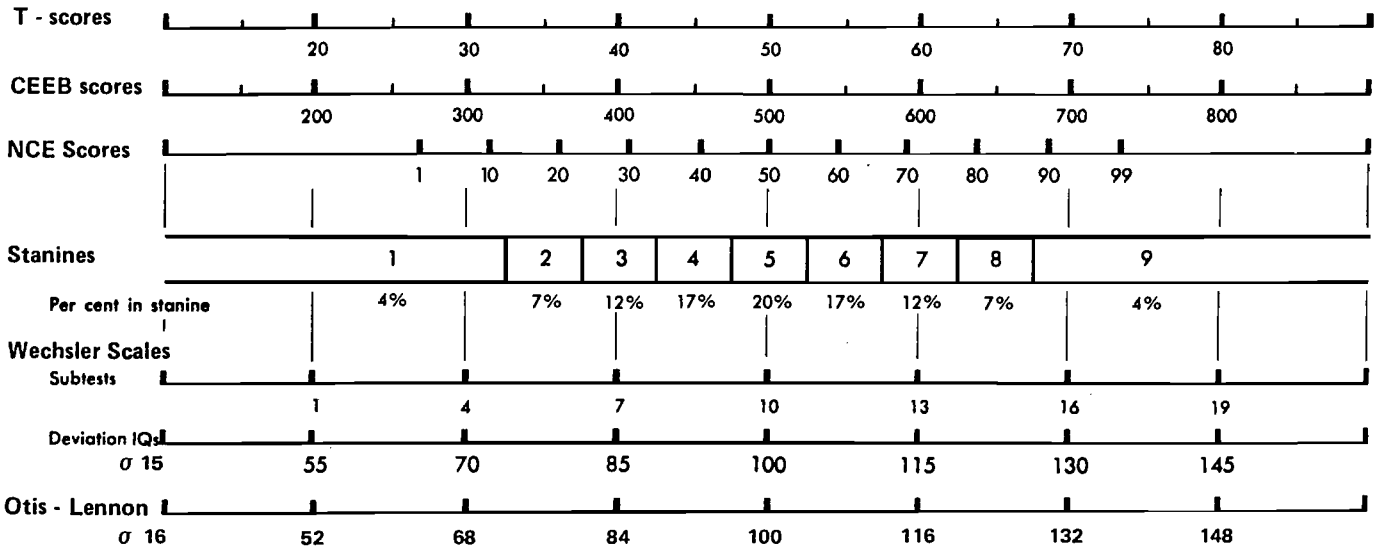
Raw scores, or the number of correct or incorrect answers that a child obtains on a test provides the examiner with relatively little information. In order to be meaningful, raw scores must be converted to a type of reference system. Norm-referenced interpretation is a relative interpretation which is based on the individual's position with respect to some group, usually called the normative group.[3] Two types of norm-referenced comparisons can be made, across ages and within ages. Developmental scores such as age equivalents and grade equivalents compare the performance of students across ages or grades. Within age (or within group) comparisons can be made using several different types of scores such as stanines, standard scores, etc., each with certain advantages and disadvantages.

The manuals accompanying standardized tests should contain tables that permit a tester to convert raw scores to various derived scores such a percentile ranks or standard scores. Some tests such as the *Bruininks-Oseretsky Test of Motor Proficiency*[28] or the *Developmental Test of Visual-Motor Integration*[8] provide one set of tables for converting raw scores to percentiles, and another set of tables for converting to age equivalents. Table 7 presents a list of types of scores provided by tests commonly used by occupational and physical therapists.

The selection of the particular type of score to use and to report depends on the purpose of testing and the sophistication of the consumer. Salvia and Ysseldyke[2] recommend against the use of developmental scores because they are readily misinterpreted. Percentile ranks have the advantages that (a) they require the fewest assumptions for accurate interpretation, (b) the scale of measurement can be ordinal, equal-interval, or ratio data, and the distribution of scores need not be normal, and (c) they are readily understood.[2] Standard scores are convenient for test authors since their use allows the author to give equal weight to various subtests. Standard

FIGURE 5. Percentage of Cases Under Portions of the Normal Curve





THE PSYCHOLOGICAL CORPORATION
HARCOURT BRACE JOVANOVICH, PUBLISHERS
555 ACADEMIC COURT, SAN ANTONIO, TEXAS 78204-0952

TABLE 6. Summary of Various Norm-Referenced Scores

Type of Score	Interpretation	Score	Examples of Interpretations
Percentile Rank	Percentage of scores in a distribution at or below this point.	PR = 60	"60% of the raw scores are at or lower than this score."
z-Score	Number of standard deviation units a score is above (or below) the mean of a given distribution.	z = +1.5	"This raw score is located 1.5 standard deviations above the mean."
		z = -1.2	"This raw score is located 1.2 standard deviations below the mean."
Stanine	Location of a score in a specific segment of a normal distribution of scores.	Stanine=5	"This raw score is located in the middle 20% of a normal distribution of scores."
		Stanine=9	"This raw score is located in the top 4% of a normal distribution of scores."

Deviation IQ	Location of a score in a normal distribution having a mean of 100 and a standard deviation of 16.	IQ = 124	"This raw score is located 1.5 standard deviations above the mean in a normal distribution whose mean is 100 and whose standard deviation is 16. This score has a percentile rank of 93."
		IQ = 84	"This raw score is located 1.0 standard deviations below the mean in a normal distribution whose mean is 100 and whose standard deviation is 16. This score has a percentile rank of 16."
Grade-Equivalent Score	The grade placement at which the raw score is average.	GE = 3.5	"This raw score is the obtained or estimated average for pupils whose grade placement is at the 5th month of the third grade."
Age-Equivalent Score	The age at which the raw score is average.	AE = 7-6	"This raw score is the average score for students whose age is 7 years 6 months."

Adapted from Nitko[9,p381]

TABLE 7. Types of Scores for Tests Commonly Used by Occupational and Physical Therapists

Test Names	Type of Score				
	Age Equivalency	Standard Score	Percentile	Stanine	Other
Bayley Scales of Infant Development [1]		*			
Peabody Developmental Motor Scales [30]	*	*	*		
Denver Developmental Screening Test [31]					*
Miller Assessment for Preschoolers [18]			*		
Bruininks Oseretsky Test of Motor Proficiency [28]	*	*	*	*	

Test of Motor Impairment [19]			*
Test of Visual Motor Integration-Revised [7]	*	*	*
Southern California Sensory Integration Tests [22]		*	
Sensory Integration and Praxis Tests [23]		*	

scores are useful for the examiner since if the distribution is normal, they can be converted to percentile ranks, and also can be used in profile analysis. According to Anastasi,[4] standard scores are the most satisfactory type of derived score. At the stage that decisions are being made about scoring systems by test developers, it is critical to employ experts in tests and measurements as consultants.

REFERENCES

1. Bayley N: *Bayley Scales of Infant Development*. New York, Psychological Corporation, 1969.

2. Salvia J, Ysseldyke JE: *Assessment in Special and Remedial Education*. Boston, Houghton Mifflin, 1985.

3. Wiersma W, Jurs SG: *Educational Measurement and Testing*. Boston, Allyn & Bacon Inc, 1985.

4. Anastasi A: *Psychological Testing*, ed 6. New York, Collier Macmillan Publishers, 1988.

5. Ayres AJ: *Interpreting the Southern California Sensory Integration Tests*. Los Angeles, Western Psychological Services, 1976.

6. Cermak S, Quintero J, Cohen P: Developmental age trends in crossing the body midline in normal children. *Am J Occp Ther*. 34:313-319, 1980.

7. Beery KE: *Developmental Test of Visual-Motor Integration, Administration and Scoring Manual*. Chicago, Follett Publishing Company, 1967.

8. Beery KE: *Revised Administration, Scoring, and Teaching Manual for the Developmental Test of Visual-Motor Integration*. Cleveland, Modern Curriculum Press, 1980.

9. Nitko, AJ: *Educational Tests and Measurement: An Introduction*. New York, Harcourt Brace Jovanovich, 1983.

10. Jastak JF, Jastak SR: *WRAT Manual: The Wide Range Achievement Test*. Wilmington, DE, Guidance Associates of Delaware, 1965.

11. Dunn LM, Markwardt FC: *Peabody Individual Achievement Test*. Circle Pines, MN, American Guidance Services Inc, 1970.

12. Popham WJ: *Modern Educational Measurement*. Englewood Cliffs, NJ, Prentice-Hall, 1981.

13. Gesell A, Amatruda CS: *Developmental Diagnosis*, ed 2. New York, Hoeber-Harper, 1947.

14. Piaget J: *The Origins of Intelligence in Children*. New York, International Universities Press, 1952.

15. Uzgiris IC, Hunt J: *Assessment in Infancy: Ordinal Scales of Psychological Development*. Urbana, University of Illinois Press, 1975.

16. Goldschmid ML, Bentler PM: *Manual: Concept Assessment Kit-Conservation*. San Diego, Educational and Industrial Testing Service, 1968.

17. Anastasi A: *Psychological Testing*, ed 5. New York, Macmillan Publishing Co Inc, 1982.
18. Miller LJ: *Miller Assessment for Preschoolers*. San Antonio, TX, Psychological Corporation, 1988, 1982.
19. Stott DH, Moyes FA, Henderson SE: *Test of Motor Impairment, Henderson Revision*. Guelph, Ontario, Brook Educational Publishing Ltd, 1984.
20. Ahmann JS, Glock MD: *Evaluating Student Progress*, ed 6. Boston, Allyn & Bacon Inc, 1981.
21. Thordike RL, Hagen EP, Sattler JM: *Stanford Binet Intelligence Scale*, ed 4. Chicago, Riverside Publishing Co, 1986.
22. Ayres AJ: *Southern California Sensory Integration Tests Revised*. Los Angeles, Western Psychological Services, 1980.
23. Ayres AJ: *Sensory Integration and Praxis Tests*. Los Angeles, Western Psychological Services, in press.
24. Walker HM: *Walker Problem Behavior Identification Checklist*. Los Angeles, Western Psychological Services, 1983.
25. Wechsler D: *Wechsler Intelligence Scales for Children-Revised*. New York, Psychological Corporation, 1974.
26. Wechsler, D: *Wechsler Adult Intelligence Scale-Revised*. New York, Psychological Corporation, 1981.
27. Grolund NE: *Measurement and Evaluation in Teaching*. New York, Macmillan, 1985.
28. Bruininks RH: *Bruininks-Oseretsky Test of Motor Proficiency, Examiner's Manual*. Circle Pines, MN, American Guidance Service, 1978.
29. Carrow E: *Screening Test for Auditory Comprehension of Language*. Austin, TX, Learning Concepts, 1973.
30. Folio MR, Fewell RR: *Peabody Developmental Motor Scales and Activity Cards Manual*. New York, Teaching Resources Corporation, 1983.
31. Frankenburg WK, Dodds JB, Fandal AW, Kazuk E, Cohrs M: *Denver Developmental Screening Test, Reference Manual*, rev ed. Denver, CO, LaDoca, 1975.

KEY POINTS

1. A norm-referenced standardized test provides norms to aid in the interpretation of individual scores.
2. A group of test performance scores can be described by examining measures of central tendency where one number (mean, median, or mode) represents the performance of the group.
3. A group of test performance scores can also be described by their variability (range or variance) which indicates whether

different groups of scores have different dispersions or distributions.

4. The normal curve is a symmetrical bell-shaped curve in which the mean, median, and mode are identical, and the distribution can always be divided into predictable proportions.
5. Derived scores are based on a transformation of the raw scores to some other unit of measurement which enables comparison to the norm.
6. Age scores are easily understood. A disadvantage is that many variables cannot be expressed meaningfully, because for many factors age norms are only appropriate within a certain period of growth. Age scores represent what a child at the 50th percentile would receive for a score (although "normal" includes lower scores).
7. Grade equivalency scores mean that a child's raw score is the average performance for that grade. They are easy to understand but have disadvantages similar to age equivalency scores.
8. Ordinal scales, a type of developmental score, are based on developmental sequences and are used to identify stages reached by a child. A major problem with them is that successful performance at one level implies successful performance at other preceding levels.
9. Percentile rank indicates an individual child's position relative to the standardization sample. Advantages include that they are easy to understand, compute, suitable to any type of test, and widely adaptable and applicable. However percentile ranks do not increment according to raw score value; therefore means cannot be calculated and they cannot be used to calculate correlations.
10. Standard scores express an individual's distance from the mean in terms of standard deviation or variability of the distribution. Raw scores are transformed in such a way that the set of scores always has the same mean and the same standard deviation. Their primary advantages are that they have uniform meaning from test to test and scores can be compared between tests. Their main disadvantage is lack of familiarity to laypersons.
11. A disadvantage of z-scores (a type of standard score) is that

they contain decimals and minus values which are sometimes confusing to interpret. T-scores can transform the z-scores so that all scores within normal limits are positive.

12. The stanine is a system of standard scores that divides the distribution of raw scores into nine parts. Stanines can be applied to any type of data that approximate a normal distribution and can be ranked from high to low. The primary disadvantage is that they reflect coarse groupings of scores.
13. It is important to use caution with between-test comparisons of scores since variables such as the standardization samples may be different and an individual's relative standing on the tests may vary as a function of the standardization sample.
14. Norm scores provide relative rather than absolute information. Norms reflect how a particular group (the norm group) performed and should not be considered as performance "standards."
15. Selection of the type of scores to use and report depends on the purpose of testing, sophistication of the intended users, and the types of interpretations to be made.