4

A PRIMER ON STATISTICS AND PSYCHOMETRICS

We conquer the facts of nature when we observe and experiment upon them. When we measure them we have made them our servants. A little statistical insight trains them for invaluable work.

—Edward L. Thorndike, American psychologist (1874–1949)

Goals and Objectives

This chapter is designed to enable you to do the following:

- Become familiar with basic statistical concepts and procedures
- Become familiar with the meaning of reliability and the procedures for evaluating it
- Understand the different forms of validity

This chapter will introduce you to basic statistical and psychometric concepts that are used for assessment. A knowledge of statistical and psychometric concepts will enhance your understanding of psychological tests and other clinical procedures and research reports. The basic concepts reviewed in this chapter will also help you understand the material covered in other chapters of this text as well as other areas of psychology and other sciences. Note that the Resource Guide contains a glossary of measurement terms.

THE WHY OF PSYCHOLOGICAL MEASUREMENT AND STATISTICS

Measurement in psychology is usually different from physical measurement. In our everyday experience, we assign numbers to the physical characteristics of objects—such as height, weight, or length—that we perceive directly. Although physical measurement may be more precise than psychological measurement because psychological characteristics are likely to be intangible, both types of measurement are important. Both psychological measurement and physical measurement consist of (a) identifying and defining a dimension (e.g., height) or behavior (e.g., cooperativeness), (b) determining the relevant measurement tool and operations, (c) specifying the rules of measurement, and (d) using a scale of units to express the measurement. Psychological measurement attains the precision of physical measurement when we measure such things as reaction time or how close someone comes to a target. Psychological measurement conveys meaningful information about people's attributes, such as their intelligence, reading ability, adaptive behavior, interests, personality traits, and attitudes, through test scores or ratings that reflect such attributes.

Statistics make life easier by reducing large amounts of data to manageable size, allowing us to study individuals and groups. Statistics also help us to communicate information about test scores, draw conclusions about those scores, and evaluate chance variations in test scores. Only by using statistics can we determine, for example, whether a child's scores on a test administered at two different times differ significantly, whether a child's scores on two different tests differ significantly, or whether the scores of different groups of children on the same test differ significantly. These kinds of determinations are important in evaluating progress and comparing ability levels both within an individual and between individuals. Individual differences are an important focus in the field of psychology. People differ: Some are bright and talented, others less bright and less talented; some are energetic, others lethargic; some are extraverted, others introverted; some are well adjusted, others less well adjusted; and some are good readers, others poor readers. Measurement helps us describe this variability in human characteristics.

Remember that test scores are imperfect and statistics help us determine the amount of error in test scores. Yet conclusions based on statistical analysis of test scores can never be absolute. Statistics tell us nothing about how the scores were obtained, what the scores mean, what effect the testing conditions had, or how motivated the child was. Other kinds of information, obtained through observation and test interpretation, can shed light on these questions. Still, measurement enables us to compare and contrast many psychological phenomena.

Measurement is a process of assigning quantitative values to objects or events according to certain rules. In physical measurement, the use of a ruler or a scale ensures that everyone follows agreed-on rules in measuring the length or weight of an object. In psychological measurement, a formal test, a rating scale, and/or a human observer plays a role similar to that played by a physical instrument. For example, after observing a child on the playground for 10 minutes, a human observer might use a five-point rating scale (e.g., from 1 = very uncooperative to 5 = very cooperative) to rate the child's level of cooperativeness. Although the human observer is following a rule to measure behavior, he or she must estimate variables without the help of a physical instrument.

SCALES OF MEASUREMENT

A *scale* is a system for assigning values or scores to some measurable trait or characteristic. The four most common scales—nominal, ordinal, interval, and ratio scales—are described below. Nominal and ordinal scales (referred to as *lower-order scales*) are used with *discrete variables.* Discrete variables are characterized by separate, indivisible categories, with no intermediate values (e.g., gender, color, or number of children in a family). Statistics known as nonparametric statistics, such as chi square and phi coefficient, are used to analyze the data obtained from nominal and ordinal scales. Interval and ratio scales (referred to as *higher-order scales*) are used with *continuous variables.* Continuous variables are characterized by an infinite number of possible values of the variable being measured (e.g., temperature, age, or height). Interval and ratio scales possess all the properties of nominal and ordinal scales but have additional properties (see Table 4-1). Parametric statistics, such as the *t* test and Pearson's product-moment correlation (*r*), are used to analyze the data obtained from interval and ratio scales.

Nominal Measurement Scale

At the lowest level of measurement is a *nominal measurement scale. Nominal* means "name." A nominal measurement scale consists of a set of categories that do not have a sequential order and that are identified by a name, number, or letter for each item being scaled. The names, numbers, or letters usually represent mutually exclusive categories, which cannot be arranged in a meaningful order and are merely labels or classifications. An example of nominal scaling is the assigning of numbers to baseball players (the numbers do not reflect the players' abilities) or the assigning of names or numbers to schools. Although nominal scales are of limited

Table 4-1
Properties of Scales of Measurement

	Property					
Scale	*Classification*	*Order*	*Equal intervals*	*True zero*	*Arithmetical operations possible*	*Examples of variables*
Nominal	X	—	—	—	None possible; scale useful only for classification	Gender, ethnicity, marital status
Ordinal	X	X	—	—	Greater than or less than operations	SES, movie ratings, intelligence test scores
Interval	X	X	X	—	Addition and subtraction of scale values	Temperature, sea level
Ratio	X	X	X	X	Multiplication and division of scale values	Height, weight, age, length

Note. Scores on intelligence tests are often considered to be on an interval scale, but in fact they are on an ordinal scale.

usefulness because they allow only for classification, they are still valuable. Some variables, such as gender, ethnicity, and geographic area, can be described only by nominal scales.

Ordinal Measurement Scale

At the next level of measurement is an *ordinal measurement scale.* Like a nominal measurement scale, an ordinal measurement scale classifies items, but it has the additional property of order (or magnitude). The variable being measured is ranked or ordered along some dimension, without regard for the distances between scores. One example of ordinal scaling is the ranking of students from highest to lowest, based on class standing. An ordinal scale tells us who is first, second, and third; it does not tell us, however, whether the distance between the first- and second-ranked scores is the same as the distance between the second- and third-ranked scores or the nineteenth- and twentieth-ranked scores. The difference between the first- and second-ranked grade point averages could be .10 (e.g., 3.30 versus 3.20), and the difference between the nineteenth- and twentieth-ranked grade point averages could be .20 (e.g., 2.00 versus 1.80). Another variable that can be measured using an ordinal scale is socioeconomic status (SES). For example, 1 could represent the lowest income level and 7 the highest income level. A third type of ordinal scale is a Likert rating scale, such as

No Anxiety	Mild Anxiety	Moderate Anxiety	Severe Anxiety	Extreme Anxiety
1	2	3	4	5

One cannot assume that a one-point increase in anxiety anywhere along this five-point scale equals a one-point increase anywhere else on the scale. Finally, standardized intelligence test scores that are designed to follow a normal distribution (see the discussion later in the chapter), such as the Wechsler Intelligence Scale for Children, Fourth Edition (WISC–IV), use ordinal measurement scales, even though intelligence test scores are often said to use interval measurement scales (Thomas, 1982). For example, a 15-point increase in IQ score from 100 to 115 may not mean the same thing as a 15-point increase from 115 to 130.

Interval Measurement Scale

At the third level of measurement is an *interval measurement scale.* It classifies, as a nominal scale does, and orders, as an ordinal scale does, but it adds an arbitrary zero point and equal units between points. An example of an interval measurement scale is the Fahrenheit scale, which measures temperature. On the Fahrenheit scale, the interval between 10°F and 20°F is the same as the interval between 60°F and 70°F. However, the zero point on such a scale is arbitrary, because a temperature reading of 0°F does not mean a complete lack of temperature. In addition, there are numbers below zero (e.g., –10°F) as well as above zero.

Ratio Measurement Scale

At the highest level of measurement is a *ratio measurement scale.* It has a true zero point, has equal intervals between adjacent units, and allows ordering and classification. Because there is a meaningful zero point, there is true equality of ratios between measurements made on a ratio scale. Weight is one example of a characteristic measured on a ratio scale; someone who weighs 150 pounds is twice as heavy as someone who weighs 75 pounds. Like weight, reaction time is measured on a ratio scale with a true zero point and equal ratios; a reaction time of 2,000 milliseconds is exactly twice as long as one of 1,000 milliseconds. Ratio scales are rarely used in psychology, because most psychological characteristics do not have an absolute zero point. Often we must be content with interval scales or the statistically weaker ordinal and nominal scales.

Table 4-2
Common Statistical and Psychometric Symbols and Abbreviations

Symbol	Definition
a	Intercept constant in a regression equation
b	Slope constant in a regression equation
c	Any unspecified constant
CA	Chronological age
cf	Cumulative frequency
DQ	Developmental quotient
f	Frequency
F	Test statistic in analysis of variance or covariance
IQ	Intelligence quotient
M	Mean (see also $\bar{X}$)
MA	Mental age
Mdn or *Md*	Median
n	Number of cases in a subsample
N	Number of cases in a sample
p	Probability or proportion
P	Percentile
Q	Semi-interquartile range; half the difference between Q_3 and Q_1
Q_1	First quartile score (25th percentile score)
Q_3	Third quartile score (75th percentile score)
r	Pearson correlation coefficient
r^2	Coefficient of determination; the proportion of variance in Y attributable to X
r_{pb}	Point biserial correlation coefficient
r_s or ρ	Spearman rank-difference correlation coefficient (also referred to as rho)
r_{xx}	Reliability coefficient
r_{xy}	Validity coefficient (x represents the test score and y the criterion score)
R	Coefficient of multiple correlation
rel. f	Relative frequency
S, s, or SD	Standard deviation of the sample
S^2	Variance of the sample
SE_E, SE_{est}	Standard error of estimate
SEM, SE_m, SE_{meas}, S_m, s_m, s_{meas}, or s_{err}	Standard error of measurement
t	t test
T	T score; standard score with a mean of 50 and standard deviation of 10
x	Deviation score ($X - \bar{X}$); indicates how far a particular score falls above or below the mean of the group
X	Raw score
$\bar{X}$	Mean (see also M)
Y	A second raw score
z	z score; standard score with a mean of 0 and standard deviation of 1
σ	Standard deviation of a population
σ^2	Variance of a population
Σ	"Sum of"
ΣX	"Sum of X"; ΣX means to add up all the Xs (scores)
ΣX^2	Sum of squared Xs (square first, then add)
$(\Sigma X)^2$	Squared sum of Xs (add first, then square the total)
ΣXY	Sum of cross products of X and Y (multiply each $X \times Y$, then add)
ϕ	Phi coefficient; a correlation coefficient for a 2 × 2 contingency table
χ^2	Chi square
<	Less than
>	Greater than
≥	Greater than or equal to
≤	Less than or equal to
±	Plus or minus
$\sqrt{\ }$	Square root
≠	Not equal to

DESCRIPTIVE STATISTICS

Descriptive statistics summarize data obtained about a sample of individuals. Examples of descriptive statistics are frequency distributions, normal curves, standard scores, measures of central tendency, and measures of dispersion, correlation, and regression. Some descriptive statistics are covered below; others are discussed later in the chapter.

Table 4-2 shows symbols and abbreviations commonly used in statistics and psychometrics. These symbols are shorthand for important characteristics of a test or norm group. (The list is for reference; it is not necessary to memorize the symbols.) As you gain experience in the field, the symbols will become more familiar to you.

Measures of Central Tendency

Measures of central tendency identify a single score that best describes the scores in a data set. The three most commonly used measures of central tendency are the mean, the median, and the mode. These statistics describe the average, the middle, and the most frequent score(s) of a set of scores, respectively.

Mean. The *mean* (M or $\bar{X}$) is the arithmetic average of all the scores in a set of scores. To compute the mean, divide the sum of all the scores by the total number of scores in the set (N). The formula is

$$M = \frac{\Sigma X}{N}$$

where M = mean of the scores
ΣX = sum of the scores
N = number of scores

Example: The mean for the four scores 2, 4, 6, and 8 is 5:

$$M = \frac{2+4+6+8}{4} = \frac{20}{4} = 5$$

The mean depends on the exact position of each score in a distribution, including extreme scores. However, it may not be the best measure of central tendency if there are too many scores that deviate extremely from the other scores in the set (such extreme scores are referred to as *outliers*). For example, three people with incomes of $30,000, $40,000, and $2,000,000 have an average income of $900,000. Yet it is unlikely that any of them have anywhere near a $900,000 life style. When there are few outliers in a distribution of scores, the mean is the preferred measure of central tendency. It can be calculated for both interval and ratio scale data.

Median. The *median* (*Mdn* or *Md*) is the middle point in a set of scores arranged in order of magnitude. Fifty percent of the scores lie at or above the median, and 50% of the scores lie at or below the median. If there are an even number of scores, the median is the number halfway between the two middlemost scores and, therefore, may not be any of the actual scores, unless the two middlemost scores are the same. If there are an odd number of scores, the median is simply the middlemost score.

To compute the median, arrange the scores in order of magnitude from highest to lowest. Then count up (or down) through half the scores. Table 4-3 illustrates the procedure for calculating the median of an even number and an odd number of scores in a distribution. In the first column, there are eight scores. To obtain the median, count up four scores from the bottom and then calculate the number halfway between the fourth and fifth scores (the two middlemost scores). In the second column, there are seven scores. To obtain the median, count up four scores from the bottom; the median is the fourth score. The median divides a distribution into two equal halves; the number of scores above the median is the same as the number below.

Table 4-3
Calculation of the Median

X *(even number of scores)*	*X* *(odd number of scores)*
130	130
128	128
125	125
124	124 ← 124 median
← 123.5 median	
123	123
120	120
110	110
108	

When a distribution is "skewed" (i.e., most of the scores are at either the high end or the low end of the set), the median is a better measure of central tendency than the mean. The median is not affected disproportionately by outliers and is an appropriate measure of central tendency for ordinal, interval, or ratio scale data. Suppose we wished to compare salaries at Harvard University with those at the University of Minnesota. The median salary would be a better single measure of the salaries of all employees at a university than the mean, because the salaries include those of professors, janitors, and all others.

Mode. The *mode* is the score that occurs most frequently in a set of scores. If there is only one score that occurs most frequently, we say the distribution is *unimodal.* If two scores occur with the same frequency and more often than any other score, we say that the distribution is *bimodal*—there are two modes in the set. When more than two scores occur with the same frequency and more frequently than any other score, we say that the distribution is *multimodal*—there are multiple modes in the set.

The mode tells us what score is most likely to occur and is therefore useful in analyzing nominal scale data (e.g., "What was the most frequently occurring classification in the group?"). However, it is greatly affected by chance and has little or no mathematical usefulness.

Measures of Dispersion

Dispersion refers to the variability of scores in a set or distribution of scores. The three most commonly used measures of dispersion are the range, the variance, and the standard deviation.

Range. The *range* is the difference (or distance) between the highest and lowest scores in a set; it is the simplest measure of dispersion. To compute the range, subtract the lowest score in the set from the highest score. The formula is

$$R = H - L$$

where R = range
H = highest score
L = lowest score

Example: The range for the distribution 50, 80, 97, and 99 is 49:

$$R = 99 - 50 = 49$$

The range is easily calculated; however, it is an insensitive measure of dispersion because it is determined by the locations of only two scores. The range tells us nothing about the distribution of scores located between the high and low scores, and a single score can markedly increase the range. Still, the range can be useful. It provides a preliminary review of a distribution and a gross measure of the spread of scores.

Variance. The *variance* (S^2) is a measure of the amount of variability of scores around the mean—the greater the variability, the greater the variance. Unlike the range, the variance takes into account every score in a group. When two different sets of scores have the same mean but different variances, it means that the scores in one set are more widely dispersed than the scores in the other. The variance is obtained by comparing every score in a distribution to the mean of the distribution. The variance is the average squared deviation of scores from the mean. To compute the deviation of an individual score (i.e., how far an individual score is from the mean of the group), subtract the mean from that score. Scores that have values greater than the mean will yield positive values, whereas scores that have values less than the mean will yield negative values. To compute the variance of a sample, use the following formula:

$$S^2 = \frac{\Sigma(X - \bar{X})^2}{N - 1}$$

where S^2 = variance of the scores
Σ = sum
X = raw score
$\bar{X}$ = mean
N = number of scores

Example: The variance for the four scores 2, 4, 6, and 8 is 6.67:

$$S^2 = \frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4-1}$$

$$= \frac{9+1+1+9}{4} = \frac{20}{3} = 6.67$$

Squaring the distance from the mean has two important benefits: It makes all the variances positive so that they can be summed (rather than canceling each other out), and it gives greater weight to values farther from the mean and thereby signals the accuracy and precision of the mean (i.e., how far scores fall from their central indicator). This is a quality captured by the standard error of measurement, a concept discussed later in the chapter.

Standard deviation. The *standard deviation* (*SD*, *S*, or *s*) is also a measure of how much scores vary, or deviate, from the mean. It is the square root of the variance, representing the average distance of the data values from the mean. The standard deviation is always a positive number (or zero) and is measured in the same units as the original data. The standard deviation is often used in the field of testing and measurement. To compute the standard deviation of a sample, use the following formula:

$$SD = \sqrt{\frac{\Sigma(X - \bar{X})^2}{N - 1}}$$

Example: The standard deviation for the four scores 2, 4, 6, and 8 is 2.58:

$$S^2 = \frac{(2-5)^2 + (4-5)^2 + (6-5)^2 + (8-5)^2}{4-1}$$

$$= \frac{9+1+1+9}{4} = \frac{20}{3} = \sqrt{6.67} = 2.58$$

Normal Curve

The *normal curve* is a frequency distribution that, when graphed, forms a bell-shaped curve (see Figure 4-1). It is also called a *Gaussian distribution,* after Carl Friedrich Gauss, who developed it in 1809 (see Figure 4-2). Many human characteristics—such as height, weight, intelligence, and personality traits—have normal distributions. You can often assume that human characteristics follow a normal curve, even though the characteristics do not always fit the curve perfectly.

Let's look at some features of the normal curve. First, the normal curve is a symmetrical distribution of scores with an equal number of scores above (to the right of) and below (to the left of) the midpoint of the curve. Second, there are more scores close to the middle of the distribution than at the ends of the distribution. Third, the mean, median, and mode of a normal curve are the same. Fourth, specific percentages of scores fall at precise distances (measured in standard deviation units) from the mean. This enables us to calculate exactly how many cases fall between any two points under the normal curve (see below). Finally, tables in statistics books present the proportion of scores above and below any point on the *abscissa* (i.e., the value of a coordinate on the horizontal, or *X*, axis), expressed in standard deviation units.

Figure 4-1 shows the precise relationship between the standard deviation and the proportion of cases under a normal curve. It also shows the percentages of cases that fall within one, two, and three standard deviations above and below the mean. In a distribution of scores that follows a normal curve, approximately 68% of the cases fall within +1 *SD* and –1 *SD* of the mean (approximately 34% of the cases are between the mean and 1 *SD* above the mean, and approximately 34% of the cases are between the mean and 1 *SD* below the mean). As we move away from the mean, the number of cases diminishes. The areas between +1 *SD* and +2 *SD* and between –1 *SD* and –2 *SD* each contain approximately 14% (13.59%) of the cases. Between +2 *SD* and +3 *SD* and between –2 *SD* and –3 *SD*, there are even fewer cases—each area represents approximately 2% (2.14%) of the cases. The areas beyond +3 *SD* and –3 *SD* represent only .13% of the cases.

These percentages are also useful because the scores along the abscissa can be translated into percentile ranks (discussed

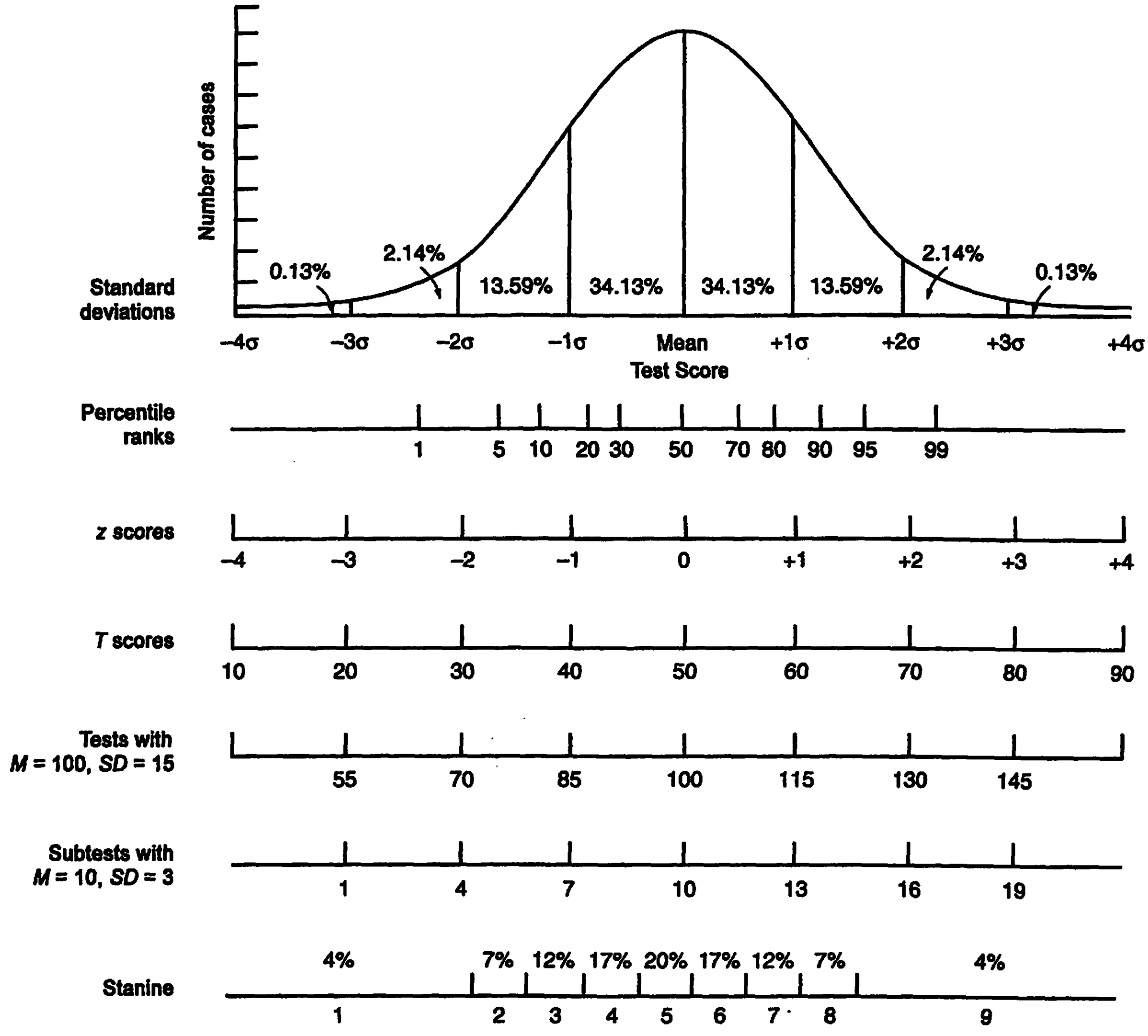


Figure 4-1. Relationship of the normal curve to various types of standard scores.

later in the chapter). Thus, a score of 115 in a distribution with $M = 100$ and $SD = 15$ represents the 84th percentile rank. And a score of 85 represents the 16th percentile rank. A score of 115 is +1 *SD* above the mean, while a score of 85 is −1 *SD* below the mean. Other percentile ranks can be computed in a similar manner. Table BC-1 on the inside back cover gives the percentile ranks associated with standard scores in a distribution with $M = 100$ and $SD = 15$. We will return to the normal curve when we consider standard scores.

CORRELATION

Correlation coefficients (*r*) tell us about the degree of relationship between two variables, including the strength and direction of their relationship. The strength of the relationship is expressed by the absolute magnitude of the correlation coefficient. The sign of the coefficient reflects the direction of the relationship. A positive correlation (+) indicates that higher scores on one variable are associated with higher scores on the second variable (e.g., more hours spent studying are associated with a higher GPA) and thus that lower scores on one variable are associated with lower scores on the second variable (e.g., fewer hours spent studying are associated with a lower GPA). Conversely, a negative correlation (−) signifies an inverse relationship—that is, high scores on one variable are associated with low scores on the other variable (e.g., a large number of days absent tends to be associated with a low GPA). Correlation coefficients range in value from −1.00 to +1.00.

Correlations are used in prediction. The higher the correlation between two variables, the more accurately we can predict the value of one variable when we know the value of the other variable. A correlation of +1.00 (or −1.00) means that

Figure 4-2. Gauss, a great mathematician, honored by Germany on their 10 Deutsche mark bill (note the normal curve to the left of his picture).

we can perfectly predict a person's score on one variable if we know the person's score on the other variable (e.g., weight in pounds perfectly predicts weight in kilograms). In contrast, a correlation of .00 indicates that knowing the score on one variable does not help at all in predicting the score on the other variable (e.g., comparing weight and annual income). Finally, a correlation of .50 indicates that knowing the score on one variable partially predicts the score on the other variable (e.g., comparing IQ and GPA).

It is important to distinguish between the strength of the correlation and the direction of the correlation. A correlation above .50, either negative or positive, indicates a moderate to strong relationship between the two variables. When we consider only the *strength* of the relationship, it doesn't matter whether the correlation is positive or negative (e.g., whether $r = +.50$ or $r = -.50$). However, we also need to know the *direction* of the relationship between the scores—that is, whether it is positive or negative.

Variables can be related linearly or curvilinearly. A *linear relationship* between two variables can be portrayed by a straight line. A *curvilinear relationship* between two variables can be portrayed by a curve. If two variables have a curvilinear relationship, a linear correlation coefficient will underestimate the true degree of association.

Variables can also be continuous or discrete. A continuous variable is divisible into an infinite number of parts (e.g., temperature, height, age). In contrast, a discrete variable has separate, indivisible categories (e.g., the number of heads in a series of coin tosses). A *dichotomous variable* is a discrete variable that has two possible values (e.g., head or tail, pass or fail, male or female). As discussed earlier in the chapter, the scale of measurement used will depend on whether the variables being measured are continuous or discrete. Essentially, variables must be continuous in order for ratio and interval scales of measurement to be used; ordinal and nominal scales of measurement must be used with discrete variables.

Figure 4-3 shows scatterplots (plots of individual scores on a graph) of eight different relationships. A *scatterplot* presents a visual picture of the relationship between two variables. Each point in a scatterplot represents a pair of scores for one individual on two different variables (e.g., height and weight). That is, a data point represents a single score on the X variable and a single score on the Y variable.

Graph (a) in Figure 4-3 shows a perfect positive linear relationship between X and Y ($r = +1.00$); the dots fall in a straight line from the lower left (low X, low Y) to the upper right (high X, high Y). Graph (b) shows a perfect negative linear relationship ($r = -1.00$); the dots fall in a straight line from the upper left (low X, high Y) to the lower right (high X, low Y). Graphs (c) through (f) show varying degrees of relationship between X and Y. Graph (g) shows a totally random relationship (i.e., no relationship) between X and Y ($r = .00$). And graph (h) shows a nearly perfect curvilinear relationship between X and Y; the dots fall along a curved line.

The most common correlation coefficient is the *Pearson correlation coefficient,* symbolized by *r.* Pearson's r should be used only when the following conditions are met: (a) The two variables are continuous and normally distributed, (b) there is a linear relationship between the variables, and (c) the predictor variable predicts as well at the high-score ranges as at the low-score ranges. Note that the Pearson correlation coefficient, which is calculated on the assumption that two variables are linearly related, would incorrectly indicate that there was no relationship between the two variables shown in graph (h) of Figure 4-3. When the conditions for using Pearson's r cannot be met (e.g., the data are ordinal), the *Spearman* r_s (rank-difference) method can be used (see Table 4-4). This method uses the ranks of the scores instead of the scores themselves. A rank is a number given to a score to represent its order in a distribution. For example, in a set of 10 scores, the highest score receives a rank of 1, the fifth score from the top receives a rank of 5, and the lowest score receives a rank of 10.

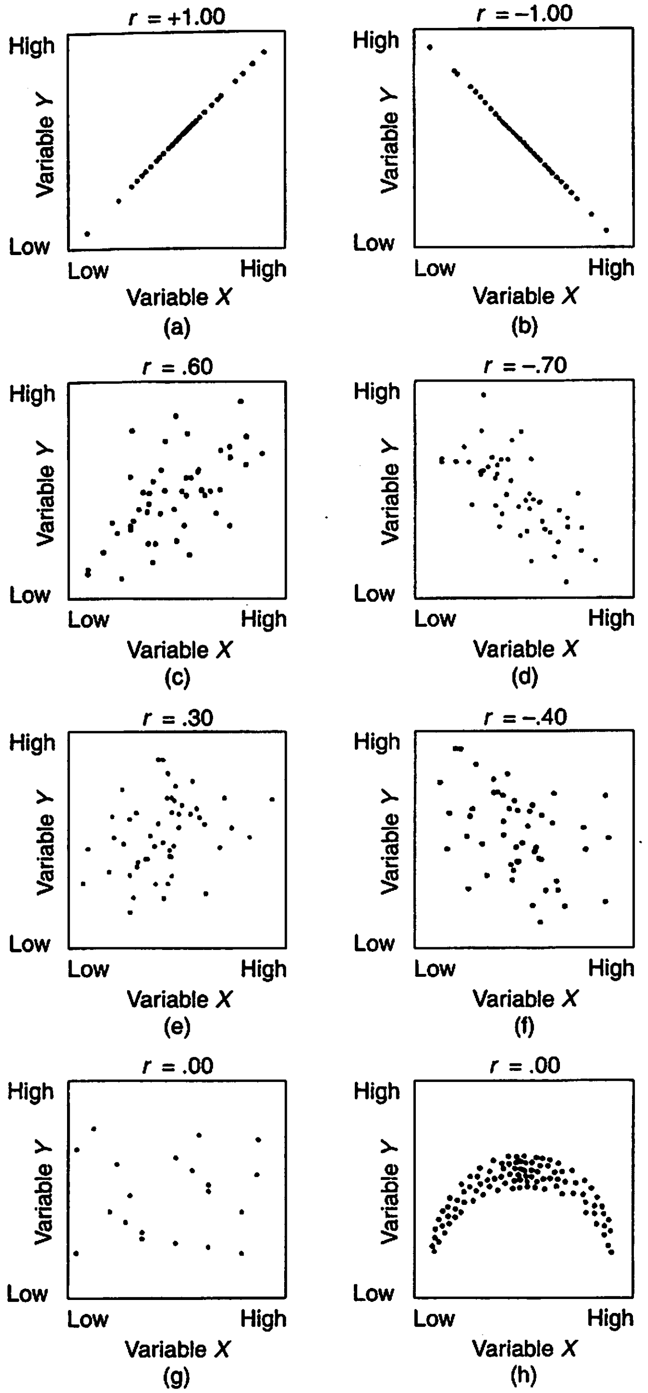


Figure 4-3. Scatter diagrams illustrating various degrees of relationship.

The following are useful terms to describe the strength of a correlation:

- .20 to .29: low
- .30 to .49: moderately low
- .50 to .69: moderate
- .70 to .79: moderately high
- .80 to .99: high

When the sample size is large, a correlation coefficient may be statistically significant but reflect only a weak association between the two variables. For example, a Pearson correlation coefficient of .20 may be significant when the sample size is 100, but the level of variance explained is low (.20 = 4%). In contrast, a Pearson correlation of .70 may not be significant when the sample size is small, but the level of variance explained is high (.70 = 49%). Correlations also can be lower when there is a restriction of range—that is, when scores are very close to each other (e.g., 20, 21, 22, 24, 26, as opposed to 4, 6, 8, 22, 25, 30) and thus have less variability—or when there is a large amount of measurement error. (We will discuss measurement error in a later section.) Outliers are scores that are extreme, atypical, and infrequent and that unduly influence the size and direction of the correlation coefficient (i.e., such scores markedly increase or decrease the size of the correlation coefficient and its direction, either positive or negative). A single outlier can have a powerful effect on the correlation coefficient when the sample size is small.

Sometimes test publishers (or researchers) attempt to minimize the effect of measurement error by *correcting for attenuation.* This correction results in an estimate of what the correlation between two variables would be if both variables were perfectly reliable. However, an estimated r based on a correction for attenuation may not give a true picture of the relationship between the variables (e.g., it may inflate the relationship), because variables are never perfectly reliable.

Correlations should not be used to infer cause and effect. For example, although there is a correlation between hot, wet climates and the occurrence of malaria, climate is not the cause of malaria; the relationship between hot climates and malaria is only an indirect one. For a long time, people believed that "bad air" caused malaria. (The ancient Romans named the disease for this reason: *Mal aria* means "bad air" in Latin.) We now know that the disease is actually carried by mosquitoes, which flourish in stagnant water in hot climates.

When we want to know how much variance in one variable is explained by its relationship to another variable, we must square the correlation coefficient. The resulting value,

Table 4-4
Formulas for Computing a Variety of Correlation Coefficients

Name	*Description of variables*	*Formula*
Pearson product-moment correlation coefficient (r)	Both variables continuous (on interval or ratio scale)	$r = \frac{N\,\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\,\Sigma X^2 - (\Sigma X)^2][N\,\Sigma Y^2 - (\Sigma Y)^2]}}$ where r = correlation coefficient N = number of paired scores ΣXY = sum of the products of the paired X and Y scores ΣX = sum of the X scores ΣY = sum of the Y scores ΣX^2 = sum of the squared X scores $(\Sigma X)^2$ = square of the sum of the X scores ΣY^2 = sum of the squared Y scores $(\Sigma Y)^2$ = square of the sum of the Y scores
Spearman rank-difference correlation coefficient (Spearman r, r_s, or ρ)	Both variables on an ordinal scale (rank-ordered)	$r_s = 1 - \frac{6\Sigma D^2}{N(N^2 - 1)}$ where D = difference between ranks for each person N = number of paired scores
Point biserial correlation coefficient (r_{pb})	One variable continuous (on interval or ratio scale), the other genuinely dichotomous (usually on nominal scale)	Formula for r can be used (see above). The dichotomous variable can be coded 0 or 1. For example, if sex is the dichotomous variable, 0 can be used for females and 1 for males (0 = females, 1 = males), or vice versa.
Phi (ϕ) coefficient	Both variables dichotomous (on nominal scales)	1. $\phi = \frac{BD - AD}{\sqrt{(A + B)(C + D)(A + C)(B + D)}}$ where A, B, C, and D are the four cell frequencies in a contingency table 2. $\phi = \sqrt{\frac{\chi^2}{N}}$ where χ^2 = chi square N = total number of observations

r^2, is known as the *coefficient of determination.* For example, if we want to know how much variance in school grades is accounted for by knowing the scores on a measure of intelligence, we first compute a correlation coefficient for the two measures. Let's say r = .60. Squaring r gives .36, or 36%. Consequently, we can say that knowing the scores on the measure of intelligence allows us to account for 36% of the variance in school grades. This value may not seem large, but given that other factors (such as the student's motivation, effort, and previous instruction in various subject areas) account for some of the variance in school grades as well, a score on a measure of intelligence is a significant predictor of academic achievement. However, like a correlation coefficient, the coefficient of determination only describes an association between two variables. It does not establish a cause-and-effect relationship between the two variables.

REGRESSION

Regression Equation

You can use the correlation coefficient, together with other information, to construct a linear equation for predicting the score on one variable when you know the score on another variable. A *linear equation* describes a linear relationship between variables, as discussed earlier in the chapter. This type of relationship can be represented on a graph by a straight line that fits all of the scores in that graph. This equation, called the *regression equation,* has the following form:

$$Y_{\text{pred}} = bX + a$$

where Y_{pred} = predicted score on Y
b = slope of the regression line

X = known score on X
a = Y intercept of the regression line

The slope of the regression line, b, is defined as

$$b = r\frac{SD_Y}{SD_X}$$

where r = Pearson correlation between the X and Y scores
SD_Y = standard deviation of the Y scores
SD_X = standard deviation of the X scores

The formula for calculating b directly from raw data is

$$b = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{N\Sigma X^2 - (\Sigma X)^2}$$

where N = number of paired scores

The intercept a, or regression constant, is determined as follows:

$$a = \bar{Y} - b\bar{X}$$

where $\bar{Y}$ = mean of the Y scores
b = slope of the regression line
$\bar{X}$ = mean of the X scores

Example: To find the regression equation and correlation coefficient for the following pairs of scores (X, Y), we first calculate $X^{2,}$ $Y^{2,}$ and XY.

	X	Y	X^2	Y^2	XY
	7	9	49	81	63
	2	3	4	9	6
	6	4	36	16	24
	6	5	36	25	30
	3	1	9	1	3
Σ	24	22	134	132	126

$$\bar{X} = 4.80 \qquad \bar{Y} = 4.40$$

The slope of the regression line is then given by

$$b = \frac{5(126) - 24(22)}{5(134) - (24)^2} = \frac{630 - 528}{670 - 576} = \frac{102}{94} = 1.09$$

and the regression constant is given by

$$a = 4.40 - 1.09(4.80) = 4.40 - 5.23 = -.83$$

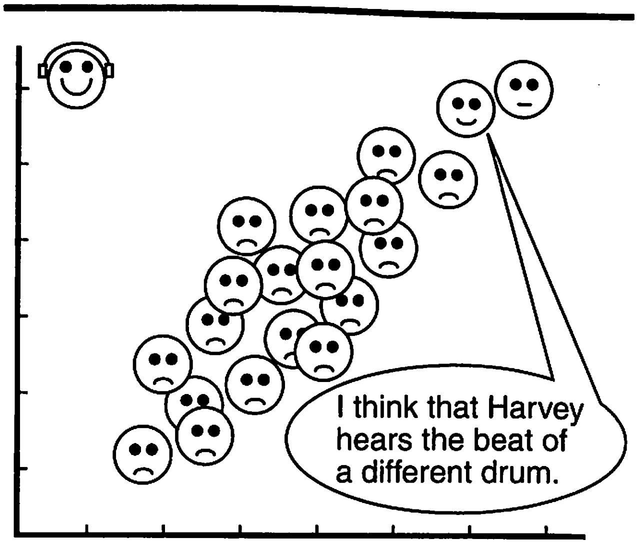
These values can now be substituted into the regression equation:

$$Y_{\text{pred}} = 1.09X - .83$$

The Pearson correlation coefficient (see Table 4-4 for formula) for these data is

$$r = \frac{5(126) - 24(22)}{\sqrt{[5(134) - (24)^2][5(132) - (22)^2]}}$$

$$= \frac{102}{\sqrt{94(176)}} = \frac{102}{\sqrt{16{,}544}} = \frac{102}{128.62} = .79$$



Scatterplot: n = 21; r = +0.63

The Outlier

Courtesy of David Likely.

Standard Error of Estimate

A measure of the accuracy of the predicted Y scores in a regression equation is the standard error of estimate:

$$SE_{\text{est}} = SD_Y\sqrt{1 - r_{XY}^2}$$

where SD_Y = standard deviation of the Y scores
r_{XY}^2 = square of the correlation between the X and Y scores

The *standard error of estimate* is the standard deviation of the error scores, a measure of the amount by which the observed or obtained scores in a sample differ from the predicted scores. The higher the correlation between X and Y, the smaller the standard error of estimate and, hence, the greater the average accuracy of the predictions. When you have a perfect correlation between scores (that is, $r = +1.00$), the standard error of estimate becomes zero, as you can see by substituting 1.00 for r in the above equation. Thus, a +1.00 correlation coefficient means that you can make perfect predictions of Y if you know X. A .00 correlation means that knowledge of X does not improve your prediction of Y. In this case, the standard error of estimate is exactly the same as the standard deviation of the Y scores, and the best you can do is simply to guess that each Y score falls at the mean of the score distribution.

Example: The standard error of estimate for a test with a standard deviation of 15 and a .60 correlation between X and Y is

$$\begin{aligned} SE_{\text{est}} &= 15\sqrt{1 - (.60)^2} \\ &= 15\sqrt{1 - .36} \\ &= 15(.80) = 12 \end{aligned}$$

This means that we can say at a 68% confidence level that the predicted score Y will be within ±12 of the actual value of Y. We will return to this measure again when we discuss confidence levels.

MULTIPLE CORRELATION

Multiple correlation is a statistical technique for determining the relationship between one variable and two or more other variables. An example is predicting a student's GPA based on his or her IQ plus the average number of hours spent daily on homework. The symbol for the coefficient of multiple correlation is R, and its values range from .00 to +1.00. When we use several variables for a prediction, the prediction is likely to be more accurate and powerful than if we based it on a single variable only. A principal drawback to using multiple correlation, however, is that large samples are generally required when several variables are used in the analysis—usually over 100 individuals or at least 20 individuals per variable. Thus, if 10 variables were being studied, we would need 200 individuals to arrive at a stable prediction equation.

One example of the use of multiple correlation is in the prediction of college performance. High school grades, intelligence test scores, and educational attainment of parents are measures that correlate positively with performance in college. Another example is in the prediction of success in counseling. Personality test scores, teacher ratings of behavior pathology, and intelligence test scores correlate with successful outcomes. By using these measures in a multiple correlation, we can predict the outcome of academic performance or therapy with more accuracy than by using any individual measure alone.

NORM-REFERENCED MEASUREMENT

In *norm-referenced measurement,* a child's performance on a test is compared with the performance of a representative group of children, referred to as a *norm group* or a *standardization sample.* Norms are needed because the number of correct responses the child makes is not very meaningful in itself. For example, knowing that a child obtained a raw score of 21 on a 30-item test (i.e., answered 70% of the items correctly) is of little use unless we also know how other children performed on the same test; we need a relevant normative population. We could compare the child's score with scores from a representative population of children in the United States, with scores from children in the child's school, or with scores from a special population. Such comparisons are made by converting the child's raw score into some relative measure, called a derived score. A *derived score* indicates the child's standing relative to the norm group and allows us to compare the child's performance on one measure with his or her performance on other measures. Norm-referenced tests are also called "standardized tests," because they require standardized administration and scoring procedures and the scores are transformed, or "standardized," relative to the norm group.

Four concepts related to norm-referenced measurement are population, representative sample, random sample, and reference group. The *population* is the complete group or set of cases. A *representative sample* is a group drawn from the population that represents the population accurately. A *random sample* is a sample obtained by selecting members of the population based on random selection (such as the flip of a coin) so that each person in the population has an equal chance of being selected. And the *reference group* is the norm group that serves as the comparison group for computing standard scores, percentile ranks, and related statistics.

Representativeness

The *representativeness* of a norm group reflects the extent to which the group's characteristics match those of the population of interest. For psychological and psychoeducational assessment, the most prominent of these characteristics are typically age, grade level, gender, geographic region, ethnicity, and socioeconomic status (SES). SES is usually determined by ascertaining the educational attainment and/or occupational level of the client or of the client's parents if the client is a child. We also need to know when the norms were established in order to determine whether the norms are still relevant.

"I could have done better, but I didn't want to depart too far from the accepted norm."

Courtesy of Germaine Vanselow, Cartoonist, Bill Vanselow.

Size

A norm group should be large enough to ensure that the test scores are stable and representative of the population—that is, that the subgroups in the population are adequately represented. Usually, the larger the number of individuals in the norm group, the more stable and representative the norms. If a test is going to be used for several age groups, then ideally the sample should contain at least 100 individuals in each age group.

Relevance

To interpret the *relevance* of a child's scores properly, an examiner needs a reference group against which to evaluate the scores. For most assessment purposes, large nationally representative samples are preferred, because they provide stable and reliable scores against which to compare a child's test scores. If you use a reference group that is different from the customary one, clearly say so in your report.

DERIVED SCORES

The major types of derived scores used in norm-referenced measurement are standard scores, percentile ranks, normal-curve equivalents, stanines, age-equivalent scores, grade-equivalent scores, and ratio IQs. As the following discussion indicates, the various derived scores differ in their usefulness.

Standard Scores

Standard scores are raw scores that have been transformed so that they have a predetermined mean and standard deviation. They are expressed as an individual's distance from the mean in terms of the standard deviation of the distribution. Once transformed, a child's score can be expressed as a value on this standardized scale.

One type of standard score is a *z score,* which has $M = 0$ and $SD = 1$. Almost all z scores lie between –3.0 and +3.0. A z score of –2.5 would indicate that a raw score fell 2½ standard deviations below the mean. We frequently convert z scores to other standard scores to eliminate the + and – signs. For example, a T score is a standard score from a distribution with $M = 50$ and $SD = 10$. T scores almost always fall between 20 and 80; a z score of 0 is equivalent to a T score of 50.

Table 4-5 shows formulas for computing various standard scores. A general formula for converting standard scores from one system to another is

$$\text{New standard score} = \left(\frac{X_{\text{old}} - M_{\text{old}}}{SD_{\text{old}}}\right)SD_{\text{new}} + M_{\text{new}}$$

where X_{old} = score on old system
M_{old} = mean of old system
SD_{old} = standard deviation of old system
SD_{new} = standard deviation of new system
M_{new} = mean of new system

Example: A standard score of 60 in a T distribution ($M = 50$, $SD = 10$) is converted to a standard score in a distribution with $M = 100$ and $SD = 15$ as follows:

$$\begin{aligned}\text{New standard score} &= \left(\frac{60-50}{10}\right)15 + 100\\ &= \left(\frac{10}{10}\right)115 + 100 = (1)15 + 100 = 115\end{aligned}$$

Percentile Ranks

Percentile ranks are derived scores that permit us to determine an individual's position relative to the standardization sample or any other specified sample. A percentile rank is a point in a distribution at or below which the scores of a given percentage of individuals fall. If 63% of the scores fall at or below a given score, then that score is at the 63rd percentile rank. That is, a student at the 63rd percentile rank on a particular test performed as well as or better than 63% of the students in the norm group and not as well as the remaining 37% of the students. *Quartiles* are percentile ranks that divide a distribution into four equal parts, with each part containing 25% of the norm group. *Deciles,* a less common percentile rank, contain 10 bands, with each band containing 10% of the norm group. Exhibit 4-1 shows some procedures for calculating percentile ranks.

Interpretation of percentile ranks is straightforward. For example, a child with a percentile rank of 35 on a measure of memory has scored as high as or higher than 35% of the children in the norm sample. However, the psychometric properties of percentile ranks limit their usefulness in data analysis. A major problem with percentile ranks is that we can't assume that the units along the percentile-rank distribution are equal. Raw score differences between percentile ranks are smaller near the mean than at the extremes of the distribution. For example, the difference between a person at the 51st percentile rank and one at the 55th percentile rank may be very small. However, there are fewer cases at the extremes (people are more spread out), and so here small differences in percentile ranks (e.g., between the 95th and 99th percentile ranks) may be meaningful (see Figure 4-1). Percentile ranks cannot be added, subtracted, multiplied, or divided. In order to use them in statistical tests, you must normalize percentile ranks by converting them to another scale. Percentile ranks are often used in discussing results with parents, but you must always keep this problem of imprecise units in mind.

Normal-Curve Equivalents

Normal-curve equivalents (NCEs) are standard scores with $M = 50$ and $SD = 21.06$. NCEs divide the normal curve into 100 equal units (see Table BC-1 on the inside back cover). Unlike percentile ranks, which cannot be used for statistical

Table 4-5
Formulas for Computing Various Standard Scores

Score	Example
z score $z = \frac{X - \bar{X}}{SD}$ where z = z score corresponding to the individual raw score X X = individual raw score $\bar{X}$ = mean of sample SD = standard deviation of sample	The z score for an individual with a raw score of 50 in a group having a mean of 30 and standard deviation of 10 is calculated as follows: $z = \frac{50 - 30}{10} = 2$ Thus, the z score for this individual is 2.
***T* score** $T = 10(z) + 50$ where T = T score corresponding to the individual raw score X 10 = standard deviation of the T distribution z = z score corresponding to the individual raw score X 50 = mean of the T distribution	The T score for an individual with a z score of 2 is calculated as follows: $T = 10(2) + 50 = 70$ Thus, the T score for this individual is 70.
Standard score $SS = 15(z) + 100$ where SS = standard score corresponding to the individual raw score X 15 = standard deviation of the standard score distribution z = z score corresponding to the individual raw score X 100 = mean of the standard score distribution	The standard score for an individual with a z score of 2 is calculated as follows: $SS = 15(2) + 100 = 130$ Thus, the standard score for this individual is 130.

analyses, NCEs can be used for such purposes because they can legitimately be added, subtracted, multiplied, and divided.

Stanines

Stanines (a contraction of "standard nine") provide a single-digit scoring system with $M = 5$ and $SD = 2$. Stanine scores are expressed as whole numbers from 1 to 9. When we convert scores to stanines, the shape of the original distribution is converted into an approximately normal curve. The percentages of scores at each stanine are 4, 7, 12, 17, 20, 17, 12, 7, and 4, respectively (refer to Figure 4-1). Stanines have drawbacks, such as loss of information associated with large categories and categories that are not equal intervals.

Age-Equivalent Scores and Grade-Equivalent Scores

Age-equivalent scores are obtained by computing the average raw scores obtained on a test by children at different ages. (Other terms for age-equivalent scores are *test-age equivalent, test age,* and *mental age,* or *MA.*) For example, if the average raw score of a group of 10-year-old children on a test is 15 items correct out of 25, any child obtaining a raw score of 15 receives an age-equivalent score of 10-0 (10 years, 0 months). Similarly, *grade-equivalent scores* are obtained by computing the average raw scores obtained on a test by children in different grades. If the average score of seventh graders on an arithmetic test is 30, we say that a child with a score of 30 has arithmetical knowledge at the seventh-grade level (or a grade-equivalent score that equals the seventh-grade level).

Grade-equivalent scores are expressed in tenths of a grade (e.g., 5.5 refers to average performance of children at the middle of the fifth grade). This is in contrast to age-equivalent scores, which are expressed in years and months. A grade-equivalent score, therefore, refers specifically to the performance of an average student at that grade level on that test. It is important to note that the score does not mean that the performance of the student who achieved it is consistent with all curricular expectations for that grade level at his or her particular school. Note that a hyphen is usually used for age equivalents (e.g., 10-0) and a decimal for grade equivalents (e.g., 5.5).

Age-equivalent and grade-equivalent scores must be interpreted carefully, because they can be misleading for the following reasons:

1. Scores in age-equivalent or grade-equivalent distributions may not represent equal units. For example, the differ-

Exhibit 4-1
Calculating Percentile Ranks

The following formula is used to determine the percentile rank for a score in a distribution:

$$\text{Percentile rank} = \frac{\left(\frac{X - \text{lrl}}{i}\right)\text{fw} + \Sigma\text{fb}}{N} \times 100$$

where X = raw score
lrl = lower real limit of the target interval or score
i = width of the target interval or score
fw = frequency within the target interval or score
Σfb = sum of frequencies (number of scores occurring) below the target interval or score
N = total number of scores

To compute the lower real limit of a whole number, simply subtract .5 from the number; to get the upper real limit, add .5 to the number. The width of the target interval or score (i) is obtained by subtracting the lower real limit from the upper real limit.

Example 1
Let's compute the percentile rank for a score of 110 in the following distribution:

	X	f
	120	5
	119	10
Target interval for a score of 110 →	110	20
	100	40
	90	10
	80	5
	$N =$	90

where $\text{lrl} = 109.5$
$i = 1$
$\text{fw} = 20$
$\Sigma\text{fb} = 55$
$N = 90$

Substituting these values into the percentile rank formula yields the following:

$$\begin{aligned}\text{Percentile rank} &= \frac{\left(\frac{X - \text{lrl}}{i}\right)\text{fw} + \Sigma\text{fb}}{N} \times 100\\ &= \frac{\left(\frac{110 - 109.5}{1}\right)20 + 55}{90} \times 100\\ &= \frac{\left(\frac{.5}{1}\right)20 + 55}{90} \times 100\\ &= \frac{10 + 55}{90} \times 100\\ &= \frac{65}{90} \times 100\\ &= .72 \times 100\end{aligned}$$

The percentile rank is the 72nd percentile, or 72. Thus, a score of 110 exceeds 72% of the scores in the distribution.

The formula given here for calculating percentile rank can be used with both grouped (organized into classes of more than one value) and ungrouped (organized into classes of single values) data. When the distribution is ungrouped and all the intervals are 1, a simplified version of the formula can be used:

$$\text{Percentile rank} = \frac{.5\text{fw} + \Sigma\,\text{fb}}{N} \times 100$$

Example 2
Let us compute the percentile rank for a score of 4 in the following distribution:

	X	f
	5	3
Target interval for a score of 4 →	4	5
	3	4
	2	3
	1	2
		$N = 17$

where $\text{fw} = 5$
$\Sigma\text{fb} = 9$
$N = 17$

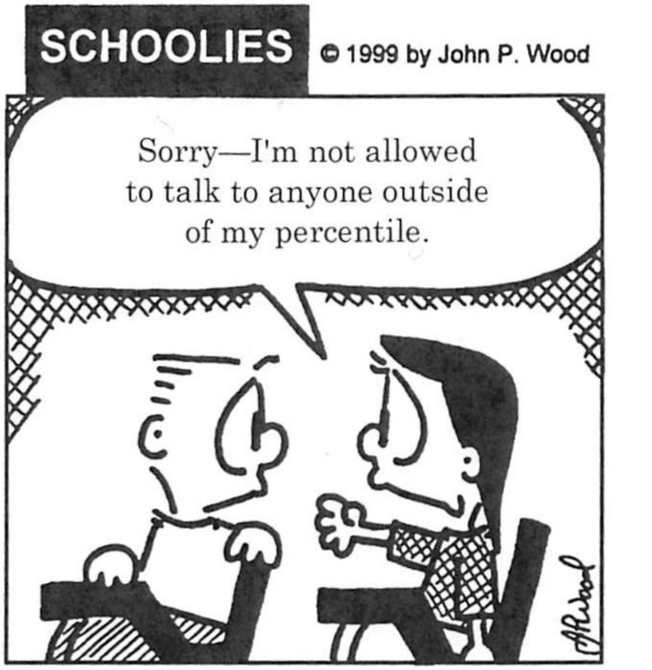
Substituting these values into the percentile rank formula for ungrouped data with intervals of 1 yields

$$\begin{aligned}\text{Percentile rank} &= \frac{.5\text{fw} + \Sigma\text{fb}}{N} \times 100\\ &= \frac{(.5)5 + 9}{17} \times 100\\ &= \frac{2.5 + 9}{17} \times 100\\ &= \frac{11.5}{17} \times 100\\ &= .68 \times 100\end{aligned}$$

The percentile rank is the 68th percentile, or 68. Thus, a score of 4 exceeds 68% of the scores in the distribution.

ence between second grade–equivalent and third grade–equivalent scores may not be the same as the difference between eleventh grade–equivalent and twelfth grade–equivalent scores. This happens because many skills (such as vocabulary and visual-motor skills) are acquired more rapidly at younger ages than at older ages.

2. Because many grade equivalents are obtained by interpolation (estimating a value between two given values or


SCHOOLIES
© 1999 by John P. Wood
Sorry—I'm not allowed to talk to anyone outside of my percentile.

Copyright © 1999 by John P. Wood. Reprinted with permission.

points) and extrapolation (extending norms to scores not actually obtained in the standardization sample), particular interpolated or extrapolated scores may not actually have been obtained by any children.

3. Grade equivalents sometimes encourage comparison with inappropriate groups. For example, we should not say that a second grader who obtains a grade equivalent of 4.1 in arithmetic is functioning in all ways like a fourth grader; fourth graders are the wrong comparison group. The second-grade student shares with the average fourth grader the number of items right on the test—not other attributes associated with fourth-grade mathematical skills. A grade equivalent of 4.1 on a specific test should be interpreted only in reference to the child's second-grade comparison group.
4. Identical grade-equivalent scores on different tests may mean different things. For example, grade-equivalent scores of 4.6 on two different tests of mathematics may mean that the child has mastered different mathematical content assessed by the two tests.
5. Expressing student performance in terms of grade equivalents could be seen as suggesting that growth is constant throughout the school year, an assumption that may not be warranted.
6. At junior and senior high school levels, age equivalents and grade equivalents may have little meaning for school subjects not taught at those levels or for skills that reach their peak at an earlier age.
7. Grade equivalents exaggerate small differences in performance; a score slightly below the median may result in a grade level equivalent one or two years lower.
8. Grade equivalents vary from test to test, from subtest to subtest within the same test, and from percentile to percentile, thereby complicating any type of comparison.
9. Grade-equivalent scores depend on promotion practices in different schools and on the particular curricula being used in different grades and in different schools.
10. Age-equivalent and grade-equivalent scores tend to be based on ordinal scales that cannot support the computation of important statistical measures, such as the standard error of measurement.

Age-equivalent and grade-equivalent scores are psychometrically impure; nevertheless, they may be useful when you discuss assessment findings. Age-equivalent and grade-equivalent scores place performance in a developmental context, and they provide information that consumers of the findings (e.g., parents and the public) can easily understand. If age-equivalent and grade-equivalent scores are used, consumers should be educated in their use. The Administration Manual of the WISC–IV and the Wechsler Preschool and Primary Scale of Intelligence–III (WPPSI–III), and the Examiner's Manual of the Stanford-Binet Intelligence Scale: Fifth Edition present test-age equivalents of total raw scores for each of their subtests.

Ratio Intelligence Quotients

Intelligence tests designed during the early part of the twentieth century used ratio IQs. *Ratio IQs* were defined as ratios of mental age (MA) to chronological age (CA), multiplied by 100 to eliminate the decimal: IQ = MA/CA × 100. For example, substituting an MA of 12 and a CA of 10 into the formula yields a ratio IQ of 120 (IQ = 12/10 × 100 = 120). Mental age represented the age of the group of children who obtained, on average, the given number of raw score points. Thus, for example, if 87 raw score points was the average obtained by 12-year-old children in the standardization sample of a test, then all children subsequently tested with the instrument who scored 87 were assigned a mental age of 12-0.

Ratio IQs are problematic for at least two reasons. First, because raw scores on intelligence tests increase linearly with age only up to about 16 years, the conversion of raw scores to a mental age beyond age 16 years is problematic. And we still do not know precisely when mental development reaches a ceiling level. Second, ratio IQs for different ages are not comparable because the standard deviation of the ratio IQ distribution does not remain constant with age. The same ratio IQ has different meanings at different ages.

Contemporary intelligence tests do not use mental age to calculate IQ. Instead, the IQ represents a standard score and, in most cases, has a mean of 100 and a standard deviation of 15 (see the section on standard scores later in the chapter). Standard scores avoid the two problems described above. However, as noted above, some current intelligence tests do provide age equivalents, which also can be thought of as mental-age scores.

We do not recommend the use of ratio IQs except when standard scores are not available and it is necessary to make a crude approximation of a child's level of ability. This may happen, for example, when the child being assessed is chronologically too old for one test and mentally too young for another test. A ratio IQ would allow you to conclude, for example, that a child with a CA of 10-0 and an MA of 12-0 has performed at an above-average level, whereas a child with a CA of 10-0 and an MA of 8-0 has performed at a below-average level.

Relationships Among Derived Scores

All derived scores are obtained from raw scores. The different derived scores are merely different expressions of a child's performance. Which derived score is used in a given field is more or less an arbitrary historical convention:

- Scores on cognitive measures tend to be expressed as standard scores with $M = 100$ and $SD = 15$.
- Scores on personality and behavioral measures tend to be expressed as T scores with $M = 50$ and $SD = 10$.
- Scores on other assessment measures, such as those used by occupational therapists, tend to be expressed as z scores with $M = 0$ and $SD = 1$.

The mathematical formulas described in this section make it easy to transform one type of derived score to another. The most frequently used conversion in the area of intelligence testing is from standard scores to percentile ranks (see Figure 4-1). Although standard scores are the preferred derived scores, percentile ranks—and, on occasion, age equivalents—also are useful, as they can help describe a child's performance to parents or teachers. Percentile ranks, however, are often misinterpreted as indicating the percentage of questions that the child answered correctly. Do not use the abbreviation "%" or "%tile" for "percentile rank" because these abbreviations may be understood as "percent correct." Instead, we recommend that you spell out the words *percentile rank* in your report.

Figure 4-1 shows the relationships among various derived scores. If a test has a standard score mean IQ of 100, a standard deviation of 15, and scores that are normally distributed, we can precisely determine the percentile ranks associated with each IQ. To illustrate, we will determine the percentile ranks associated with Wechsler IQs at several standard deviation points.

Let's begin by determining the percentile rank associated with an IQ of 115. An IQ of 115 is at the point that is +1 *SD* away from the mean. Although there are mathematical procedures for computing percentile ranks precisely, you can simply look at Figure 4-1 and determine the percentile rank associated with an IQ of 115—the 84th percentile—by adding 34% to 50%. The 50% is the proportion of the population below the mean of 100, and the 34% is the proportion of the population between the mean and +1 *SD* away from the mean. The key is to recognize that an IQ of 115 is +1 *SD* above the mean because 15 is the standard deviation of the distribution in this example.

You can also look at Figure 4-1 to determine the percentile ranks of other IQs. Note that an IQ of 130 is +2 *SD* away from the mean. We know that the area below the mean represents 50% of the population, the area from the mean to +1 *SD* represents approximately 34% of the population, and the area from +1 *SD* to +2 *SD* represents approximately 14% of the population. To find the percentile rank for an IQ of 130, we add 50 + 34 +14 to get the 98th percentile rank.

To figure out the percentile rank associated with an IQ of 85, subtract 34 from 50, because an IQ of 85 corresponds to the point that is –1 *SD* away from the mean. The answer is the 16th percentile rank. An IQ of 70 is associated with the second percentile rank (50 – 34 – 14 = 2). Note that the above examples hold only for tests with $M = 100$ and $SD = 15$ (e.g., WISC–IV, WPPSI–III, WAIS–III, and SB5). A glance at Table BC-1 on the inside back cover will show you the percentile ranks associated with IQs based on $M = 100$ and $SD = 15$.

INFERENTIAL STATISTICS

Inferential statistics are used in drawing inferences about a population based on a sample drawn from the population. Consider an experiment in which the scores obtained on a fluency reading test by 100 children who were enrolled in a 10-week speed-reading program were 25 points higher than those of 100 children who were not enrolled in the program. Is the difference significant or is it just due to chance? And what about the real difference for the population—how much larger or smaller is it likely to be than the 25 points found in the sample? These are questions that can be answered by inferential statistics.

Statistical Significance

When we want to know whether the difference between two or more scores can be attributed to chance or to some systematic or hypothesized cause, we run a test of statistical significance. *Statistical significance* refers to whether scores differ from what would be expected on the basis of chance alone. Statisticians have generally agreed that a reasonable criterion for deciding that something is not a chance occurrence is that it would happen by chance only 5% of the time or less. The expression $p < .05$ means that the results have a probability level of less than .05 (or 5 or fewer times in 100) of occurring by chance, whereas the expression $p > .05$ means that the results have a probability level of greater than .05 (or more than 5 times in 100) of occurring by chance. By convention, the first is considered statistically significant; the second is not. Thus, the .05 significance level indicates that we can have confidence that an observed difference would occur by chance only 5% of the time.

There also are more stringent levels of significance, such as the .01 (1 time in 100) and the .001 (1 time in 1,000) levels. Researchers choose a more or less stringent level of significance depending on how confident they need to be about results. Tests of significance are used to evaluate differences between two or more means, differences between a score and the mean of the scale, and differences of correlations from zero (or chance).

Effect Size

Tests of significance, although highly useful, don't tell us the complete story. Because tests of significance are highly dependent on sample size, larger sample sizes are more likely to provide statistically significant results. Consequently, two seemingly similar studies will yield apparently inconsistent outcomes if one study uses a small sample and the other a large sample.

We need to consider not only statistical significance, but also the values of the means, the degree to which the means differ, the direction of the mean difference, and whether the results are meaningful—that is, whether they have important practical or scientific implications. The difference between the means of two groups may be statistically significant and yet have no practical significance. For example, if one group of 200 individuals has a mean of 100 and another group of similar size has a mean of 101, the significance test may yield a p value less than .05 because the groups are large, but the difference of only 1 point may have little practical meaning. In another study, if one group of 20 individuals has a mean of 100 and the other group of 20 has a mean of 110, the significance test may yield a p value larger than .05 (nonsignificant) because the groups are small, yet the difference of 10 points could be meaningful.

Effect size (*ES*) is a statistical index based on standard deviation units, independent of sample size. It measures the degree or magnitude of a result—that is, the difference between two group means (or treatment effects)—rather than the probability that the result is due to chance. Effect size statistics provide a standard context for interpreting "meaningful" results independent of sample size and statistical significance. We recommend that both effect size and statistical significance tests be reported in research reports. Effect size is also used in meta-analysis, which is discussed later in the chapter.

Cohen's *d*. Cohen's *d*, a statistic in standard deviation units, provides one way to compute effect size (Cohen, 1988). This statistic represents the distance between the means of two groups in standard deviation units. To compute *d*, use the following formula:

$$d = \frac{M_1 - M_2}{SD_{\text{pooled}}}$$

where M_1 = mean of group 1
M_2 = mean of group 2
SD_{pooled} = square root of the average of the two squared standard deviations, or

$$SD_{\text{pooled}} = \sqrt{\frac{SD_1^2 + SD_2^2}{2}}$$

Cohen (1988) defined effect size as small if $d = .20$, medium if $d = .50$, and large if $d = .80$, although not all researchers agree with these descriptive terms (e.g., Hopkins, 2002). You may want to consider using the following terms to describe the strength of effect sizes based on their corresponding correlation coefficients (see the formula for converting r to d later in the chapter):

- 2.68 or higher: very high
- 1.51 to 2.67: strong
- .88 to 1.50: moderate
- .42 to .87: low
- .41 or lower: very low

Because effect size values are in standard deviation units, we can use the normal curve to find out how many percentile points are represented by any effect size. (Note that most statistics books have a table that shows the areas of the normal curve.) Let's take three examples:

1. An effect size of .60 represents a difference of 23 percentile points (the area covered in a normal curve between 0.00 and .60 standard deviation units is .2257).
2. An effect size of 1.34 represents a difference of 41 percentile points (the area covered in a normal curve between 0.00 and 1.34 standard deviation units is .4099).
3. An effect size of 2.00 represents a difference of 48 percentile points (the area covered in a normal curve between 0.00 and 2.00 standard deviation units is .4772).

Let's look at an example. A psychologist wants to determine whether a new speed-reading program improves reading comprehension scores. She randomly assigns children with reading problems to a speed-reading program group and to a control group. Pre- and post-tests are administered. She finds that both groups had similar scores at the beginning of the study, whereas at the end of the study the mean score of children who took part in the speed-reading program was 11 points higher than the mean score of the control group. This difference was significant ($p < .05$). In addition, she finds an effect size of .60, which is a medium effect by Cohen's criteria (but a low effect based on the correlation coefficient related to it), and she concludes that the program made somewhat of a difference by improving mean reading comprehension by 23 percentile points.

Now let's compare effect size statistics with traditional significance test statistics for a study designed to improve written expression skills. Suppose that 60% of the children in the study were at grade level in written expression at the beginning of the study. Of a sample of 1,000 children, 65% would need to be at grade level (an increase of 5 percentage points) at the end of the study in order to produce a statistically significant finding. However, if the sample size were 50, then 79% would need to be at grade level (an increase of 19 percentage points) at the end of the study to reach this same level of significance. In contrast, if the improvement goal were to meet a minimum effect size of $d = .2$, the increase needed to reach this level is 10 percentage points (to 70%), regardless of whether the sample size was 50 or 1,000.

Correlation coefficient (r). Significance testing for correlation coefficients also doesn't tell the whole story. In addition to indicating the coefficient's level of significance, r can also be used to evaluate effect size (Hunter & Schmidt, 2004). The correlation coefficient can be converted to d by use of the following formula:

$$d = \frac{2r}{\sqrt{1 - r^2}}$$

RELIABILITY

Theory of Reliability of Measurement

If we administer the same test to children on several occasions, they will likely earn different scores. Sometimes the scores change systematically (i.e., there is a regular increase or decrease in scores), and sometimes the scores change randomly or unsystematically (i.e., there is no discernable pattern to the increase or decrease in scores). A reliable test is one that is consistent in its measurements. In contrast, a test is unreliable if scores are subject to large random, unsystematic fluctuations; obviously, a test is not dependable if the scores change significantly on re-administration after a short time during which children receive no intervention. Technically, reliability of measurement refers to the extent to which random or unsystematic variation affects the measurement of a trait, characteristic, or quality.

According to classical psychometric theory, a test score is composed of two components: a true score and an error score. (The word *true* refers to the measurement process, not to the underlying content of the test.) A true score represents a combination of all the factors that lead to consistency in measurement of a characteristic. A child's true score is a hypothetical construct; we cannot measure it precisely. However, we can hypothesize that if we repeatedly gave the child the same test, his or her scores would be distributed around the true score. The mean of this assumed normal distribution would approximate the true score. An *error score* represents random factors that affect the measurement of the true score. The theory assumes that (a) the child possesses stable traits, (b) errors are random, and (c) the observed test score is the sum of the true score and the error score. The reliability coefficient is the ratio of the true score variance to the observed score variance.

Reliability Coefficients

The *reliability coefficient,* which expresses the degree of consistency in the measurement of test scores, is denoted by the letter r with a subscript consisting of identical letters (e.g., r_{xx} or r_{tt}). Reliability coefficients range from 1.00 (indicating perfect reliability) to .00 (indicating the absence of reliability). The four major types of reliability are internal consistency reliability, test-retest reliability, alternate-forms reliability, and interrater reliability. We use the Pearson product-moment correlation formula (see Table 4-4) to compute test-retest and alternate-forms reliability coefficients, specialized formulas to compute internal consistency reliability coefficients, and several different methods to compute interrater reliability coefficients. Table 4-6 shows some procedures for determining reliability.

Reliability is essential in a psychological measure. Low levels of reliability signify that unknown but meaningful sources of error are operating in the measure and that the measure is not stable across time or consistent across situations. Test results need to be reliable—that is, dependable, reproducible, and stable. Imagine the chaos if, when a student took two equivalent forms of the SAT on the same day, the student scored at the 85th percentile rank on one form and at the 40th percentile rank on the "equivalent" second form. Clearly the reliability value of such a test would not be satisfactory. Reliabilities above .80 are preferred for tests used in individual assessment; reliabilities should be at or above .90 for test results to be used in decision making.

The following are useful ways to describe reliability coefficients (Murphy & Davidshofer, 2005):

- .00 to .59: very low or very poor reliability
- .60 to .69: low or poor reliability
- .70 to .79: moderate or fair reliability
- .80 to .89: moderately high or good reliability
- .90 to .99: high or excellent reliability

Internal consistency reliability. *Internal consistency reliability* is based on the scores that individuals obtain during a single administration of a test. The most general measure of reliability is *Cronbach's coefficient alpha,* which can be used for different scoring systems and is based on the variance of the test scores and the variance of the item scores. Coefficient alpha measures the uniformity, or homogeneity, of items throughout the test (see Table 4-6). The values obtained by using the *Kuder-Richardson formula 20 coefficient,* a special case of coefficient alpha, are useful for tests whose items are scored as pass/fail or right/wrong. The values obtained from the *Spearman-Brown correction formula,* used to estimate reliability by the split-half method, are interpreted in the same way as coefficient alpha. (The *split-half method* involves correlating pairs of scores obtained from equivalent halves of a test administered only once.) Internal consistency reliability estimates are not appropriate for timed tests, and they do not take into account changes over time. Generally, the size of the internal consistency coefficient increases with test length; the longer the test, the higher the coefficient.

Test-retest reliability. *Test-retest reliability* is computed from the scores that individuals obtain on the same test on two different occasions. The obtained correlation—sometimes called the *coefficient of stability*—provides an index of

Table 4-6
Some Procedures Used to Determine Reliability

Procedure	*Description*
Cronbach's coefficient alpha (α) formula $r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{S_t^2 - \Sigma S_i^2}{S_t^2}\right)$ where r_{tt} = coefficient alpha reliability estimate n = number of items on the test S_t^2 = variance of the total scores on the test ΣS_i^2 = sum of the variances of individual item scores	An *internal consistency reliability* formula used when a test has no right or wrong answers. This formula provides a general reliability estimate. It is an efficient method of measuring internal consistency. Coefficient alpha essentially indicates the average intercorrelation between test items and any set of items drawn from the same domain.
Kuder-Richardson formula 20 (KR$_{20}$) $r_{tt} = \left(\frac{n}{n-1}\right)\left(\frac{S_t^2 - \Sigma pq}{S_t^2}\right)$ where r_{tt} = reliability estimate n = number of items on the test S_t^2 = variance of the total scores on the test Σpq = sum of the product of p and q for each item p = proportion of people getting an item correct q = proportion of people getting an item incorrect	An *internal consistency reliability* formula used for calculating the reliability of a test in which the items are scored 1 or 0 (or right or wrong). It is a special form of the coefficient alpha formula for use with dichotomous items.
Spearman-Brown correction formula $r_{nn} = \frac{kr_{tt}}{1 + (k-1)r_{tt}}$ where r_{nn} = estimated reliability coefficient k = number of items on the revised version of the test divided by number of items on the original version of the test r_{tt} = reliability coefficient before correction	An *internal consistency reliability* formula used to evaluate the effect that lengthening or shortening a test will have on the reliability coefficient. The formula increases the reliability estimate when the test is lengthened.
Product-moment correlation coefficient formula See Table 4-4 for the formula.	A formula used to estimate *test-retest reliability* or *parallel-forms reliability*

the consistency, or replicability, of test scores over relatively short intervals, during which scores would not be expected to change. The test-retest method is useful for evaluating the reliability of ability tests; it is less useful with behavioral checklists and scales, observational procedures, and related forms of measurement. Because the latter instruments tend to provide different readings each time measurement is conducted, lower test-retest reliability coefficients may result when they are re-administered. This does not necessarily mean that the instruments are faulty—that is, that there is measurement error. Rather, the behaviors being measured may have changed. Consequently, you should carefully consider whether low test-retest reliabilities are associated with poorly designed instruments or with actual changes (as a result of life changes, tutorials, or interventions) in children's behavior, attitudes, temperament, or other characteristics being measured.

Test-retest correlation is affected by factors associated with the specific administrations of the test and with what children remember or have learned in the interim. Any variables that affect children's performance on one occasion but not on the other will affect the test-retest reliability. Typical influencing variables include differences in administration (e.g., different examiners, different rooms, different times of the day) and differences in the children themselves (e.g., fatigue, mood, motivation). Generally, the shorter the retest interval, the higher the reliability coefficient, because within a shorter span of time there are fewer such reasons for children's scores to change. With individual intelligence tests, test-retest reliabilities are generally higher when the retest interval is less than 10 months and when the children are older adolescents (Schuerger & Witt, 1989).

Alternate-forms reliability. *Alternate-forms reliability* (also referred to as *parallel-forms reliability* or *equivalent-forms reliability*) is determined by creating two different but parallel forms of a measure and administering the two forms to the same group of children. The extent of agreement of a group's scores on the two forms, sometimes referred to as a *coefficient of equivalence,* is used as an index of reliability. For example, two forms of a measure of intelligence might be created, with different items in the two forms measuring the same construct.

The two forms would then be given to a large sample. Half of the sample would receive form A followed by form B, and the other half of the sample would receive form B followed by form A. Scores from the two forms would then be correlated, yielding a reliability coefficient.

If the two forms of a test are equivalent, they should yield the same means and variances, be highly correlated, and have high reliability coefficients (.80 or higher). If there were no error in measurement, children should earn the identical score on both forms of the test. For the forms to be truly parallel, each equivalent item on the two forms should have the same response split (number of individuals answering each item right or wrong) and the same correlations with other tests. This level of test equivalence is difficult, if not impossible, to achieve.

Alternate-forms reliability coefficients are subject to some of the same influences as test-retest reliability coefficients, such as decreased reliability as the interval between the tests increases. Because children are not tested twice with the same items, however, there is less chance than with the test-retest method that memory for specific item content will affect the scores. Constructing alternate forms is usually easier for tests that measure intellectual ability or specific academic abilities than for those that measure personality, temperament, or motivation, as the latter constructs are more difficult to define.

Interrater Reliability

Interrater reliability (also called *examiner reliability* or *scorer reliability*) refers to the degree to which the raters agree. The most common measure of interrater reliability is *percentage agreement.* This statistic tells us the percentage of items on which two or more raters gave the identical rating to the behavior or criterion being judged (e.g., raters gave the same rating to 80% of the items). Percentage agreement is not a reliability coefficient, because it provides no information about the measurement procedure itself. Furthermore, percentage agreement does not take into account that chance alone would lead to some agreement. However, because percentage agreement does indicate the extent to which two or more raters gave the same score or rating, it contributes to our understanding of the objectivity of the scoring, a factor related to reliability. Other ways to evaluate interrater reliability are with kappa and the intraclass correlation coefficient—both of which account for chance agreement—and the product-moment correlation coefficient.

Factors Affecting Reliability

The following factors affect the reliability of a test (also see the discussion of repeated evaluations and practice effects later in the chapter):

1. *Test length.* The more items there are on a test, the greater the internal consistency reliability is likely to be.

2. *Homogeneity of items.* The more homogeneous or similar to each other the items on a test are, the greater the reliability is likely to be.

3. *Test-retest interval.* The smaller the interval between administration of two tests, the smaller the chance of change in the child taking the test and, hence, the higher the test-retest reliability is likely to be.

4. *Variability of scores.* The greater the variance of scores on a test, the higher the reliability estimate is likely to be. Small changes in performance have a greater impact on the reliability of a test when the range, or spread, of scores is narrow than when it is wide. Therefore, on a given test, homogeneous samples (those with a small variance) will probably yield lower reliability estimates than heterogeneous samples (those with a large variance).

5. *Guessing.* The less guessing that occurs on a test (i.e., the less often children respond to items randomly), the higher the reliability is likely to be. Even guessing that results in correct answers introduces error into the score.

6. *Variation in the test situation.* The fewer variations there are in the test situation, the higher the reliability is likely to be. Child factors, such as misunderstanding instructions, illness, and daydreaming, and examiner factors, such as misreading instructions and making scoring errors, introduce an indeterminate amount of error into the testing procedure.

7. *Sample size.* Reliability coefficients are more meaningful when the sample represents a large group, as well as when the children closely resemble the sample on which the reliability coefficient was based. Although the standard error of measurement (see below) is not directly related to reliability, the sampling error associated with the reliability coefficient will be smaller when the sample size is large. For example, a reliability estimate of .80 based on a sample of 26 yields an estimated standard error of .07, whereas one based on a sample of 201 yields an estimated standard error of .03, a value less than half as large. Larger samples thus provide a more dependable estimate of reliability.

Standard Error of Measurement

The *standard error of measurement* (SEM), or standard error of a score, is an estimate of the amount of error inherent in a child's obtained score. It is important to consider this estimate, because some measurement error is associated with every test score and thus there is almost always some uncertainty about a child's true score. The standard error of measurement directly reflects the reliability of a test: the lower the reliability, the higher the standard error of measurement; conversely, the higher the reliability, the lower the standard error of measurement. Large standard errors of measurement reflect less stable measurements. Of course, the size of the SEM is also related to the standard deviation of the metric (or standard of measurement): the larger the standard deviation, the larger the SEM. Thus, for example, the SEM will be larger when the total score has a mean of 100 and a standard deviation of 15 than when the total score has a mean of 50 and a standard deviation of 10.

The standard error of measurement represents the standard deviation of the distribution of error scores. You can also think of the SEM as an estimate of how one person's repeated scores on the same measure tend to be distributed around his or her true score. We compute the SEM by multiplying the standard deviation (*SD*) of the test by the square root of 1 minus the reliability coefficient (r_{xx}) of the test:

$$\text{SEM} = SD\sqrt{1 - r_{xx}}$$

This equation indicates that as the reliability of a test increases, the standard error of measurement decreases. With a reliability coefficient of 1.00, the standard error of measurement would be zero. With a reliability coefficient of .00, the standard error of measurement would be equal to the standard deviation of the scores in the sample.

Confidence Intervals for Obtained Scores

When we report a test score, we also should report a *confidence interval*—a band, or range, of scores around the obtained score that likely includes the child's true score. The confidence interval may be large or small, depending on the degree of certainty we desire (how likely we want it to be that the interval around the child's obtained score contains his or her true score). Traditionally, we select points that represent the 68%, 95%, or 99% level of confidence, although we also can use the 85% or 90% level. A 95% confidence interval can be thought of as the range in which we will find a child's true score 95% of the time. With a 95% confidence interval, the statistical chances are only 5 in 100 that a child's true score lies outside the range encompassing the obtained score. It is not possible to construct a confidence interval within which a child's true score is certain to lie unless the entire distribution of scores is known.

Although you can usually use confidence intervals for various scores obtained by a child on a test (such as subtest scaled scores), we recommend that you use confidence intervals primarily for the overall score, such as the WISC–IV Full Scale IQ, because the overall score is usually the score used for diagnosis and classification. ***Individuals who use the test findings need to know that the IQ and other major scores used to make decisions about a child are not perfectly accurate because they inherently contain measurement error.*** Consequently, you should report confidence intervals associated with the IQ and other similar total or overall scores.

There are two methods for obtaining confidence intervals. One is based on the child's obtained score and the conventional standard error of measurement. The other is based on the estimated true score and the standard error of measurement associated with the estimated true score (also called the *standard error of estimate*). The following guidelines will help you to determine which type of confidence interval to use. Note that in all of the examples in this section, the confidence intervals have been rounded up to the next whole number.

Confidence interval based on obtained score and conventional standard error of measurement (SEM). When you base the confidence interval solely on the child's obtained score, without reference to his or her estimated true score, use the SEM for obtained scores.

You obtain the confidence interval by using the following formula:

$$\text{Confidence interval} = \text{obtained score} \pm (z)(\text{SEM})$$

The formula shows that two values are needed in addition to the child's test score: the z score associated with the confidence level chosen and the standard error of measurement. You can obtain the z score from a normal distribution table, found in most statistics textbooks. We used a normal distribution table to obtain the following values for the five most common levels of confidence:

68% level,	$z = 1.00$
85% level,	$z = 1.44$
90% level,	$z = 1.65$
95% level,	$z = 1.96$
99% level,	$z = 2.58$

You can usually find the SEM in the manual that accompanies a test, or you can compute it using the formula given previously. You compute the upper limit of the confidence interval by adding the product (z)(SEM) to a child's score, and the lower limit by subtracting the product from a child's score (thus the plus-or-minus symbol, ±, in the equation for the confidence interval).

Here is an example of how to construct a confidence interval, given a standard error of measurement of 3 and an IQ of 100. First we need to select a confidence level. Let's say that we select the 95% level. The z score associated with the 95% level is 1.96. To obtain the confidence interval, we multiply this value by the standard error of measurement, 3, and add a ± sign to the result to represent the upper and lower limits of the interval. Thus, the confidence interval is approximately 100 ± 6. The value 6 is then added to and subtracted from the obtained score to determine the specific band, or interval, associated with the obtained score. The upper limit of the interval is given by

$$\begin{aligned}\text{Confidence interval upper limit} &= 100 + 1.96(3)\\ &= 100 + 6 = 106\end{aligned}$$

and the lower limit of the interval is given by

$$\begin{aligned}\text{Confidence interval lower limit} &= 100 - 1.96(3)\\ &= 100 - 6 = 94\end{aligned}$$

Because the z score we used was associated with the 95% level, we can say that the chances that the child's true score is between 94 and 106 are about 95 out of 100.

For an IQ of 100 (with SEM = 3), the interval would be 100 ± 3 (97 to 103) at the 68% confidence level, 100 ± 4 (96 to 104) at the 85% confidence level, 100 ± 5 (95 to 105) at the 90% confidence level, and 100 ± 8 (92 to 108) at the 99% confidence level. The latter band indicates that the chances that the child's true score is between 92 and 108 are about 99 out of 100. Notice that we must increase the band width to increase our level of confidence (or degree of certainty).

As another example, let's construct several confidence intervals for a child who obtains an IQ of 80 on a test for which SEM = 5. We complete the equation for the 90% level of confidence in the following way:

$$\begin{aligned}\text{Confidence interval} &= \text{obtained score} \pm (z)(\text{SEM}) \\ &= 80 \pm 1.65(5) \\ &= 80 \pm 8 = 72 \text{ to } 88\end{aligned}$$

For the 99% level of confidence, the equation is as follows:

$$\begin{aligned}\text{Confidence interval} &= 80 \pm 2.58(5) \\ &= 80 \pm 13 = 67 \text{ to } 93\end{aligned}$$

Appendixes A, B, C, and F in the Resource Guide show the confidence intervals for the Composites for the WISC–IV (Table A-1), the WPPSI–III (Table B-1), the WAIS–III (Table C-1), and the SB5 (Table F-1), based on the obtained score and the conventional standard error of measurement—that is, without recourse to the estimated true score or the standard error of estimate. Use of a child's specific age group in these tables allows you to obtain the most accurate confidence interval.

Confidence interval based on estimated true score and its standard error of estimate. When you base the confidence interval on the child's obtained score with reference to his or her estimated true score, use the standard error of estimate for estimated true scores. This confidence interval will be based on statistics that take into account the effects of regression toward the mean.

Because the WISC–IV is widely used, it will be used in this section to illustrate how confidence limits are obtained with reference to the estimated true score. Table D-2 in Appendix D in the Resource Guide shows the confidence intervals, by age, for the WISC–IV Composites and Full Scale, based on the estimated true score and the appropriate standard error of measurement. You simply apply the confidence intervals in Table D-2 to the obtained score on the WPPSI–III, the WAIS–III, the SB5, and any other test with M = 100 and SD = 15 that has a reliability coefficient of .85 to .98.

The formula used to obtain the estimated true score is

$$T = r_{xx}(X - \bar{X}) + \bar{X}$$

where T = estimated true score
r_{xx} = reliability of the test
X = obtained score
$\bar{X}$ = mean of the test

Thus, the estimated true score for an obtained WISC–IV Full Scale IQ of 60 (where $r_{xx} = .97$) is

$$\begin{aligned}T &= .97(60 - 100) + 100 \\ &= -39 + 100 = 61\end{aligned}$$

The formula used to obtain the standard error of estimate (SE_{est}) is as follows:

$$SE_{\text{est}} = r_{xx}\text{SEM}$$

where SE_{est} = standard error of estimate (or standard error of measurement of the true score)
r_{xx} = reliability of the test
SEM = standard error of measurement of the test

If, in our example, the SEM was 2.68, the standard error of estimate would be

$$SE_{\text{est}} = .97(2.68) = 2.60$$

Because the confidence intervals are centered around the estimated true score, the intervals become asymmetrical when applied to the obtained score. The asymmetry is greater for values farther from the mean, because regression to the mean increases at the extremes of the distribution. In fact, for scores at or near the mean, there is no asymmetry at all—the confidence intervals are equal around the mean. For example, as Table 4-7 shows, for the WISC–IV Verbal Scale at age 16 (Section O) at the 95% confidence level, the confidence interval for an IQ of 40 is from 40 – 3 to 40 + 9 (37 to 49), whereas the confidence interval for an IQ of 91 is from 91 – 6 to 91 + 7 (85 to 98). The procedure used to obtain the confidence intervals in Table D-2 in Appendix D in the Resource Guide is the same one used by The Psychological Corporation in the construction of the confidence intervals in the WISC–IV Administration Manual.

To use Table D-2 in Appendix D in the Resource Guide, follow this procedure. First, use the list at the beginning of the table to find which section of the table applies to the child's age, the appropriate test (WISC–IV, WPPSI–III, WAIS–III, or SB5), and the appropriate Composite. Then select one confidence level from the columns labeled 68%, 85%, 90%, 95%, and 99%. The values in the table under the appropriate confidence level will allow you to calculate the lower (L) and upper (U) limits of the confidence interval for the obtained IQ. If the value is positive (when no sign precedes the value, the + sign is understood), add the absolute value to the obtained IQ. If the value is negative (a – sign precedes the absolute value), subtract the absolute value from the obtained IQ. Usually, you will find the lower limit by subtracting an absolute value from the obtained IQ, and you will find the upper limit by adding an absolute value to the obtained IQ.

For example, to calculate the confidence interval for a 12-year-old child who obtains a WISC–IV Full Scale IQ of 46, see Table D-2, Section O, in Appendix D in the Resource Guide. Section O shows that the values at the 68% confidence level for the lower and upper limits of the confidence interval are 0 and 6, respectively. (Table 4-7 shows the Section O part of Table D-2.) Because both values are positive, you can obtain the lower and upper limits of the confidence interval by adding the absolute values to the obtained IQ. The resulting confidence interval is 46 to 52 (lower limit is 46 + 0 = 46; upper limit is 46 + 6 = 52).

Table 4-7
Part of Table D-2 in Appendix D in the Resource Guide Showing Confidence Intervals Based on the Estimated True Score for Wechsler Scales and Stanford-Binet Fifth Edition for $r_{xx} = .95$

O. $r_{xx} = .95$
WISC–IV: Verbal Comprehension Index, Ages 12, 14, 15, and 16
WPPSI–III: Verbal IQ, Ages 2½, 3, 3½, Average for Ages 2½–3$^{11}/_{12}$, 4, 4½, 5½, 7, and Average for Ages 4–7¼
WPPSI–III: Performance IQ, Age 7
WPPSI–III: Full Scale IQ, Ages 2½, 3, and Average for Ages 2½–3$^{11}/_{12}$
WAIS–III: Performance Scale IQ, Ages 25–29, 55–64, and 70–74
SB5: Nonverbal IQ, Ages 2, 4, 5, 9, 14, 30–39, and Average
SB5: Verbal IQ, Ages 2, 7, 10, and 13

68%			*85%*			*90%*			*95%*			*99%*		
IQ	*L*	*U*	*IQ*	*L*	*U*	*IQ*	*L*	*U*	*IQ*	*L*	*U*	*IQ*	*L*	*U*
40–46	0	6	40–41	−2	8	40–44	−2	8	40–45	−3	9	40–45	−5	11
47–53	−1	6	42–58	−2	7	45–55	−3	8	46–54	−4	9	46–54	−6	11
54–66	−1	5	59–61	−3	7	56–64	−3	7	55–65	−4	8	55–65	−6	10
67–73	−2	5	62–78	−3	6	65–75	−4	7	66–74	−5	8	66–74	−7	10
74–86	−2	4	79–81	−4	6	76–84	−4	6	75–85	−5	7	75–85	−7	9
87–93	−3	4	82–98	−4	5	85–95	−5	6	86–94	−6	7	86–94	−8	9
94–106	−3	3	99–101	−5	5	96–104	−5	5	95–105	−6	6	95–105	−8	8
107–113	−4	3	102–118	−5	4	105–115	−6	5	106–114	−7	6	106–114	−9	8
114–126	−4	2	119–121	−6	4	116–124	−6	4	115–125	−7	5	115–125	−9	7
127–133	−5	2	122–138	−6	3	125–135	−7	4	126–134	−8	5	126–134	−10	7
134–146	−5	1	139–141	−7	3	136–144	−7	3	135–145	−8	4	135–145	−10	6
147–153	−6	1	142–158	−7	2	145–155	−8	3	146–154	−9	4	146–154	−11	6
154–160	−6	0	159–160	−8	2	156–160	−8	2	155–160	−9	3	155–160	−11	5

Note. L = lower confidence interval; U = upper confidence interval.

Note that, although you calculate the values for the confidence intervals for the estimated true score, they are applied to the obtained score. Also note that you do not provide the estimated true score in the report; it is used only to generate the confidence interval.

Table D-2 in Appendix D in the Resource Guide is based on the child's age and not on average values for the total sample; in contrast, confidence intervals in the WISC–IV Administration Manual are based on the total sample. *Use of the child's specific age group allows you to obtain the most accurate confidence interval.*

Comment on confidence intervals. In clinical and psychoeducational assessments, questions usually center on how a child is functioning at the time of the referral. *Therefore, we recommend that you use the confidence interval based on the child's obtained score, without recourse to the child's estimated true score.* If you follow this recommendation, use the confidence interval for the obtained score and the conventional standard error of measurement—see Table A-1 in Appendix A in the Resource Guide. Be aware that the WISC–IV Administration Manual does not provide a similar table. However, when you want to know how a child might perform over a longer period in relation to a specific reference group, use the confidence interval based on the estimated true score—see Table D-2 in Appendix D in the Resource Guide. Again, the confidence intervals shown in Table D-2 are more appropriate than those shown in the WISC–IV Administration Manual because they are based on the child's specific age and not on the total sample. For most purposes, we recommend using confidence intervals at the 95% level of confidence. Note again that confidence bands will be broader with higher levels of confidence (e.g., 95% vs. 68%).

Confidence Intervals for Predicted Scores

Earlier in the chapter we discussed regression equations and the standard error of estimate associated with the predicted score. The standard error of estimate allows us to establish a confidence interval around a predicted score. This confidence interval is obtained in the following way:

$$\text{Confidence interval} = Y_{\text{pred}} \pm (z)(SE_{\text{est}})$$

The confidence interval for predicted scores is similar to the confidence interval for obtained test scores. If we use a z score of 1, then the standard error of estimate tells us that we can expect the predicted score to fall within the range bounded by the standard error of estimate about 68% of the time. If we want to have more confidence in the prediction, we can use a z score associated with, for example, the 95% confidence level (z = 1.96) or the 99% confidence level (z = 2.58). However, with higher levels of confidence, we expand the band (or range) around the predicted score.

The following three examples illustrate how to establish confidence intervals. In each case, let's assume that SE_{est} = 5 and Y_{pred} = 85.

- For the 68% level of confidence, the confidence interval is 85 ± 1.00(5). Thus, the confidence interval associated with the predicted score of 85 is 80.00 to 90.00 (there is a 68% chance that Y falls within this range).
- For the 95% level of confidence, the confidence interval is 85 ± 1.96(5). Thus, the confidence interval associated with the predicted score of 85 is 75.20 to 94.80 (there is a 95% chance that Y falls within this range).
- For the 99% level of confidence, the confidence interval is 85 ± 2.58(5). Thus, the confidence interval associated with the predicted score of 85 is 72.10 to 97.90 (there is a 99% chance that Y falls within this range).

Repeated Evaluations and Practice Effects

When a test is re-administered, retest scores may differ from those obtained on the initial test. Let's look at some findings on such changes in retest scores, known as *practice effects.*

1. *Practice effects may be related to prior exposure to the test.* Children may be particularly likely to obtain higher retest scores on items that require speed of performance, especially when the retest interval is short. Scores may also change if, between tests, children look up answers that they were unsure of during the first testing.

2. *Practice effects may occur because of intervening events between the two administrations.* Retest scores might be affected by such factors as a different examiner, setting, or time of day; traumatic events in the child's life and family; or changes in the child's health, motivation, or attention.

3. *Practice effects may not occur to the same extent in all populations.* Practice effects typically seen among children with average ability may not occur among children with mental retardation or children who are gifted. Practice effects may also differ as a function of the child's age or other variables, such as cultural and linguistic backgrounds.

4. *Practice effects vary for different types of tasks.* Nonverbal tasks (such as those found on the Wechsler Perceptual Reasoning Composite) usually show more practice effects than do verbal tasks (such as those found on the Wechsler Verbal Comprehension Composite; see Chapters 9 through 11 for a discussion of the Wechsler tests). Even tasks within the same performance or verbal area may show different practice effects.

5. *Practice effects may be affected by regression toward the mean.* Regression toward the mean is a statistical phenomenon whereby students with low scores on a first test tend to get higher scores on retest and students with high scores on a first test tend to get lower scores on retest. The idea of regression toward the mean is captured in everyday expressions such as "the law of averages," "things will even out," or "we are due for a good day after a string of bad ones." Regression toward the mean occurs because, on the first test, the low scores probably have negative errors of measurement (i.e., have been depressed) and the high scores probably have positive errors of measurement (i.e., have been inflated). Regression toward the mean does not affect scores at the center of the distribution because these scores probably have an equal number of negative and positive errors of measurement.

6. *Practice effects may be difficult to interpret when the initial test and the retest are different.* If you measure intelligence with test A on the first occasion and with test B on the second, changes in IQ may occur because of differences between the two tests, not because of changes in the child. An understanding of the properties of different tests, including how they are related to each other, is critical in evaluating retest changes.

7. *Practice effects may depend on the item content covered throughout the test.* A test of ability that covers a wide age range may actually tap different abilities at different ages, even though the test is said to measure only one ability or skill. For example, an intelligence test that covers ages 2 years through 18 years will usually measure different components of intelligence at 2 years than at 18 years. In such cases, it will be difficult to compare test results at these two ages and know precisely what any changes in test scores mean.

When a child obtains higher scores on retest, we don't know for sure whether the improvement was due to prior exposure to the material or to the child's improved cognitive functioning. When a child is expected to show gains on retest but does not, he or she may have a subtle learning deficit. This can happen, for example, with children who are brain injured or who are being reevaluated after brain surgery or chemotherapy.

For the results of repeated evaluations to be most useful, we need data on the differential effects of practice in relation to such factors as item content, age, gender, ability level, and illness (type, location, and chronicity). A database that provided normative retest changes on various tests for diverse normal and clinical populations would be extremely helpful in evaluating practice effects. Any clinical significance attributed to changes in test scores should be corroborated by other assessment and clinical data; validity data would be particularly important in this regard. Until such data become available for each test that you use, be careful in interpreting retest findings.

ITEM RESPONSE THEORY

Test developers traditionally look at certain values for each item on a test to see whether the item is performing properly, a process referred to as *item analysis.* One value is *item difficulty:* the percentage of children who answer an item correctly. It ranges from 0.0 for an item with maximum difficulty (everyone in the sample answers incorrectly) to +1.0 for an item with no difficulty (everyone in the sample answers correctly). A second value is *item discrimination:* how an item discriminates between children who do well on the test as a whole and those who do poorly. It ranges from −1.0 to +1.0. A value of +.8 for an item reflects excellent discrimination, whereas values from −.2 to +.2 indicate poor discrimination. A negative value, such as −.9, indicates that an item is a reverse discriminator—children who perform poorly on the test answer the item correctly more often than children who do well on the test. This may occur when the item is keyed incorrectly, when there is more than one correct answer (as in a multiple-choice test), or when the item is ambiguous.

In addition to item discrimination and item difficulty, *item response theory* (IRT), or the *latent trait model* (LTM), adds a third parameter: a "guessing" parameter, which reflects the probability that a correct response will occur by chance. A test developer places information on responses into mathematical equations, which then guide the construction of a test. IRT provides useful information about the relationship between the attribute being measured and the test responses. The mathematical relationship can be illustrated graphically with an *item characteristic curve*—a line representing the probability of passing the item for children with different total scores on the construct being measured.

Figure 4-4 shows item characteristic curves for two items on an intelligence test. Curve *a* reflects a good item; children with higher total test scores are more likely to answer that item correctly than are children with lower scores. In contrast, Curve *b* reflects a difficult item that has less discriminating power, because children with low total test scores are almost as likely to pass the item as are those with high total test scores. The slope of the curve tells you how effective the item is. A positive slope (i.e., one that rises from the lower left to the upper right) means that the item is a good discriminator, whereas a flat slope means that the item is a poor discriminator.

Here is an example of an application of an item characteristic curve:

> Item characteristic curves can be useful in identifying items that perform differently for different groups of children. For example, suppose a test developer was concerned that some reading-comprehension items dealing with farms might measure different processes for rural children than for urban children. To examine this question, the test developer would administer the test to groups of rural and urban children and determine the item characteristic curve for each item in each group. If an item is measuring the same thing in both groups, the item characteristic curves for that item should look the same in both groups. If the item is measuring different things in the two groups, the item characteristic curves can appear different. Items whose item characteristic curves are substantially affected by the group membership of the children can be revised or deleted from the test. (Adapted from Allen & Yen, 1979, pp. 129–130)

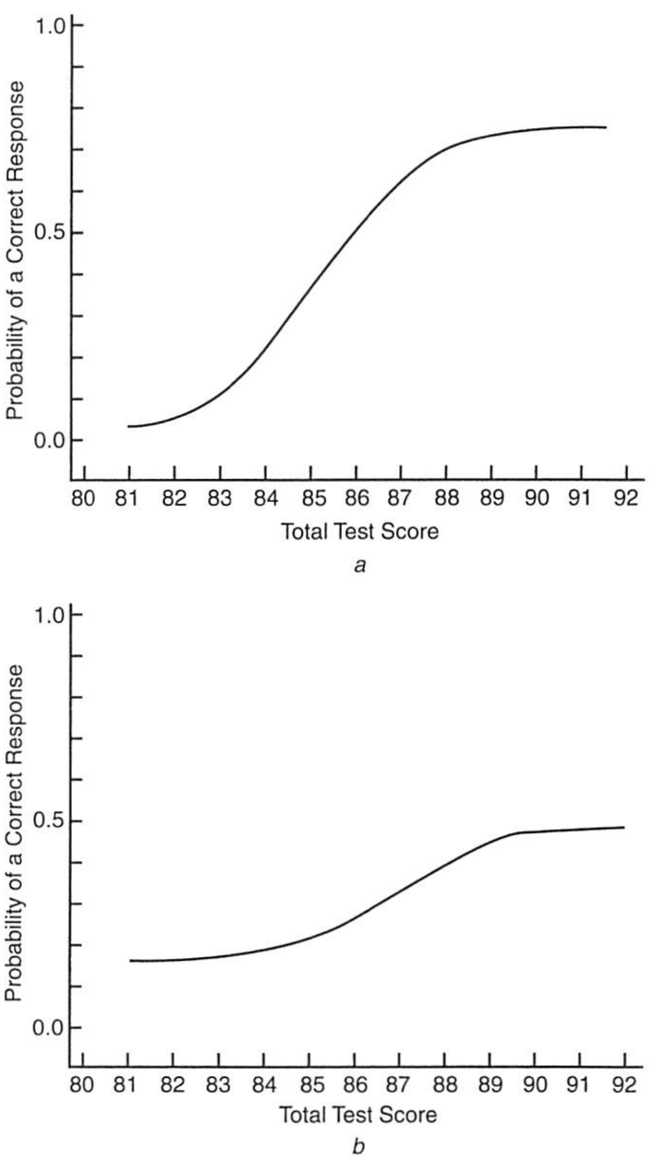


Figure 4-4. Two item characteristic curves.

Item response theory is also useful in adaptive testing:

> One of the most important applications of item response theory is to be found in computer-administered adaptive testing, also described as individualized, tailored, and response-contingent testing. This procedure adjusts the items to be administered to the responses actually given by each child to the preceding items. As the child responds to each item, the computer chooses the next item on the basis of the child's previous responses up to that point. Essentially each

child takes a test-item sequence and mix of items that is tailor-made to fit his or her performance. The test stops when enough information is available to reach a pass-fail level on the items. The child's test score is based not on the number of items passed, but on the predetermined score of each of the items passed, as determined by its difficulty level, discriminative value, and susceptibility to guessing. The item "score" represents the best estimate of the ability level at which the likelihood of passing the item is 50-50. Adaptive testing is thus made possible by the use of item response theory in developing the item pool. (Adapted from Anastasi, 1989, p. 479)

DIFFERENTIAL ITEM FUNCTIONING

The assessment of *differential item functioning* (DIF) is a statistical procedure designed to reveal whether test items function differently in different groups (Zumbo, 1999). The procedure is based on the principle that if different groups of children have the same level of ability, they should perform similarly on individual test items, regardless of their group membership. Differential item functioning occurs when children from different groups show differing probabilities of success on a test item after the groups have been matched on the underlying ability (i.e., overall score on the test) that the item is intended to measure.

Differential item functioning can occur uniformly or nonuniformly. It occurs uniformly if the difference in the probability of success is consistent across all levels of ability (e.g., the item favors all females regardless of ability). It occurs nonuniformly if the difference in the probability of success between the groups is not constant across ability levels—that is, if there is an interaction effect (e.g., the item favors females of low ability and males of high ability). The assessment procedure is useful for detecting item bias, but it is based on several questionable assumptions: that the test item measures a single trait, that the overall test is fair, and that the abilities measured by the test are equivalently distributed across all groups. Removing items judged as biased may not result in a fairer test if the groups being compared are not equal in the underlying construct being measured (Camilli, 1993).

VALIDITY

The *validity* of a test refers to whether it measures what it is supposed to measure. Validity determines the appropriateness of inferences or conclusions that are based on the test results. We use test results for such purposes as educational placements, program training, job qualification, and diagnosis. However, a test can't be used with confidence unless it is valid for the purpose for which it is used. Because tests are used for many different purposes, there is no single type of validity that is appropriate for all assessment purposes.

Validity is more difficult to define than reliability (Messick, 1989a, 1989b, 1995). Unlike reliability, validity has no single definition. A related problem is that the terminology used in the literature on validity is inconsistent. We will employ one set of terms in our discussion, but you should understand that these terms are not universal (although the definition of construct validity given below is widely accepted).

A good way to determine the validity of a test is to understand what it measures and then decide what measures should and should not be correlated with it. For example, a valid test of memory might have a negligible correlation with a measure of social intelligence, a moderate correlation with a measure of anxiety, and a high correlation with a measure of attention.

Two issues are addressed in validating tests: what a test measures and how well it measures it. Below, we will consider procedures that reflect different strategies for analyzing validity. Recognize that no test is valid for all purposes or valid in the abstract; a test is valid only for a specific purpose. Furthermore, validity is not a matter of all or nothing, but a matter of degree. When you evaluate a test, consider the various lines of evidence that support its validity. Select tests that are valid for your purposes. For example, to select the best applicants for a job, use a test with the best available criterion-related validity for that occupation. Or, to measure achievement, select a test with good content validity. Studies of test validity should continue long after publication of the test. The test publisher is responsible for furnishing evidence that the test is valid for specific purposes, and the examiner is responsible for the appropriate use of test results, for evaluating the publisher's evidence, and for studying subsequent research on the test. Let's now consider various types of validity: content validity, face validity, construct validity, and criterion-related validity.

Content Validity

Content validity refers to whether the items within a test or other measure represent the domain being assessed. In evaluating content validity, we must consider the appropriateness of the type of items, the completeness of the item sample, and the way in which the items assess the content of the domain involved. Questions relevant to these considerations include the following: (a) Does the test measure the domain of interest? (b) Are the test questions appropriate? (c) Does the test contain enough information to cover appropriately what it is supposed to measure? (d) What is the level of mastery at which the content is being assessed? If we can answer these four questions satisfactorily, the test has good content validity. For example, a mathematics test designed for children from ages 6 to 17 years quite likely would have good content validity if the test systematically sampled the material found in several mathematics books used in preschool through beginning college level.

The concept of content validity applies not only to intelligence and achievement tests but also to rating scales, checklists, and observational measures. We might ask, for example, whether the content of a behavioral rating scale designed to measure aggressive behavior actually corresponds to a generally recognized definition of the aggression construct.

We can build content validity into a test by including only items that measure the trait or behavior of interest. Content validity does not require that a test measure all possible elements of a content area, just representative ones. The initial part of the validation process for any educational or psychological test is to determine the representativeness of the test items in the test.

Although some achievement tests are based on a detailed chart of objectives that can be used to assess validity, content validity is usually evaluated through relatively subjective and unsystematic procedures. That is, we examine the content of a measure and attempt to determine whether it corresponds with our understanding of the concept it measures. This is a good starting point in assessing a measure, but more systematic procedures are also required to evaluate a measure's validity; these include assessing construct validity and criterion-related validity.

To define the domain of interest (what is to be measured), test developers may ask experts to nominate items and/or to rate items as to their acceptability and then test these items. Items are administered to a sample and evaluated for such factors as content, clarity, complexity of language, readability level, and cultural and gender bias. Items are then modified as needed and administered to another sample. Items are evaluated again on criteria including their difficulty level (i.e., percentage of examinees passing each item) and their discriminative power (i.e., ability to differentiate between high and low achievers). Discriminative ability is studied by evaluating, for example, whether the proportion of the highest 27% in the sample who answered a particular item correctly is greater than the proportion of the lowest 27% in the sample who answered the item correctly.

Face Validity

Face validity refers to whether a test looks valid "on the face of it." In evaluating face validity, we are asking whether examiners and those taking the test perceive the instrument as a reasonable measure of what it is supposed to measure. This involves judgment, but face validity is important if an individual is to be motivated to participate in the assessment process. For example, employers sometimes run into resistance in employment screening situations because potential employees believe that the assessment tools have no relevance to the job in question. However, face validity is the least important form of validity, because its assessment requires a subjective judgment, does not depend on established theories for support, and may give the respondent a false sense of what the test measures.

Construct Validity

Construct validity establishes the degree to which a test measures a specified psychological construct (i.e., an inferred entity) or trait. For example, what does a score in the gifted range on an intelligence test tell us about the intellectual functioning of the child? Similarly, what does it mean to say that a child has a low or high competence score on a teacher rating measure? What does the score tell us about the child's functioning? These are the kinds of questions that arise in connection with construct validity. Examples of cognitive constructs are intelligence, concept formation, short-term memory, speed of information processing, developmental delay, nonverbal reasoning, and mechanical aptitude.

Two components of construct validity are convergent validity and discriminant validity. *Convergent validity* refers to how well measures of the same domain in different formats—such as tests in multiple-choice, essay, and oral formats—correlate with each other. *Discriminant validity,* sometimes called *divergent validity,* refers to the extent to which measures of different domains do not correlate with each other. Discriminant validity is the flip side of convergent validity. When you assess a test's construct validity, you need to consider both convergent validity and discriminant validity along a continuum.

Although construct validity is important, it is difficult to evaluate because constructs are difficult to define and empirical procedures for evaluating them are limited. Still, we have some useful ways to evaluate how the items in a test relate to the theoretical constructs that the test purports to measure. They include specifying the meaning of the construct, distinguishing the construct from other constructs, and specifying how measures of the construct relate to other variables.

Following are some examples of ways we can obtain evidence for construct validity.

- We find a relationship between test scores and a theory related to how the test items were selected. For example, we can say that a test of intelligence has construct validity if, compared to children who have low scores, children who obtain high scores on the test also have better recall, understanding of concepts, imagination, grades in school, teacher ratings of scholarship, and parental ratings of intelligence.
- We find that scores from one test correlate with related measures. For example, suppose we give a test of leadership quality to a sample of college students, place them in groups of six students, and give each group a task to perform. We then have raters who are unfamiliar with the students' leadership test scores rate each student on his or her leadership qualities. A positive correlation between the test scores and the observers' ratings provides evidence that the test has construct validity.
- We find that scores from a test correlate very highly with related measures (the test has convergent validity) and not highly with unrelated measures (the test has discriminant validity). Thus, for example, when a test of reading correlates very highly with other tests of reading and does not correlate highly with tests of mathematics, we say that the reading test has convergent and discriminant validity.

- We conduct a factor analysis and find that the test measures the constructs underlying it. If we intercorrelate the subtests in a test and conduct a factor analysis, the results will provide information about which subtests share common variance or communality (described below) and thus measure the same construct. For example, suppose a factor analysis of the WISC–IV indicated that the subtests in the test share common variance and that the test has meaningful verbal comprehension, perceptual reasoning, working memory, and processing speed components; this finding would support the use of separate Verbal Comprehension, Perceptual Reasoning, Working Memory, and Processing Speed Composites (see Chapter 9).
- We show that there are developmental changes in scores derived from a measure of a trait or skill by finding increases in magnitude with age or experience. For example, suppose we develop a 20-item vocabulary test with items ordered according to their difficulty level. To do this, we select words from first-, second-, third-, fourth-, fifth-, and sixth-grade reading books. We then test 100 children from first through sixth grades. If the percentage passing each item (i.e., defining words correctly) increases with grade level, we have shown that the test reflects developmental changes.

Criterion-Related Validity

Criterion-related validity is based on how positively test scores correlate with some type of criterion or outcome (such as ratings, classifications, or other test scores). The criterion, like the test, must possess adequate psychometric properties: It should be readily measurable, reliable, and relevant to the purposes of the test. The test and the criterion should have a complementary relationship; otherwise, the criterion could not be used to determine whether the test measures the trait or characteristic it was designed to measure. The two forms of criterion-related validity are concurrent validity and predictive validity.

Concurrent validity is based on correlations of scores on one measure with those on a related measure. To establish concurrent validity, we administer the two measures to the same group of people, one right after the other. We might, for example, administer a measure of phonics ability and a measure of reading ability. If the phonics measure has good concurrent validity, people who obtain high scores on it will also obtain high scores on the measure of reading ability. Likewise, people who obtain low scores on the phonics measure will also obtain low scores on the measure of reading ability. If a measure has low concurrent validity, there will be an erratic and unpredictable relationship between scores on it and scores on the related measure.

Predictive validity is based on correlations of scores on one measure with those on a criterion measure taken at a later time. For example, we might compare scores on a reading readiness test administered at the beginning of the first grade (the predictor measure) to scores on a measure of reading ability administered at the end of the first grade (the criterion measure). If the reading readiness test possesses high predictive validity, children who score high on it will perform well on the later criterion measure. Likewise, those scoring low on the initial test will perform poorly on the later criterion measure. If the predictive validity of a test is low, there will be an erratic and unpredictable relationship between the two sets of scores.

Results from criterion-related validity studies are usually expressed as correlation coefficients. For example, a relationship between a teacher rating measure of social maturity and scores on a standardized social maturity test might be expressed as $r = .53, p < .01$. The correlation of .53 provides us with information about the degree of association between the predictor and the criterion, and the confidence index (p value) tells us that there is less than 1 chance in 100 of obtaining an association of that magnitude by chance (given a particular number of observations). Applying the formula for effect size given earlier, we find that $d = 1.25$, which is a moderate effect for the predictive association.

Predictive Power

Predictive power is a special type of predictive validity. It assesses the accuracy of a decision made on the basis of a given measure. Thus, predictive power refers the extent to which a test (or another measure, such as a rating scale or an observation form) agrees with an outcome criterion measure used to classify individuals in a particular category or to determine whether or not they have a particular trait or condition. For example, suppose a preschool inventory (the test criterion) is administered to a group of children at 5 years of age. The cut-off score selected by the investigator for classifying children as "at risk" for reading problems is the 15th percentile rank. Those falling at or below the 15th percentile rank are assigned to the "at risk" category, and those falling above the 15th percentile rank are assigned to the "not at risk" category. Three years later, at the end of the third grade, the children are given an achievement test (the outcome criterion). The investigator again selects the 15th percentile rank as the cut-off score for determining which children should be classified as having reading problems. The predictive power of the preschool inventory administered to the 5-year-old children is determined by how well the inventory predicts categorization based on the achievement test. For screening instruments in particular, it is valuable to have information about both predictive validity and predictive power.

All predictions must be compared to the *base rate* of a condition, an attribute, or a disease in a specific population. Base rates are important, because they are the rates against which we judge the accuracy of a prediction. The utility of a measure depends on whether it improves predictions beyond what would be expected from predictions using base rates alone. For example, if the base rate of a condition is 90%, we could be 90% accurate by simply predicting the presence of the condition for every person. Or, if the base rate of a condition is 1%, we could be 99% accurate by simply predicting the absence of the condition each time. When base rates are either very high or very low, the accuracy of predictions using base rates alone is high. When a base rate nears 50%, the accuracy of predictions using the base rate alone can potentially be improved greatly by using a relevant measure. The further away the base rate gets from 50%, the more difficult it becomes to develop measures that will increase the accuracy of predictions.

We compute the predictive power of a test by determining the percentages of correct and incorrect classifications that it makes. To do this, we might assign individuals to either an "at risk" or a "not at risk" category based on their test scores, and to a "poor outcome" or a "good outcome" category based on their scores on an outcome criterion measure. As in the above example, let's choose the 15th percentile rank for the test criterion and the outcome criterion. We can depict the results in a 2 x 2 matrix, as shown in Figure 4-5.

The four cells in the matrix represent the following types of agreement (alternative terminology for agreement type is shown in parentheses):

(a) *True positive (hit).* The test classified the child as being at risk of having a poor outcome (referred to as the positive classification); the outcome criterion measure indicated that the child actually did have a poor outcome. Thus, the outcome criterion measure confirmed the way the test classified the child. *Positive* here means that the child is classified as being at risk for having problems (or a poor outcome). In medicine, a true positive result occurs when a diagnostic test returns a positive result (indicating that a condition is present) and the condition is in fact present.

(b) *False positive (false alarm).* The test classified the child as being at risk for having a poor outcome; however, the child had a good outcome on the outcome criterion measure. Thus, the outcome criterion measure disconfirmed the way the test classified the child. In medicine, a false positive result occurs when a diagnostic test returns a positive result (indicating that a condition is present) but in fact the condition is not present.

(c) *False negative (miss).* The test classified the child as not being at risk for having a poor outcome (referred to as the negative classification); however, the child had a poor outcome on the criterion outcome measure. Thus, the outcome criterion measure disconfirmed the way the test classified the child. *Negative* here means that the child is classified as not being at risk for having problems (or a poor outcome). In medicine, a false negative result occurs when the diagnostic test returns a negative result (indicating that a condition is not present) but the condition is in fact present.

(d) *True negative (correct rejection).* The test classified the child as not being at risk for having a poor outcome; the outcome criterion measure indicated that the child actually did have a good outcome. Thus, the outcome criterion measure confirmed the way the test classified the child. In medicine, a true negative result occurs when a diagnostic test returns a negative result (indicating that a condition is not present) and the condition is in fact not present.

Combinations of individual cells in Figure 4-5 provide the following 10 different measures of predictive power.

1. *True positive rate, a*/(*a* + *c*). The true positive rate reflects the probability that a test correctly identifies people who will have a poor outcome. This is the rate at which people predicted by the test to have a poor outcome in fact did have a poor outcome. It is also referred to as the *index of sensitivity,* the *valid positive rate,* or the *hit rate.*

2. *False positive rate, b*/(*b* + *d*). The false positive rate reflects the probability that a test incorrectly identifies people who will have a poor outcome. This is the rate at which people

		Outcome criterion		
		Poor outcome	**Good outcome**	**Total**
Test criterion	**At risk**	True positive (hit) (*a*)	False positive (false alarm) (*b*)	*a* + *b*
	Not at risk	False negative (miss) (*c*)	True negative (correct rejection) (*d*)	*c* + *d*
	Total	*a* + *c*	*b* + *d*	*a* + *b* + *c* + *d* = *N*

Figure 4-5. Model for assessing the predictive utility of a test.

predicted by the test to have a poor outcome instead had a good outcome. It is also referred to as the *false alarm rate.*

3. *False negative rate, c/(a + c).* The false negative rate reflects the probability that a test incorrectly identifies people who will have a poor outcome. This is the rate at which people predicted by the test to have a good outcome instead had a poor outcome. It is also referred to as the *miss rate* or the *underreferral rate.*

4. *True negative rate, d/(b + d).* The true negative rate reflects the probability that a test correctly identifies people who will have a good outcome. This is the rate at which people predicted by the test to have a good outcome did in fact have a good outcome. It is also referred to as the *index of specificity,* the *valid negative rate,* or the *correct rejection rate.*

5. *Positive predictive power, a/(a + b).* The positive predictive power reflects the proportion of people whom the test correctly identified as being at risk for having a poor outcome. It is also referred to as the *efficiency rate.*

6. *Negative predictive power, d/(c + d).* The negative predictive power reflects the proportion of people whom the test correctly identified as not being at risk for having a poor outcome.

7. *Overall accuracy rate, (a + d)/N.* The overall accuracy rate reflects the proportion of people in the total sample whom the test correctly identified as being either at risk (true positive) or not at risk (true negative) for having a poor outcome. It is also referred to as the *overall hit rate,* the *correct classification rate,* the *observed proportion of overall agreement,* or the *effectiveness rate.* Although useful and informative, the overall accuracy rate does not distinguish between the number of true positive ratings and the number of true negative ratings.

8. *Overall inaccuracy rate, (b + c)/N.* The overall inaccuracy rate reflects the proportion of people in the total sample whom the test incorrectly identified as being either at risk (false positive) or not at risk (false negative) for having a poor outcome. It is also referred to as the *overall error rate,* the *incorrect classification rate,* the *observed proportion of overall disagreement,* or the *misclassification rate.* Although useful and informative, the overall inaccuracy rate does not distinguish between the number of false positive ratings and the number of false negative ratings.

9. *Base rate, (a + c)/N.* The base rate reflects the proportion of people in the total sample who had a poor outcome. It is also referred to as the *prevalence rate* or the *true proportion.*

10. *Odds ratio, ad/bc.* The odds ratio is the ratio of the odds of individuals with a poor outcome being identified as at risk to the odds of individuals with a good outcome being identified as at risk. The odds ratio provides an index that is not influenced by the base rate of individuals with a poor outcome.

Table 4-8
Different Measures of Predictive Power

Measure	*Calculation*
True positive rate (index of sensitivity)	*a*/(*a* + *c*)
False positive rate (false alarm rate)	*b*/(*b* + *d*)
False negative rate (miss rate)	*c*/(*a* + *c*)
True negative rate (index of specificity)	*d*/(*b* + *d*)
Positive predictive power (efficiency rate)	*a*/(*a* + *b*)
Negative predictive power	*d*/(*c* + *d*)
Overall accuracy rate (overall hit rate)	(*a* + *d*)/*N*
Overall inaccuracy rate (overall error rate)	(*b* + *c*)/*N*
Base rate	(*a* + *c*)/*N*
Odds ratio	*ad*/*bc*

Table 4-8 summarizes the 10 different measures of predictive power.

Let's look at how to compute the overall accuracy rate, the overall inaccuracy rate, and the base rate. If the four cells had the frequencies $a = 45$, $b = 15$, $c = 5$, and $d = 35$, these rates would be as follows:

$$\text{Overall accuracy rate} = \frac{45 + 35}{45 + 15 + 5 + 35} = .80, \text{ or } 80\%$$

$$\text{Overall inaccuracy rate} = \frac{15 + 5}{45 + 15 + 5 + 35} = .20, \text{ or } 20\%$$

$$\text{Base rate} = \frac{45 + 5}{45 + 15 + 5 + 35} = .50, \text{ or } 50\%$$

We can measure whether a test adds to predictive accuracy by determining whether the ratio of the base rate of the poor outcome (the rate of occurrence) to the base rate of the good outcome (the rate of nonoccurrence) exceeds the ratio of the rate of false positives to the rate of true positives. Using the labels in Figure 4-5, this relationship can be expressed as $a/d > b/a$. For the frequencies in the previous example ($a = 45$, $b = 15$, $c = 5$, $d = 35$), the relationship is as follows:

$$\begin{aligned}\text{Increase in predictive accuracy} &= 45\%/35\% \text{ vs. } 15\%/45\% \\ &= 1.29 \text{ vs. } .33\end{aligned}$$

Because 1.29 is considerably greater than .33, using a test with the indicated frequencies would lead to more correct decisions than merely following the base rate predictions. That is, the test adds to predictive accuracy.

Factors Affecting Validity

Validity coefficients are affected by the same factors that affect correlation coefficients, as well as by other factors such as the following:

1. *Range of attributes being measured.* Narrowing the range of scores of either the test or the criterion measure will reduce the size of the validity coefficient; this is referred to as restriction of range. For example, math achievement test scores would have a higher correlation with intelligence test scores in

a general population sample than in a sample composed of only children who are gifted or children with mental retardation.

2. *Length of the interval between administration of the test and of the criterion measure.* Lengthening the time interval tends to lower the size of the validity coefficient.

3. *Range of variability in the criterion measure.* If there were no variability in the criterion measure used to assess the validity of an intelligence test (e.g., all students obtained 90% accuracy on the achievement test), the validity coefficient for the intelligence test would be zero; however, this would be a poor test of validity. We cannot say that an intelligence test is not valid when the achievement test scores have no variability—it is a case of trying to predict the unpredictable or of trying to predict differences where none exist. What is needed in order to find out whether the intelligence test is valid is a more heterogeneous sample. However, there are also instances when the criterion group may be too heterogeneous. For example, if we administer the criterion measure to a group that is more heterogeneous than the population for which a test in intended, validity estimates will be spuriously (falsely) high. Suppose we use a random sample of school children to validate a test of artistic ability that is designed to screen children nominated by their teachers as showing artistic talent; the random sample will be more heterogeneous than the group for whom the test was originally intended (i.e., children nominated for having artistic talent). The resulting validity coefficient is likely to be spuriously high, showing that the test has good discrimination (i.e., that it differentiates children who have artistic ability from those who do not). We can determine the amount of overestimation by comparing the validity coefficient obtained by using the random sample with the one obtained by using a sample of children nominated for their artistic talent.

Judging the Validity of an Individual Child's Test Scores

The validity of a child's test scores can be affected by such factors as the child's test-taking skills, anxiety, fatigue, transient medical conditions, confusion, limited attention, degree of rapport with the examiner, motivation, speed, understanding of test instructions, physical handicaps, temporary hearing impairments, language skills, educational opportunities, and familiarity with the test material. Deficiencies in any of these areas will decrease validity. Thus, for example, test results are not valid when children are uncooperative or highly distractible, when they don't understand the test instructions or the wording of the test questions, when they have physical handicaps that interfere with their ability to take the tests (and no adjustments have been made by the examiner), or when they have limited comprehension of English.

Validity can also be affected by intervening events and contingencies. You will need to consider everything you know about a child in evaluating different types of validity. For example, does an emotionally disturbed child have an acute or a chronic condition? An acute disturbance might lower his or her performance on an intelligence or achievement test, resulting in nonrepresentative test results. If an intervention—such as drugs, psychotherapy, placement in a foster home, or environmental manipulation—improves the child's performance, the validity or representativeness of the initial test results is likely questionable. However, if a child has a chronic condition, such as irreversible brain damage or an autistic disorder, his or her test results may not be invalid, because in such cases the child's level of ability may not change over time.

Deficiencies in the robustness of the criterion might affect the validity of tests. For example, achievement test scores, a popular criterion, may be affected by the quality of the teaching, of textbooks, and/or of the curriculum. Scores also might be affected by the children's levels of ability, effort, classroom behavior, study skills, relationships with teachers and peers, and home environment (e.g., parent encouragement, study facilities, and resources in the home such as a computer and access to the Internet).

If you have any reason to question the validity of test results (even though you have used a psychometrically sound test), state your reservations in the psychological report. And if you seriously question the validity of the results, consider destroying the test protocol or writing *Invalid* on the face sheet. The fact that a child deviates from some earlier level of functioning may not invalidate the results—his or her current level of functioning may be different from the earlier level. In some cases, you may need to estimate the earlier level of functioning based on prior test results, school grades, or parental reports. In cases of brain injury, the earlier level of functioning is referred to as the *premorbid* (or preinjury) *level*—that is, the level at which the child was functioning prior to the brain injury.

META-ANALYSIS

A single study seldom provides definitive answers to research questions. Instead, scientific progress is achieved through the accumulation of findings from numerous studies on a particular issue. Traditionally, researchers relied on narrative literature reviews to help them arrive at generalizations. However, these reviews were often flawed: Narrative reviews of the same body of research sometimes led to different conclusions because of subjective judgments, preferences, and reviewer bias.

Meta-analysis is an alternative to the narrative literature review and avoids many of its flaws. It summarizes the results of many studies. Meta-analysis uses rigorous research techniques (including quantitative methods) to sum up and integrate the findings of a body of studies covering similar topics. Because the individual studies reviewed are likely to have used different statistical techniques, meta-analysis uses a standard measure of effect size (usually Cohen's *d* or *r*, discussed in this chapter). Researchers have successfully applied meta-analysis to studies in the social, behavioral, and biomedical sciences.

Meta-analysis is particularly useful in validity generalization studies. Researchers examine a large number of studies that present evidence on the validity of a particular test. The empirical findings from these validity studies (e.g., validity coefficients and scores showing between-group differences) are converted to a common metric and then evaluated for consistency (i.e., generalizability or robustness) across different populations, test conditions, criterion measures, and the like. Findings from meta-analyses highlight trends in data and inform researchers and practitioners about the validity of the test or other measure under study. Although meta-analysis has many potential benefits and is widely used to synthesize research findings, its conclusions may be compromised by the variety of studies reviewed and their shortcomings, such as poor design and inadequate sampling.

FACTOR ANALYSIS

Factor analysis is a mathematical procedure used to explain the pattern of intercorrelations among a set of variables (such as individual test items, entire tests, subtests, or rating scales) by deriving the smallest number of meaningful variables or factors. A *factor* is a statistically derived, hypothetical dimension that accounts for part of the intercorrelations among a set of variables. The aim of factor analysis is to explain the pattern of intercorrelations by identifying the smallest number of meaningful underlying variables or factors that could account for the observed intercorrelations. Identifying the minimum number of factors reduces a mass of information to more manageable proportions and is more economical than proposing a different factor to explain every correlation.

Factor analysis is also used to delineate patterns in a complex set of data (see Chapter 9), to discover the basic structure in a data set, to develop an empirical typology, to develop scales and weight factors in the scales, to test hypotheses, to transform data, to explore new relationships in a data set, and to construct theories (see Chapter 7). In constructing and evaluating psychological tests and measures, factor analysis focuses on the number of factors needed to explain the pattern of relationships among the variables, the nature of the factors, how well the hypothesized factors explain the observed data, and how much purely random or unique variance each observed variable includes.

Factor analysis is based on the assumption that a significant correlation between two variables indicates a common underlying factor shared by both variables. Factor analysis starts with a correlation matrix that shows the intercorrelations between several variables (see Table 4-9). Intercorrelations are the correlations between all variables in the matrix. For instance, if there are four variables in the matrix, the correlations would be between *a* and *b*, *a* and *c*, *a* and *d*, *b* and *c*, *b* and *d*, and *c* and *d*.

The first step in a factor analysis is to calculate the factor loading of each variable on each factor, which reflects the extent to which each variable "loads" on the factor (see the group factors in Table 4-10). *Factor loadings* are simply the correlation coefficients between variables and factors. The loadings indicate the weight of each factor in determining performance on each variable.

The next step is to name each factor. For example, suppose a factor shows high loadings for variables involving vocabulary, information, and knowledge of word similarities. The theoretical factor underlying these three subtests, which is assumed to be explained by a higher-order factor, might be called "verbal ability." Some variables may load on more than one factor, and some variables may have minimal loadings on the factors. Different investigators and test publishers might use different names for the same factor. For example, one investigator might call a factor "verbal ability"; another might use the term "lexical knowledge" or "crystallized intelligence." Or, one investigator might label a factor "verbal comprehension," whereas another might identify it as "receptive oral language."

Factors, like the variables from which they are derived, only describe the relationships observed in the data. There is no implication that the observed scores are somehow caused by the factors or vice versa. Factors do not represent underlying causal entities.

The two major types of factor analysis are exploratory factor analysis and confirmatory factor analysis. An *exploratory factor analysis* (EFA) is used to explore the underlying structure of a collection of variables when there are no a priori hypotheses about the factor structure. A *confirmatory factor analysis* (CFA) is used to confirm a hypothesized factor structure. The variables for a confirmatory factor analysis are selected on the basis of prior theory.

Methods Used in Factor Analysis

There are different methods for extracting factors. Two common ones are principal component analysis (PCA) and principal factor analysis (PFA). When there are many factors, the results of the two methods are somewhat similar. *Principal component analysis* seeks the set of factors that can account for all common and unique variance in a set of variables. In contrast, *principal factor analysis,* which incorporates prior communality estimates, seeks the smallest set of factors that can account for the common variance in a set of variables.

Most factor analysis programs begin by extracting first the factor that accounts for the largest proportion of variance, then the factor that accounts for the next largest proportion, and so on. Usually, the *first unrotated factor* is a general factor on which most variables have high loadings. We find a *general factor*—a factor on which all the variables load—when all subtests overlap (e.g., are positively intercorrelated), such as in an intelligence test. In an intelligence test, the first general factor is considered to reflect general intelligence, called *g*. In other cases, such as in a multidimensional test of personality, there may be two or three important personality factors but no single personality factor on which all variables load.

Rather than attempting to interpret the original factors, however, researchers usually rotate the matrix of factor loadings to make the factor structure clearer. The rotation rearranges the factors so that, ideally, for every factor there are some variables with high loadings on the factor and other variables with low loadings on the factor. The order in which the factors were originally extracted is not always preserved in the rotation; in particular, researchers usually cannot discern the first unrotated factor. One popular type of rotation is *varimax rotation,* in which the factors are *orthogonal*—that is, neither overlapping nor correlated. Another popular type of rotation is *oblim rotation,* in which the factors are allowed to be correlated. We call the factors resulting from the rotation *group factors.* It is up to the researcher or test developer to name or interpret each factor by looking at the contents of the variables that have high loadings on the factor.

After all of the common factor variance has been extracted and the rotation completed, there still may be a significant amount of unanalyzed variance. Variance that is present in one variable but not in the other variables under study is known as *specific factor variance, specific variance,* or *specificity.*

Components of Variance

In a factor analysis, we can divide the variance associated with a variable into three categories: communality, specificity, and error variance.

Communality. Communality refers to that part of the total variance that can be attributed to common factors (those that appear in more than one variable). The formula for obtaining communality is as follows:

$$h_t^2 = a_{t1}^2 + a_{t2}^2 + \cdots + a_{tm}^2$$

where h_t^2 = communality of test t

$a_{t1}^2, \ldots, a_{tm}^2$ = loading of test t on factor 1, ..., factor m

For the WISC–IV data in Table 4-10 in the next part of the chapter, the communality estimate for the Similarities subtest is

$$h_t^2 = .73^2 + .09^2 + .04^2 + .01^2 = .54$$

Specificity. Specificity refers to that part of the total variance that is due to factors specific to a particular variable, not to measurement error or common factors. We obtain the proportion of specific variance in the following way:

$$s_t^2 = r_{tt} - h_t^2$$

where s_t^2 = variance specific to test t

r_{tt} = reliability of test t

h_t^2 = communality of test t

The proportion of specific variance for the WISC–IV Similarities subtest (see Table 4-10) is

$$s_t^2 = .86 - .54 = .32$$

Error variance. Error variance refers to that part of the total variance that remains when we subtract the reliability of the variable from the total variance. We obtain it by using the following formula:

$$e_t^2 = 1 - r_{tt}$$

where e_t^2 = error variance of test t

r_{tt} = reliability of test t

Error variance for the Similarities subtest is

$$e_t^2 = 1 - .86 = .14$$

When specific variance exceeds error variance, we can conclude that the variable has some specificity. In the example above, we conclude that Similarities has adequate specificity. This means that Similarities measures a specific construct not measured by other subtests.

Illustration of Factor Analysis

Let's examine how we might apply factor analysis to the WISC–IV. Table 4-9 shows a partial set of WISC–IV subtest intercorrelations (for 4 of the 15 WISC–IV subtests). These correlations are based on the entire standardization group (N = 2,200). If the WISC–IV measures general intellectual ability, children with an abundance of this ability should perform well on each of the subtests and those with a small amount of this ability should do poorly. With respect to the intercorrelations in Table 4-9, this means that children who do well on Similarities should also do well on Vocabulary and, to a somewhat lesser degree, on Block Design and Picture Concepts. In contrast, those who do poorly on Similarities should also do poorly on Vocabulary and, to a lesser degree, on the other two subtests. If children's scores on the four subtests are highly correlated, we can reasonably conclude that the four subtests measure something in common.

Subtests correlate with each other to different degrees. When specific abilities are more pronounced than general or group abilities, the correlations among subtests should be lower. Since the correlations in Table 4-9 are moderate to strong, we might conclude that there is a general ability factor in these four subtests of the WISC–IV. Something more than a general factor may be present when the correlations are not consistently high—when some abilities are important for some subtests but not for others.

The factor analytic findings for the entire WISC–IV are discussed in Chapter 9. They indicate that both a general factor and group factors are present in the test. Additionally, several subtests have adequate subtest specificity. Table 4-10 shows the median general factor and group factor loadings, reliability, communality, specificity, and error variance for the Similarities, Vocabulary, Block Design, and Picture Concepts subtests. Loadings of .70 and above on the general

Table 4-9
Average Intercorrelations for Four WISC–IV Subtests

Subtest	*SI*	*VC*	*BD*	*PCn*
SI	—	.74	.50	.50
VC	.74	—	.48	.42
BD	.50	.48	—	.41
PCn	.50	.42	.41	—

Note. Abbreviations: SI = Similarities, VC = Vocabulary, BD = Block Design, PCn = Picture Concepts.
Source: Adapted from Wechsler (2003b, p. 51).

factor are considered substantial, as are loadings of .30 or .40 and above on the group factors. The loadings indicate that Similarities, Vocabulary, and Block Design are good measures of the general factor and that Picture Concepts is a fair measure of the general factor. Additionally, Similarities and Vocabulary load highly on the Verbal Comprehension group factor, Block Design loads highly on the Perceptual Reasoning group factor, and Picture Concepts loads moderately on the Perceptual Reasoning group factor. None of these four subtests loads highly on either the Working Memory group factor or the Processing Speed group factor. Of the four subtests, the Similarities, Block Design, and Picture Concepts subtests have adequate specificity, because specific variance (specificity) exceeds error variance on these three subtests. In contrast, Vocabulary does not have adequate specificity, because error variance exceeds specific variance.

Comment on Factor Analysis

Factor analysis is a complex statistical method. The same set of data can yield different results depending on the factor analytic method used, the number of factors retained, and the rotations of the factors. In addition, the naming of factors is arbitrary, as noted earlier in the chapter. Thus, although factor analysis is a useful procedure, results obtained from it must be interpreted very carefully.

OTHER USEFUL PSYCHOMETRIC CONCEPTS

Occasionally, you will find that two or more tests believed to measure the same ability give different results for the same child. Different results might occur, for example, because of characteristics of the child, testing conditions, examiner characteristics, or the psychometric properties of the tests. Chapter 6 discusses the first three issues in more detail. Here we will discuss how the psychometric properties of two supposedly similar tests might lead to different results (Bracken, 1987, 1988; Wasserman & Bracken, 2002):

1. *Floor effect differences.* The lower limits of scores may differ on different tests. The test floor is the lowest possible score obtainable on a test. Floor effects thus refer to the number of easy items available at the lowest level of a test to distinguish among children with below-average ability. You need to consider the test floor because it indicates how well the instrument can discriminate among children in the lower ranges of functioning; it tells you which populations can and cannot validly be tested with the instrument. You also need to consider whether the test floor is relevant to actual practice. If it isn't, the test scores should be questioned.

Let's see how floor effects operate on the WISC–IV. The lowest possible Full Scale IQ obtainable on the WISC–IV is 40 (see WISC–IV Administration Manual, Table A.6, p. 239). Thus, the WISC–IV does not provide IQs for children functioning more than four standard deviations below the mean of the test (which is 100). It is also important to consider the subtest floors, especially for profile analysis (i.e., comparing profiles of subtest scaled scores). For example, for a raw

Table 4-10
General Factor Loadings, Group Factor Loadings, Reliability, Communality, Specific Variance, and Error Variance for Four WISC–IV Subtests

		Group factors							
WISC–IV subtests	*General factor*	*Factor A, Verbal Comprehension*	*Factor B, Perceptual Reasoning*	*Factor C, Working Memory*	*Factor D, Processing Speed*	*Reliability* (r_{tt})	*Communality* (h_t^2)	*Specificity* (s_t^2)	*Error* (e_t^2)
SI	.81	.73	.09	.04	.01	.86	.54	.32	.14
VC	.83	.90	.02	–.02	–.01	.89	.81	.08	.11
BD	.70	–.02	.70	.08	.08	.86	.50	.36	.14
PCn	.61	.15	.35	.13	.07	.83	.17	.66	.17

Note. Abbreviations: SI = Similarities, VC = Vocabulary, BD = Block Design, PCn = Picture Concepts.

score of 1, subtest floors for the 15 WISC–IV subtests for the age group 6 years, 0 months (6-0) to 6 years, 3 months (6-3) vary between scaled scores of 1 and 4 (see WISC–IV Administration Manual, Table A.1, p. 204). Therefore, when you examine a WISC–IV profile of a child who is functioning at the lower levels of the test, you must consider the available range of subtest scaled scores. If you are concerned that WISC–IV subtest floors are not adequate, consider using another instrument that allows lower test scores.

2. *Ceiling effect differences.* The upper limits of scores may differ on different tests. Analogous to the test floor, the test ceiling is the highest possible score on a test. Ceiling effects thus refer to the number of difficult items available at the highest level of a test to distinguish among children with above-average ability. You need to consider the test ceiling because it indicates how well the instrument discriminates among children in the upper ranges of functioning and which populations can and cannot validly be tested with the instrument. You also need to consider whether the test ceiling is relevant to actual practice. If it isn't, the test scores should be questioned.

Let's now look at ceiling effects on the WISC–IV. The test ceiling of the WISC–IV Full Scale IQ is 160 (see WISC–IV Administration Manual, Table A.6, p. 240). This indicates that the WISC–IV does not provide IQs for individuals functioning more than four standard deviations above the mean of the test. As with subtest floors, knowledge of subtest ceilings is important for profile analysis. Subtest ceilings on the WISC–IV show little variability—they are at a scaled score of 19 for all subtests except Word Reasoning. On Word Reasoning, the highest scaled score is 18 at ages 14-0 years to 15-11 years and 17 at ages 16-0 years to 16-11 years. Therefore, when you examine a WISC–IV profile of a child who is functioning at the upper limits of the test, you can compare subtests using essentially the same range of scaled scores. However, if you are concerned that the WISC–IV subtests have a limited ceiling in a particular assessment, consider using an instrument that has higher test ceilings.

3. *Item gradient differences.* Item gradients may differ on different tests. Item gradients refer to the ratio of item raw scores to standard scores, or the number of raw score points required to earn 1 standard score point. In other words, item gradients help us see "how rapidly standard scores increase as a function of a child's success or failure on a single test item" (Bracken, 1987, p. 322). Item gradients tell us how steeply items are arranged within a test. Tests with steep gradients (that is, tests in which the difficulty level of items changes rapidly, so a change of a single raw score point produces a large change in the standard score) are less sensitive to small or moderate differences in ability or skill development than are tests with gradual gradients. This means that tests with steep gradients are less effective in assessing a child's abilities or skills than tests with more gradual gradients (Bracken, 1987).

Now let's look at both the WISC–IV and the WPPSI–III to see how item gradients operate. We will use the Block Design subtest as an illustration. For the WISC–IV and WPPSI–III Block Design subtest, Table 4-11 shows the raw scores required to earn scaled scores of 1 to 6 (obtained from the WISC–IV Administration Manual, Table A.5, p. 204 and from the WPPSI–III Administration Manual, Table A.1, p. 227). On the WISC–IV, a 6 year, 0 month child with one correct answer obtains a scaled score of 2 and a child with three correct answers obtains a scaled score of 4. This means that the WISC–IV Block Design subtest has a gradual gradient and discriminates well among children who are functioning at the low end of the subtest. The WPPSI–III shows a similar pattern of item gradients, beginning with a scaled score of 2. Every one-point or two-point increase in raw scores results in an increase of one scaled score.

Table 4-11
Scaled-Score Equivalents of Raw Scores on the WPPSI–III and WISC–IV Block Design Subtest for a 6-0-Year-Old Child

WPPSI–III raw score	*WISC–IV raw score*	*Scaled score*
0–13	0	1
14–15	1	2
16–17	2	3
18–19	3	4
20	4	5
21–22	5	6

Source: Adapted from Wechsler (2002a, 2003a).

4. *Norm table layout differences.* Norm tables may have different age-span layouts on different tests. For example, age-span layouts may be in 1-month, 3-month, or 4-month intervals; these differences may lead to divergent scores on different tests for the same ages.

5. *Age-equivalent or grade-equivalent score differences.* Age-equivalent or grade-equivalent scores on different tests may not coincide, even though the standard scores are similar on the two tests.

6. *Reliability differences.* Tests with low reliability will produce less stable scores than tests with high reliability.

7. *Differences in skill areas assessed.* Different tests may measure different skills, even though they have the same label for a skill area (e.g., "reading"). One test may measure word recognition (i.e., simply reading the word aloud), whereas another test may measure reading comprehension (i.e., understanding what one reads).

8. *Test content differences.* Different tests may measure the same skill area but contain different content. For example, tests measuring arithmetic may sample different arithmetical principles or concepts.

9. *Publication date differences.* Tests published in different years may yield scores that differ because of changes in the abilities of the norm groups.

10. *Sampling differences.* Tests normed on different samples may yield different scores because the samples are not comparable. For example, one sample might contain more educated people than another, which would tend to make the average score of that sample higher.

The above considerations indicate that you must carefully study the psychometric properties of each test instrument you consider using. You must also pay attention to psychometric properties when you compare the results from two or more tests.

CONCLUDING COMMENT

Despite all the effort devoted to developing reliable and valid assessment instruments, all such instruments have their limitations. Keep in mind the following:

- No instrument is completely reliable (i.e., without error).
- Validity does not exist in the abstract; it must be anchored to the specific purposes for which the instrument is used.
- Every child's behavior fluctuates from time to time and from situation to situation (e.g., a child might perform differently with different examiners).
- Any assessment instrument contains only a sample of all possible questions or items related to the domain of interest.
- Assessment instruments purporting to measure the same construct may give different results for a particular child.
- Instruments measure samples of behavior or constructs at one point in time.
- Assessment scores will likely change to some degree over the course of a child's development.

THINKING THROUGH THE ISSUES

1. Even though you will seldom compute standard deviations and carry out significance tests when you administer and score assessment measures, you will often use standard scores and other statistical concepts to interpret results. How will knowledge of statistics and psychometric concepts be useful to you as a clinician?
2. Before you use a measure, how important is it that you become familiar with its reliability, validity, and standardization?
3. Under what circumstances would you use measures that have minimal reliability or validity?

SUMMARY

The Why of Psychological Measurement and Statistics

1. Measurement in psychology is usually different from physical measurement.
2. In our everyday experience, we assign numbers to the physical characteristics of objects—such as height, weight, or length—that we perceive directly.
3. Although physical measurement may be more precise than psychological measurement because psychological characteristics are likely to be intangible, both types of measurement are important.
4. Psychological measurement conveys meaningful information about people's attributes, such as their intelligence, reading ability, adaptive behavior, interests, personality traits, and attitudes, through test scores or ratings that reflect such attributes.
5. Statistics make life easier by reducing large amounts of data to manageable size, allowing us to study individuals and groups.
6. Statistics also help us communicate information about test scores, draw conclusions about those scores, and evaluate chance variations in test scores.
7. Remember that test scores are imperfect and statistics help us determine the amount of error in test scores.
8. Measurement is a process of assigning quantitative values to objects or events according to certain rules.

Scales of Measurement

9. A scale is a system for assigning values or scores to some measurable trait or characteristic.
10. The four scales most commonly used in psychology and education are nominal, ordinal, interval, and ratio scales.
11. Nominal and ordinal scales (referred to as lower-order scales) are used with discrete variables. Discrete variables are characterized by separate, indivisible categories, with no intermediate values (e.g., gender, color, or number of children in a family).
12. Interval and ratio scales (referred to as higher-order scales) are used with continuous variables. Continuous variables are characterized by an infinite number of possible values of the variable being measured (e.g., temperature, age, or height). Interval and ratio scales possess all the properties of nominal and ordinal scales but have additional properties.
13. A nominal measurement scale consists of a set of categories that do not have a sequential order and that are identified by a name, number, or letter for each item being scaled. The names, numbers, or letters usually represent mutually exclusive categories, which cannot be arranged in a meaningful order and are merely labels or classifications.
14. An ordinal measurement scale classifies items, but it has the additional property of order (or magnitude). The variable being measured is ranked or ordered along some dimension, without regard for the distances between scores.
15. An interval measurement scale classifies, as a nominal scale does, and orders, as an ordinal scale does, but it adds an arbitrary zero point and equal units between points.
16. A ratio measurement scale has a true zero point, has equal intervals between adjacent units, and allows ordering and classification. Because there is a meaningful zero point, there is true equality of ratios between measurements made on a ratio scale.

Descriptive Statistics

17. Descriptive statistics summarize data obtained about a sample of individuals.
18. Examples of descriptive statistics are frequency distributions, normal curves, standard scores, measures of central tendency, and measures of dispersion, correlation, and regression.

19. Measures of central tendency identify a single score that best describes the scores in a data set.
20. The three most commonly used measures of central tendency are the mean, the median, and the mode.
21. The mean is the arithmetic average of all the scores in a set of scores.
22. The median is the middle point in a set of scores arranged in order of magnitude.
23. The mode is the score that occurs most frequently in a set of scores.
24. Dispersion refers to the variability of scores in a set or distribution of scores.
25. The three most commonly used measures of dispersion are the range, the variance, and the standard deviation.
26. The range is the difference (or distance) between the highest and lowest scores in a set; it is the simplest measure of dispersion.
27. The variance is a measure of the amount of variability of scores around the mean—the greater the variability, the greater the variance.
28. The standard deviation is also a measure of how much scores vary, or deviate, from the mean.
29. The normal curve is a frequency distribution that, when graphed, resembles a bell-shaped curve.

Correlation

30. Correlation coefficients tell us about the degree of relationship between two variables, including the strength and direction of their relationship.
31. The strength of the relationship is expressed by the absolute magnitude of the correlation coefficient.
32. Correlations are used in prediction.
33. The higher the correlation between two variables, the more accurately we can predict the value of one variable when we know the value of the other variable.
34. Variables can be related linearly or curvilinearly.
35. A linear relationship between two variables can be portrayed by a straight line.
36. A curvilinear relationship between two variables can be portrayed by a curve.
37. If two variables have a curvilinear relationship, a linear correlation coefficient will underestimate the true degree of association.
38. Variables can also be continuous or discrete.
39. A continuous variable is divisible into an infinite number of parts.
40. A discrete variable has separate, indivisible categories.
41. A dichotomous variable is a discrete variable that has two possible values.
42. A scatterplot presents a visual picture of the relationship between two variables.
43. The most common correlation coefficient is the Pearson correlation coefficient, symbolized by *r*.
44. Pearson's *r* should be used only when the following conditions are met: (a) The two variables are continuous and normally distributed, (b) there is a linear relationship between the variables, and (c) the predictor variable predicts as well at the high-score ranges as at the low-score ranges.
45. When the conditions for using Pearson's *r* cannot be met, the Spearman r_s (rank-difference) method can be used.
46. When the sample size is large, a correlation coefficient may be statistically significant but reflect only a weak association between the two variables.
47. Sometimes test publishers (or researchers) attempt to minimize the effect of measurement error by correcting for attenuation.
48. This correction results in an estimate of what the correlation between two variables would be if both variables were perfectly reliable.
49. However, an estimated *r* based on a correction for attenuation may not give a true picture of the relationship between the variables, because variables are never perfectly reliable.
50. Correlations should not be used to infer cause and effect.
51. When we want to know how much variance in one variable is explained by its relationship to another variable, we must square the correlation coefficient. The resulting value, r^2, is known as the coefficient of determination.

Regression

52. You can use the correlation coefficient, together with other information, to construct a linear equation for predicting the score on one variable when you know the score on another variable.
53. A measure of the accuracy of the predicted *Y* scores in a regression equation is the standard error of estimate. The standard error of estimate is the standard deviation of the error scores, a measure of the amount by which the observed or obtained scores in a sample differ from the predicted scores.

Multiple Correlation

54. Multiple correlation is a statistical technique for determining the relationship between one variable and two or more other variables.
55. The symbol for the coefficient of multiple correlation is *R*.

Norm-Referenced Measurement

56. In norm-referenced measurement, a child's performance on a test is compared with the performance of a representative group of children, referred to as a norm group or a standardization sample.
57. Norms are needed because the number of correct responses the child makes is not very meaningful in itself.
58. A derived score indicates the child's standing relative to the norm group and allows us to compare the child's performance on one measure with his or her performance on other measures.
59. Four concepts related to norm-referenced measurement are population, representative sample, random sample, and reference group.
60. The population is the complete group or set of cases.
61. A representative sample is a group drawn from the population that represents the population accurately.
62. A random sample is a sample obtained by selecting members of the population based on random assignment so that each person in the population has an equal chance of being selected.
63. The reference group is the norm group that serves as the comparison group for computing standard scores, percentile ranks, and related statistics.
64. The representativeness of a norm group reflects the extent to which the group's characteristics match those of the population of interest.
65. For psychological and psychoeducational assessment, the most salient of these characteristics are typically age, grade

level, gender, geographic region, ethnicity, and socioeconomic status (SES).

66. A norm group should be large enough to ensure that the test scores are stable and representative of the population—that is, that the subgroups in the population are adequately represented.
67. To interpret the relevance of a child's scores properly, an examiner needs a reference group against which to evaluate the scores.

Derived Scores

68. The major types of derived scores used in norm-referenced measurement are standard scores, percentile ranks, normal-curve equivalents, stanines, age-equivalent scores, grade-equivalent scores, and ratio IQs.
69. Standard scores are raw scores that have been transformed so that they have a predetermined mean and standard deviation.
70. One type of standard score is a z score, which has $M = 0$ and $SD = 1$.
71. A T score is a standard score from a distribution with $M = 50$ and $SD = 10$.
72. Percentile ranks are derived scores that permit us to determine an individual's position relative to the standardization sample or any other specified sample.
73. A percentile rank is a point in a distribution at or below which the scores of a given percentage of individuals fall.
74. Quartiles are percentile ranks that divide a distribution into four equal parts, with each part containing 25% of the norm group.
75. Deciles, a less common percentile rank, contain 10 bands, with each band containing 10% of the norm group.
76. A major problem with percentile ranks is that we can't assume that the units along the percentile-rank distribution are equal.
77. Normal-curve equivalents (NCEs) are standard scores with $M = 50$ and $SD = 21.06$.
78. Stanines (a contraction of "standard nine") provide a single-digit scoring system with $M = 5$ and $SD = 2$. Stanine scores are expressed as whole numbers from 1 to 9.
79. Age-equivalent scores are obtained by computing the average raw scores obtained on a test by children at different ages.
80. Other terms for age-equivalent scores are test-age equivalent, test age, and mental age, or MA.
81. Grade-equivalent scores are obtained by computing the average raw scores obtained on a test by children in different grades.
82. Age-equivalent and grade-equivalent scores are psychometrically impure.
83. Ratio IQs were defined as ratios of mental age (MA) to chronological age (CA), multiplied by 100 to eliminate the decimal: IQ = MA/CA × 100.
84. All derived scores are obtained from raw scores. The different derived scores are merely different expressions of a child's performance.

Inferential Statistics

85. Inferential statistics are used in drawing inferences about a population based on a sample drawn from the population.
86. When we want to know whether the difference between two or more scores can be attributed to chance or to some systematic or hypothesized cause, we run a test of statistical significance. Statistical significance refers to whether scores differ from what would be expected on the basis of chance alone.
87. Statisticians have generally agreed that a reasonable criterion for deciding that something is not a chance occurrence is that it would happen by chance only 5% of the time or less.
88. We need to consider not only statistical significance, but also the values of the means, the degree to which the means differ, the direction of the mean difference, and whether the results are meaningful—that is, whether they have important practical or scientific implications.
89. Effect size (ES) is a statistical index based on standard deviation units, independent of sample size. It is useful in determining whether the results of a study are meaningful.
90. Cohen's d, a statistic in standard deviation units, provides one way to compute effect size.

Reliability

91. A reliable test is one that is consistent in its measurements.
92. A test is unreliable if scores are subject to large random, unsystematic fluctuations.
93. Technically, reliability of measurement refers to the extent to which random or unsystematic variation affects the measurement of a trait, characteristic, or quality.
94. According to classical psychometric theory, a test score is composed of two components: a true score and an error score.
95. The word *true* refers to the measurement process, not to the underlying content of the test.
96. An error score represents random factors that affect the measurement of the true score.
97. The reliability coefficient, which expresses the degree of consistency in the measurement of test scores, is denoted by the letter r with a subscript consisting of identical letters (e.g., r_{xx} or r_{tt}).
98. Reliability coefficients range from 1.00 (indicating perfect reliability) to .00 (indicating the absence of reliability).
99. Reliability is essential in a psychological measure.
100. Low levels of reliability signify that unknown but meaningful sources of error are operating in the measure and that the measure is not stable across time or consistent across situations.
101. Internal consistency reliability is based on the scores that individuals obtain during a single administration of a test.
102. The most general measure of reliability is Cronbach's coefficient alpha.
103. Test-retest reliability is computed from the scores that individuals obtain on the same test on two different occasions.
104. Alternate-forms reliability (also referred to as parallel-forms reliability or equivalent-forms reliability) is determined by creating two different but parallel forms of a measure and administering the two forms to the same group of children.
105. Interrater reliability (also called examiner reliability or scorer reliability) refers to the degree to which the raters agree.
106. Several factors affect the reliability of a test, including test length, homogeneity of items, test-retest interval, variability of scores, guessing, variation in the test situation, and sample size.
107. The standard error of measurement (SEM), or standard error of a score, is an estimate of the amount of error inherent in a child's obtained score.
108. The standard error of measurement directly relates to the reliability of a test: The lower the reliability, the higher the standard error of measurement; conversely, the higher the reliability, the lower the standard error of measurement.
109. The standard error of measurement represents the standard deviation of the distribution of error scores.

110. When we report a test score, we also should report a confidence interval—a band, or range, of scores around the obtained score that likely includes the child's true score.

111. The confidence interval may be large or small, depending on the degree of certainty we desire (how likely we want it to be that the interval around the child's obtained score contains his or her true score).

112. Individuals who use the test findings need to know that the IQ and other major scores used to make decisions about a child are not perfectly accurate because they inherently contain measurement error.

113. There are two methods for obtaining confidence intervals. One is based on the child's obtained score and the conventional standard error of measurement. The other is based on the estimated true score and the standard error of measurement associated with the estimated true score.

114. In clinical and psychoeducational assessments, questions usually center on how a child is functioning at the time of the referral. Therefore, we recommend that you use the confidence interval based on the child's obtained score, without recourse to the child's estimated true score.

115. When you want to know how a child might perform over a longer period in relation to a specific reference group, use the confidence interval based on the estimated true score.

116. The standard error of estimate allows us to establish a confidence interval around a predicted score.

117. When a test is re-administered, retest scores may differ from those obtained on the initial test.

118. Practice effects may be related to prior exposure to the test.

119. Practice effects may occur because of intervening events between the two administrations.

120. Practice effects may not occur to the same extent in all populations.

121. Practice effects vary for different types of tasks.

122. Practice effects may be affected by regression toward the mean.

123. Practice effects may be difficult to interpret when the initial test and the retest are different.

124. Practice effects may depend on the item content covered throughout the test.

Item Response Theory

125. Item difficulty refers to the percentage of children who answer an item correctly.

126. Item discrimination refers to how an item discriminates between children who do well on the test as a whole and those who do poorly.

127. Item response theory uses three parameters to evaluate items: item discrimination, item difficulty, and a "guessing" parameter, which reflects the probability that a correct response will occur by chance.

128. An item characteristic curve is a line representing the probability of passing the item for children with different total scores on the construct being measured.

Differential Item Functioning

129. The assessment of differential item functioning is a statistical procedure designed to reveal whether test items function differently in different groups.

Validity

130. The validity of a test refers to whether it measures what it is supposed to measure.

131. Because tests are used for many different purposes, there is no single type of validity appropriate for all assessment purposes.

132. Validity is more difficult to define than reliability. Unlike reliability, validity has no single definition.

133. A related problem is that the terminology used in the literature on validity is inconsistent.

134. A good way to determine the validity of a test is to understand what it measures and then decide what measures should and should not be correlated with it.

135. Two issues are addressed in validating tests: what a test measures and how well it measures it.

136. Content validity refers to whether the items within a test or other measure represent the domain being assessed.

137. Face validity refers to whether a test looks valid "on the face of it."

138. Construct validity establishes the degree to which a test measures a specified psychological construct (i.e., an inferred entity) or trait.

139. Convergent validity refers to how well measures of the same domain in different formats—such as tests in multiple-choice, essay, and oral formats—correlate with each other.

140. Discriminant validity, sometimes called divergent validity, refers to the extent to which measures of different domains do not correlate with each other.

141. Criterion-related validity is based on how positively test scores correlate with some type of criterion or outcome (such as ratings, classifications, or other test scores).

142. The two forms of criterion-related validity are concurrent validity and predictive validity.

143. Concurrent validity is based on correlations of scores on one measure with those on a related measure.

144. Predictive validity is based on correlations of scores on one measure with those on a criterion measure taken at a later time.

145. Results from criterion-related validity studies are usually expressed as correlation coefficients.

146. Predictive power is a special type of predictive validity. It assesses the accuracy of a decision made on the basis of a given measure. Thus, predictive power refers to the extent to which a test (or another measure, such as a rating scale or an observation form) agrees with an outcome criterion measure used to classify individuals in a particular category or to determine whether or not they have a particular trait or condition.

147. All predictions must be compared to the base rate of a condition, an attribute, or a disease in a specific population. Base rates are important, because they are the rates against which we judge the accuracy of a prediction.

148. We compute the predictive power of a test by determining the percentages of correct and incorrect classifications that it makes. At least 10 different measures of predictive power can be computed.

149. Validity coefficients are affected by the same factors that affect correlation coefficients, as well as by other factors such as range of attributes being measured, length of the interval between administration of the test and of the criterion measure, and range of variability in the criterion measure.

150. The validity of a child's test scores can be affected by such factors as the child's test-taking skills, anxiety, transient medi-

cal conditions, confusion, limited attention, degree of rapport with the examiner, motivation, speed, understanding of test instructions, physical handicaps, temporary hearing impairments, language skills, educational opportunities, and familiarity with the test material.

151. Validity can also be affected by intervening events and contingencies. Deficiencies in the robustness of the criterion might affect the validity of tests.
152. If you have any reason to question the validity of test results, state your reservations in the psychological report.

Meta-Analysis

153. Meta-analysis uses rigorous research techniques (including quantitative methods) to sum up and integrate the findings of a body of studies covering similar topics.
154. Meta-analysis is particularly useful in validity generalization studies.
155. Conclusions from meta-analysis may be compromised by the variety of studies reviewed and their shortcomings, such as poor design and inadequate sampling.

Factor Analysis

156. Factor analysis is a mathematical procedure used to explain the pattern of intercorrelations among a set of variables (such as individual test items, entire tests, subtests, or rating scales) by deriving the smallest number of meaningful variables or factors.
157. A factor is a statistically derived, hypothetical dimension that accounts for part of the intercorrelations among a set of variables.
158. Factor analysis is based on the assumption that a significant correlation between two variables indicates a common underlying factor shared by both variables.
159. An exploratory factor analysis (EFA) is used to explore the underlying structure of a collection of variables when there are no a priori hypotheses about the factor structure.
160. A confirmatory factor analysis (CFA) is used to confirm a hypothesized factor structure.
161. Principal component analysis seeks the set of factors that can account for all common and unique variance in a set of variables.
162. Principal factor analysis, which incorporates prior communality estimates, seeks the smallest set of factors that can account for the common variance in a set of variables.
163. Rather than attempting to interpret the original factors, researchers usually rotate the matrix of factor loadings to make the factor structure clearer. The rotation rearranges the factors so that, ideally, for every factor there are some variables with high loadings on the factor and other variables with low loadings on the factor.
164. In a factor analysis, we can divide the variance associated with a variable into three categories: communality, specificity, and error variance.
165. Communality refers to that part of the total variance that can be attributed to common factors (those that appear in more than one variable).
166. Specificity refers to that part of the total variance that is due to factors specific to a particular variable, not to measurement error or common factors.
167. Error variance refers to that part of the total variance that remains when we subtract the reliability of the variable from the total variance.
168. Factor analysis is a complex statistical method. The same set of data can yield different results depending on the factor analytic method used, the number of factors retained, and the rotations of the factors. In addition, the naming of factors is arbitrary.

Other Useful Psychometric Concepts

169. Occasionally, you will find that two or more tests believed to measure the same ability give different results for the same child. Different results might occur, for example, because of characteristics of the child, testing conditions, examiner characteristics, or the psychometric properties of the tests.
170. The psychometric properties of two supposedly similar tests might lead to different results because of floor effect differences, ceiling effect differences, item gradient differences, norm table layout differences, age-equivalent or grade-equivalent score differences, reliability differences, differences in skill areas assessed, test content differences, publication date differences, and sampling differences.

Concluding Comment

171. No instrument is completely reliable (i.e., without error).
172. Validity does not exist in the abstract; it must be anchored to the specific purposes for which the instrument is used.
173. Every child's behavior fluctuates from time to time and from situation to situation (e.g., a child might perform differently with different examiners).
174. Any assessment instrument contains only a sample of all possible questions or items related to the domain of interest.
175. Assessment instruments purporting to measure the same construct may give different results for a particular child.
176. Instruments measure samples of behavior or constructs at one point in time.
177. Assessment scores will likely change to some degree during a child's development.

KEY TERMS, CONCEPTS, AND NAMES

The why of psychological measurement and statistics (p. 92)
Scale (p. 92)
Lower-order scales (p. 92)
Discrete variables (p. 92)
Higher-order scales (p. 92)
Continuous variables (p. 92)
Nominal measurement scale (p. 92)
Ordinal measurement scale (p. 93)
Interval measurement scale (p. 93)
Ratio measurement scale (p. 93)
Descriptive statistics (p. 94)
Measures of central tendency (p. 94)
Mean (p. 94)
Outliers (p. 95)
Median (p. 95)
Mode (p. 95)
Unimodal (p. 95)
Bimodal (p. 95)

Multimodal (p. 95)
Dispersion (p. 95)
Range (p. 95)
Variance (p. 96)
Standard deviation (p. 96)
Normal curve (p. 96)
Gaussian distribution (p. 96)
Abscissa (p. 96)
Correlation (p. 97)
Correlation coefficients (p. 97)
Strength of the relationship (p. 97)
Direction of the relationship (p. 97)
Linear relationship (p. 98)
Curvilinear relationship (p. 98)
Dichotomous variable (p. 98)
Scatterplot (p. 98)
Pearson correlation coefficient (p. 98)
Spearman r_s (p. 99)
Correcting for attenuation (p. 99)
Coefficient of determination (p. 100)
Regression (p. 100)
Linear equation (p. 100)
Regression equation (p. 100)
Standard error of estimate (p. 101)
Multiple correlation (p. 102)
Norm-referenced measurement (p. 102)
Norm group (p. 102)
Standardization sample (p. 102)
Derived scores (p. 102)
Population (p. 102)
Representative sample (p. 102)
Random sample (p. 102)
Reference group (p. 102)
Representativeness (p. 102)
Size (p. 103)
Relevance (p. 103)
Standard scores (p. 103)
z score (p. 103)
T score (p. 103)
Percentile ranks (p. 103)
Quartiles (p. 103)
Deciles (p. 103)
Normal-curve equivalents (p. 103)
Stanines (p. 104)
Age-equivalent scores (p. 104)
Grade-equivalent scores (p. 104)
Test-age equivalents (p. 104)
Test age (p. 104)
Mental age (MA) (p. 104)
Ratio IQs (p. 106)
Inferential statistics (p. 107)
Statistical significance (p. 107)
Effect size (p. 108)
Cohen's *d* (p. 108)
Reliability (p. 109)
Theory of reliability of measurement (p. 109)
True score (p. 109)
Error score (p. 109)
Reliability coefficient (p. 109)
Internal consistency reliability (p. 109)
Cronbach's coefficient alpha (p. 109)
Kuder-Richardson formula 20 coefficient (p. 109)
Spearman-Brown correction formula (p. 109)
Split-half method (p. 109)
Test-retest reliability (p. 109)
Coefficient of stability (p. 109)
Alternate-forms reliability (p. 110)
Parallel-forms reliability (p. 110)
Equivalent-forms reliability (p. 110)
Coefficient of equivalence (p. 110)
Interrater reliability (p. 111)
Examiner reliability (p. 111)
Scorer reliability (p. 111)
Percentage agreement (p. 111)
Factors affecting reliability (p. 111)
Standard error of measurement (SEM) (p. 111)
Confidence interval (p. 112)
Confidence intervals for obtained scores without reference to the estimated true score (p. 112)
Confidence intervals for obtained scores with reference to the estimated true score (p. 113)
Confidence intervals for predicted scores (p. 114)
Repeated evaluations (p. 115)
Practice effects (p. 115)
Item response theory (IRT) (p. 116)
Item analysis (p. 116)
Item difficulty (p. 116)
Item discrimination (p. 116)
Latent trait model (LTM) (p. 116)
Item characteristic curve (ICC) (p. 116)
Differential item functioning (DIF) (p. 117)
Validity (p. 117)
Content validity (p. 117)
Face validity (p. 118)
Construct validity (p. 118)
Convergent validity (p. 118)
Discriminant validity (p. 118)
Divergent validity (p. 118)
Criterion-related validity (p. 119)
Concurrent validity (p. 119)
Predictive validity (p. 119)
Predictive power (p. 119)
True positive (hit) (p. 120)
False positive (false alarm) (p. 120)
False negative (miss) (p. 120)
True negative (correct rejection) (p. 120)
True positive rate (p. 120)
Index of sensitivity (p. 120)
Valid positive rate (p. 120)
Hit rate (p. 120)
False positive rate (p. 120)
False alarm rate (p. 121)
False negative rate (p. 121)
Miss rate (p. 121)
Underreferral rate (p. 121)
True negative rate (p. 121)
Index of specificity (p. 121)
Valid negative rate (p. 121)
Correct rejection rate (p. 121)
Positive predictive power (p. 121)
Efficiency rate (p. 121)
Negative predictive power (p. 121)

Overall accuracy rate (p. 121)
Overall hit rate (p. 121)
Correct classification rate (p. 121)
Observed proportion of overall agreement (p. 121)
Effectiveness rate (p. 121)
Overall inaccuracy rate (p. 121)
Overall error rate (p. 121)
Incorrect classification rate (p. 121)
Observed proportion of overall disagreement (p. 121)
Misclassification rate (p. 121)
Base rate (p. 121)
Prevalence rate (p. 121)
True proportion (p. 121)
Odds ratio (p. 121)
Factors affecting validity (p. 121)
Premorbid level (p. 122)
Meta-analysis (p. 122)
Factor analysis (p. 123)
Factor (p. 123)
Factor loadings (p. 123)
Exploratory factor analysis (EFA) (p. 123)
Confirmatory factor analysis (CFA) (p. 123)
Principal component analysis (PCA) (p. 123)
Principal factor analysis (PFA) (p. 123)
First unrotated factor (p. 123)
General factor (p. 123)
g (p. 123)
Varimax rotation (p. 124)
Orthogonal (p. 124)
Oblim rotation (p. 124)
Group factors (p. 124)
Specific factor variance (p. 124)
Specific variance (p. 124)
Communality (p. 124)
Specificity (p. 124)
Error variance (p. 124)
Other useful psychometric concepts (p. 125)
Floor effect differences (p. 125)
Ceiling effect differences (p. 126)
Item gradient differences (p. 126)
Norm table layout differences (p. 126)
Age-equivalent or grade-equivalent score differences (p. 126)
Reliability differences (p. 126)
Differences in skill area assessed (p. 126)
Test content differences (p. 126)
Publication date differences (p. 127)
Sampling differences (p. 127)

STUDY QUESTIONS

1. Discuss why psychological measurement and statistics are useful.
2. Compare and contrast nominal, ordinal, interval, and ratio scales.
3. Describe the three measures of central tendency.
4. Discuss measures of dispersion. Include in your discussion the range, variance, and standard deviation.
5. Discuss the normal curve.
6. Explain the importance of correlation in psychological assessment.
7. Discuss the regression equation.
8. What is the standard error of estimate?
9. What are some important features of norm-referenced measurement?
10. Discuss derived scores. Include in your discussion types of derived scores and relationships among derived scores.
11. Discuss inferential statistics. Include in your discussion the concept of statistical significance and effect size.
12. Discuss the concept of reliability. Include in your discussion the theory of reliability of measurement, reliability coefficients, internal consistency reliability, test-retest reliability, alternate-forms reliability, interrater reliability, factors affecting reliability, standard error of measurement, confidence intervals for obtained scores, confidence intervals for predicted scores, and repeated evaluations.
13. Discuss item response theory.
14. Discuss differential item functioning.
15. Discuss the concept of validity. Include in your discussion the various types of validity, predictive power, and factors affecting validity.
16. Discuss meta-analysis and describe its usefulness in validity studies.
17. Discuss factor analysis.
18. Discuss other useful psychometric concepts.