Module 6

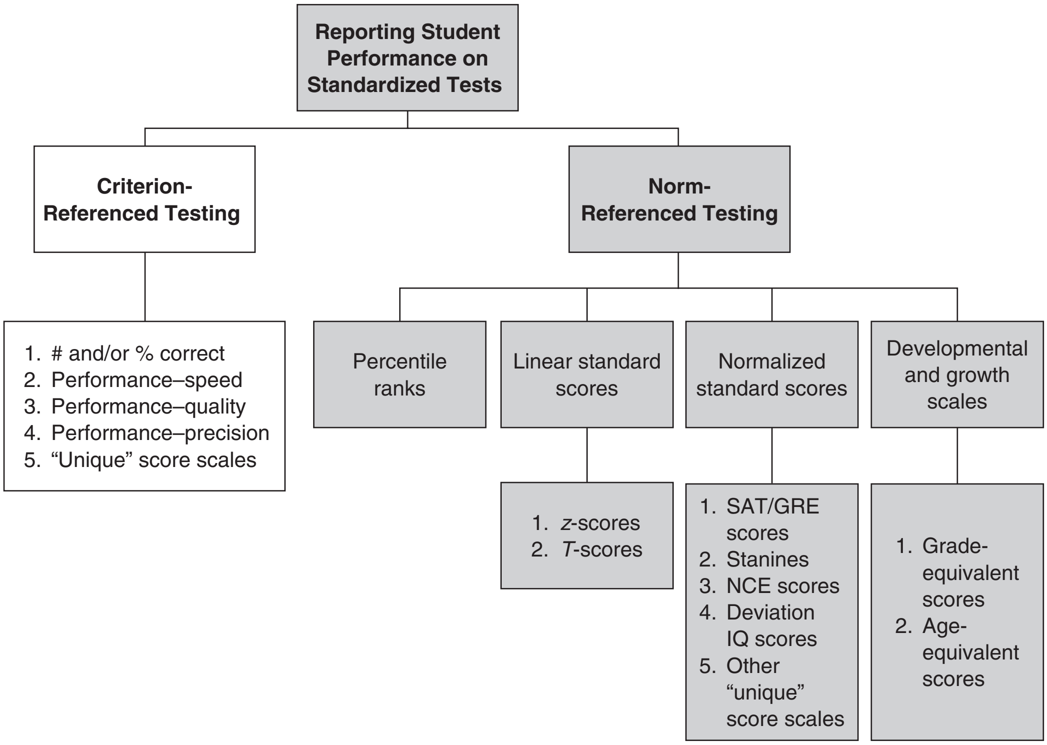
NORM-REFERENCED TEST SCORES AND THEIR INTERPRETATIONS

Compared to the number of different criterion-referenced scores used on standardized test reports, the various types of norm-referenced scores substantially outnumber those that we examined in the previous module. Figure 6.1 summarizes the various scores used for each type of test, this time highlighting those that are norm-referenced. In this module, we will explore the information that is reported by norm-referenced test scores, the various types of norm-referenced scores, and the importance of understanding what is meant by the standard error of measurement, as well as how to apply this knowledge when interpreting student test performance. Information specific to how these scores can be used to help guide or revise instruction will be discussed in Modules 7, 8, and 9.

WHAT DO NORM-REFERENCED SCORES TELL US?

When norm-referenced standardized tests are administered to students, the results are reported in a way that permits comparisons with a well-defined group of other students who have taken the same assessment (Nitko, 2004). The primary difference between criterion- and norm-referenced scores is that with norm-referenced test scores, individual student scores are entirely dependent upon the performance of other students. Norm-referenced tests and their resulting scores provide evidence that assists educators in answering the following questions:

Figure 6.1 Summary of Methods of Reporting Scores on Standardized Tests, Highlighting Types of Norm-Referenced Scores



- What is the relative standing of this student (or the students in this class or district) across this broad domain of content?
- How does the student (or do these students) compare to other similar students?

As you read earlier in Module 1, the makeup of the group functioning as the comparison students forms the basis for interpreting scores resulting from norm-referenced tests. This well defined group of students, known as a norm group, is given the same assessment under the same conditions (e.g., same time limits, same materials, same directions). Comparisons to norm groups enable teachers and administrators to describe achievement levels of students across different subject areas, to identify strengths and weaknesses across the curriculum, and to identify areas of deficiency—as well as subsequent intervention strategies—within each subject area (Nitko, 2004).

THE NATURE OF NORM GROUPS

Descriptions of norm groups are typically provided in the test manuals that accompany the actual tests. It is important to realize that the average performance of the norm group does not represent a standard to be attained or exceeded by all students in every school across the country. In contrast, the performance of the norm group on a particular standardized test is intended to represent the current level of achievement for a specific group of students, usually at a certain grade level (Nitko, 2004). Therefore, comparisons to the norm group can assist educators in making decisions about the general range or level of performance to expect from their students.

There are several types of norms that can be reported in norm-referenced tests. These include local norm groups (e.g., composed of student within a particular school district), special norm groups (e.g., composed of students who are blind or deaf), and school average norms. However, most norm-referenced tests rely on national norms (Airasian, 2005; Nitko, 2004). National norm groups are selected in order to be representative of the entire country. This representation is based on such characteristics as gender, race, ethnicity, culture, and socioeconomic status. The purpose of obtaining a representative norm group is to reduce any potential bias when comparing students from diverse backgrounds around the country. In addition to being representative, norm groups must also be current (Nitko, 2004). Test publishers work very hard at ensuring representativeness, although perfect representation can never be achieved. If the norm group for a given test is not representative and current, the resulting test scores will likely lead to misinterpretation and ultimately inappropriate educational decisions.

Publishers of norm-referenced tests have found it advantageous to sometimes transform scores so that they can be placed in some common distribution. This common distribution is called a *normal distribution*, also known as a *normal curve* or a *bell-shaped curve.* Normal distributions have three main characteristics (Mertler, 2003). These characteristics are as follows:

- The distribution is symmetrical (i.e., the left and right halves are mirror images of each other).
- The mean (or arithmetic average), median (the score that separates the upper 50% of scores from the lower 50% of scores), and the mode (the most frequently occurring score) are the same score and are located at the exact center of the distribution.
- The percentage of cases in each standard deviation (or the average distance of individual scores away from the mean) is known precisely.

The normal distribution was derived over 250 years ago (Nitko, 2004). When first originated, it was based on the belief that nearly all physical characteristics in humans were, by nature, distributed randomly around an average value. Furthermore, the vast majority of cases were located in the middle of the distribution (indicating that most people were, roughly speaking, average). A very small proportion of individuals can be found at the extreme ends of the distribution. This serves as an indication that, with respect to most characteristics, the majority of people are relatively similar to one another (e.g., approximately average height), with a minority of people at the high (i.e., very tall) and low (i.e., very short) ends. The concept of randomly and normally distributed physical characteristics has since carried over into the realm of mental measurement.

As shown in Figure 6.2, each standard deviation in a normal distribution contains a fixed percentage of cases (Nitko, 2004). The mean score plus and minus 1 standard deviation contains approximately 68% of the individuals making up the distribution; 95% of the cases are within 2 standard deviations of the mean; and over 99% of the cases are within 3 standard deviations. From the figure, it should be clear that 50% of the cases, or scores, are located above the mean (i.e., the right half of the distribution) and 50% are located below the mean (i.e., the left half); this should also make intuitive sense because in a normal distribution, the mean and median are located at the same score. Moreover, nearly 16% of the scores are greater than one standard deviation above the mean.

This information about the percentage of cases in the various segments of the distribution is key to the interpretation of scores resulting from norm-referenced standardized tests. A main purpose of the normal distribution is to help educators get a sense of how high or low a given score is in relation to an entire distribution of scores (Mertler, 2003). This serves as the basis for many of the scores we will discuss next.

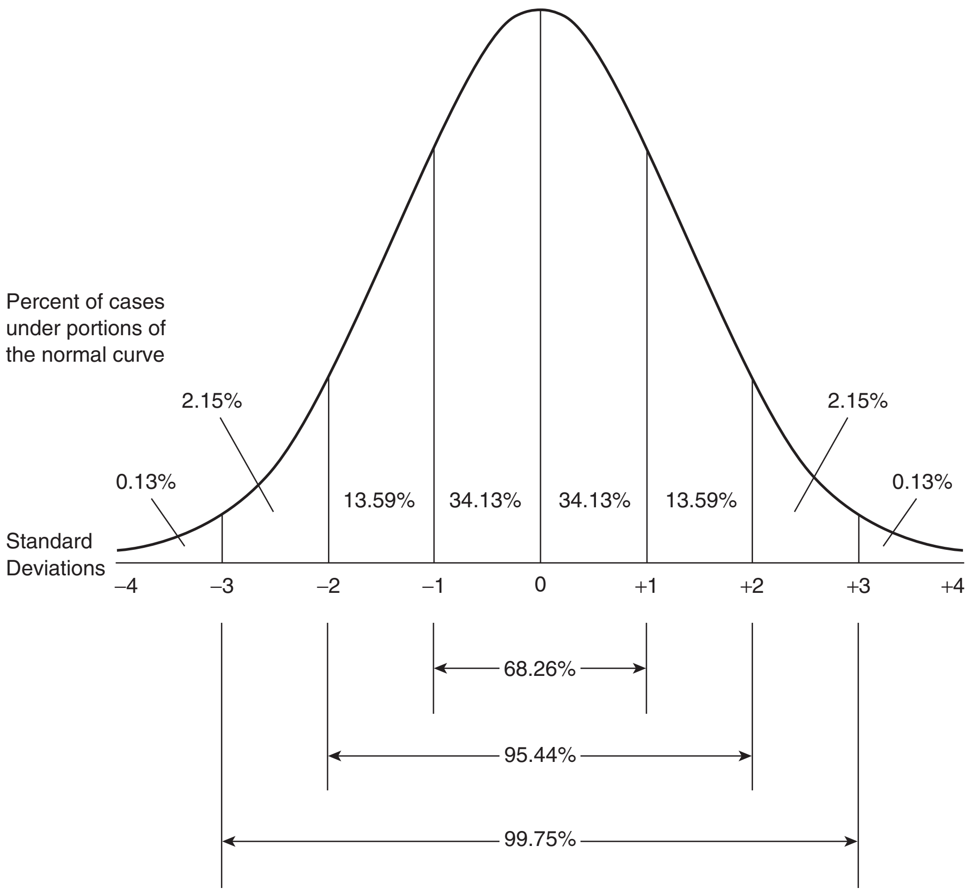
TYPES OF NORM-REFERENCED SCORES

As you saw in Figure 6.1, there are numerous types of norm-referenced scores. Most of them are based on mathematical transformations. In other words, the raw scores are changed or converted to some other scale. These new scales then conform to the characteristics of the normal distribution as were previously discussed. It is important to bear in mind that norm-referenced test scores are all based on the notion of how an individual student performs as compared to a large group of similar students. Most of these students will be average, with their performance being located near the middle of the distribution.

Raw Scores

In the previous module on criterion-referenced test scores, you read about the use of raw scores. Raw scores are the main method of reporting results of

Figure 6.2 Characteristics of the Normal Distribution



criterion-referenced tests. However, norm-referenced test reports also typically provide the raw scores (i.e., the number of items answered correctly) obtained by students on various tests and subtests. However, these scores are not very useful when interpreting the results of achievement or aptitude tests for purposes of norm-referenced comparisons, for their distributions would likely not be symmetrical or normal in nature. More importantly, teachers need to know how a particular student's raw score compares to the specific norm group. To make these types of comparisons, raw scores must first be converted to some other score scale. These new scales are referred to as *transformed* or *derived scores* and include such scores as percentile ranks, *z*-scores, *T*-scores, normal curve equivalent scores, deviation IQ scores, and stanines, among others.

Percentile Ranks

A *percentile rank* is a single number that indicates the percentage of the norm group that scored below a given raw score. Possible values for percentile ranks range from 1 to 99. However, since percentile ranks indicate various percentages of individuals above and below scores that are normally distributed, they do not represent equal units (Mertler, 2003). Percentile ranks are much more compactly arranged in the middle of the normal distribution since that is where the majority of individuals fall.

Let us consider a hypothetical example for a student, Annie, who recently took a norm-referenced achievement test. Annie's test report includes a percentile rank for all subtests appearing on the report. Let us assume that Annie correctly answered 34 out of a possible 45 items on the reading subtest of the total test battery. When converted, this raw score of 34 converts to a percentile rank equal to 86. This means that based on her raw score Annie scored higher than 86% of the other students (in the norm group) who took the test. In other words, 86% of the students who made up the norm group answered fewer than 34 items correctly.

Percentile ranks are among the most frequently reported derived scores, yet they are also among the most frequently misunderstood (Mertler, 2003). A common misinterpretation of these test scores is that they are equivalent to the percentages of items answered correctly. In our example above, Annie correctly answered 76% (i.e., 34 out of 45) of the items—clearly not the same as the 86th percentile. It is important to realize that making the assumption that a percentile rank should be interpreted in the same manner as a percentage of items answered correctly implies a criterion-referenced (i.e., the score that she actually received) rather than a norm-referenced (i.e., her score in relation to others) interpretation.

Percentile ranks indicate relative standing, but have some limitations when compared to other types of derived scores we will discuss shortly. Percentile ranks are expressed in ordinal units, which means that the distance between adjacent units on a percentile scale are not equal (Payne, 2003). The distance between the 49th and 50th percentiles is much smaller—due to the substantially large number of individuals clustered at the center of the distribution—than the distance between the 1st and 2nd percentiles. In fact, the distance between the 1st and 3rd percentiles is exactly the same as the distance between the 16th and 50th percentiles. Since units are not equal, a difference in student performance for two students located at the extreme right end of the distribution (e.g., a one-unit difference) will appear less important on a percentile rank scale than the same difference for two students located in the middle of the distribution. For example, a one raw-score unit difference for students scoring near the mean (i.e., near the middle of the distribution) may differ by several percentile ranks while two students located in the tail of the distribution with a one raw-score unit difference might both have the same percentile rank (Mertler, 2003).

There is sometimes a dangerous temptation for teachers to average percentile ranks to find a student's typical performance or to subtract them to find the difference between two scores. Since percentiles do not represent equal units, they should not be mathematically manipulated in such a manner. In other words, they cannot be added, subtracted, multiplied, or divided as a means of further comparing students' relative standings or comparing student gains or losses (Oosterhof, 2001).

Developmental/Growth Scores

Developmental scales seek to identify a student's development across various levels (e.g., grade or age) of growth (Airasian, 2005). The purpose of these scores is to compare a student's performance to a series of reference groups that vary developmentally.

Grade-Equivalent Scores

A common type of developmental score that frequently appears on norm-referenced test reports is the *grade-equivalent score*. A grade-equivalent score indicates the grade in the norm group for which a certain raw score was the median performance (Oosterhof, 2001) and is intended to estimate a student's developmental level (Airasian, 2005). Grade-equivalent scores are expressed in years and 10th of years (Spinelli, 2006); they consist of two numerical components separated by a period. The first number indicates the grade level, and the second indicates the month during that particular school year, which ranges from zero (equivalent to September) to nine (equivalent to June). For example, if a student receives a raw score of 67 on the mathematics portion of an achievement test, this score might be transformed to a grade-equivalent score of 4.2. This means that this student's performance corresponds to the performance of a typical student taking the same test in November (i.e., the 2nd month) of fourth grade.

Grade-equivalent scores are often misinterpreted as standards that all students should be expected to achieve (Oosterhof, 2001). It is again important to remember that a criterion-referenced interpretation such as this is an inappropriate use of a grade-equivalent score, which is a norm-referenced score. Similarly, grade-equivalent scores are not intended to indicate appropriate grade-level placement. If a student receives a score of 5.1 on a mathematics subtest, we should not assume that he is ready for fifth-grade math, an assumption which is again a criterion-referenced interpretation and an inappropriate use of the score. We could not possibly know where this student stands with respect to fourth-grade material since he was tested on third-grade content. An additional limitation of grade-equivalent scores is that although the scores represent months, they do not, in reality, represent equal units. For example, gains made in reading achievement between grade 1.0 and grade 1.5 are very likely greater than reading achievement gains made between grade 6.0 and grade 6.5. Because these scores can be misleading and can lead to inaccurate generalizations, they should be interpreted and used for instructional decisions with substantial caution (Spinelli, 2006). Finally, with respect to grade-equivalent scores, it is important to remember that the scores represent what is considered typical or average. If the scores for the norm group result in a normal distribution, half of the total group of students who take the test will score below the average for the group (Tanner, 2001).

Age-Equivalent Scores

Very similar to grade-equivalent scores, *age-equivalent scores* are based on the average test performances of students at various age levels as opposed to various grade levels as we saw previously (Payne, 2003). Their units are also unequal, meaning that equal age units (e.g., 6 months or 1 year) do not correspond to equal age-equivalent score units. As with grade-equivalent scores, age-equivalent scores are useful for measuring and describing growth in mental ability, reading ability, and other types of characteristics that exhibit fairly consistent growth patterns within an instructional program (Payne, 2003). They are very useful in monitoring development and growth. Age-equivalent scores are expressed in a similar fashion to their grade-equivalent counterparts.

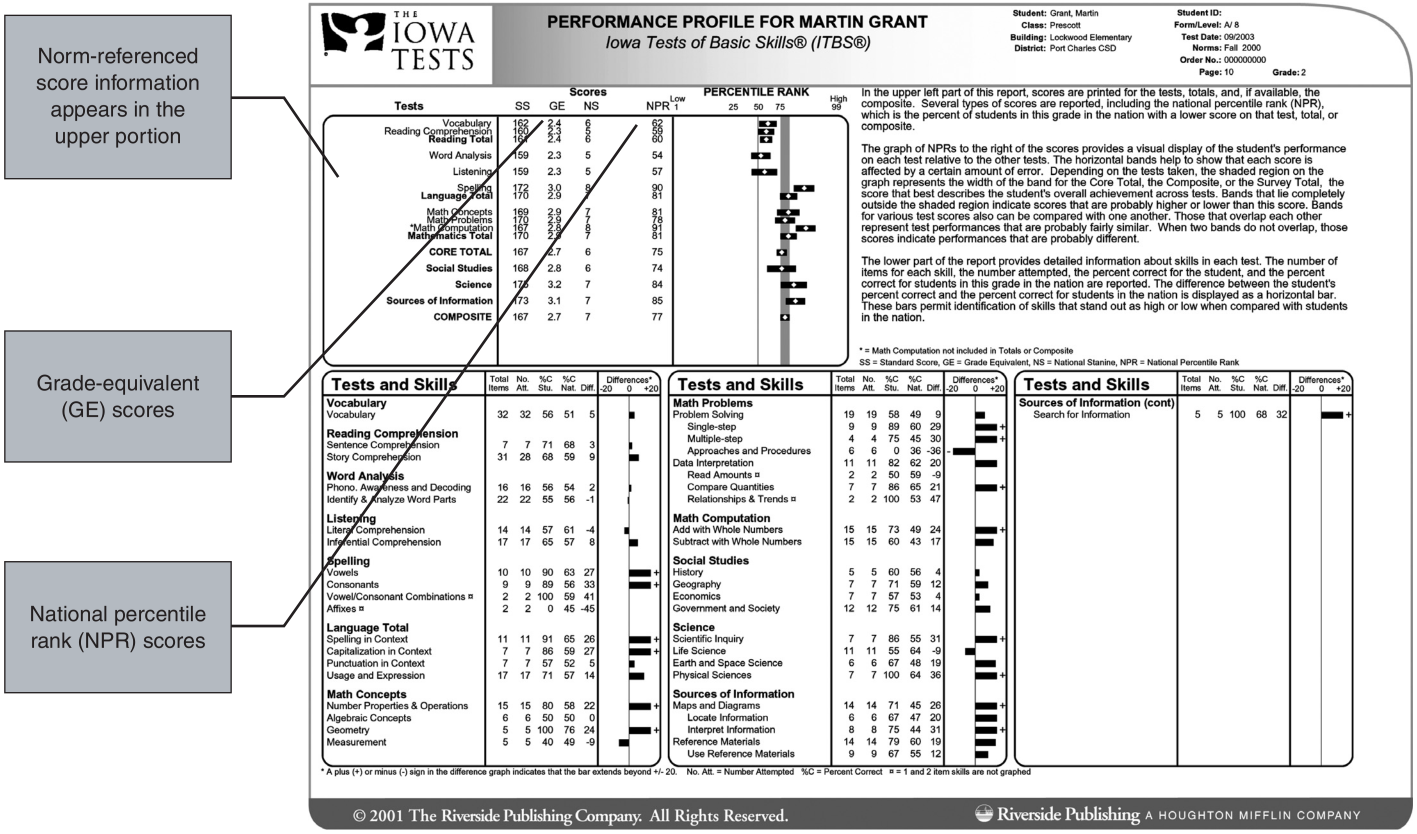
Figure 6.3 is a sample Individual Performance Profile from the Iowa Tests of Basic Skills (ITBS). On this report, the norm-referenced score information appears in the upper portion of the report. Labeled on this sample are the grade-equivalent scores and national percentile ranks.

Standardized Scores

You have seen that both percentile ranks and grade-equivalent or age-equivalent scores exist on scales with unequal units. This characteristic seriously limits the interpretability and utility of each type of score. *Standardized scores* (also known as standard scores) are obtained when raw scores are transformed to fit a distribution whose characteristics are known and fixed (Tanner, 2001). Specifically, this known distribution is the normal distribution; and the scores are reported in standard deviation units, which are equal across the entire continuum. As a result of these transformations, scores can be interpreted in a way that is unaffected by the characteristics of a particular test. Regardless of the test, standardized scores efficiently indicate whether a particular score is typical, above average, or below average as compared to others who took the test and also clearly indicate the magnitude of the variation away from the mean score (Tanner, 2001).

Moreover, standardized scores allow for comparisons of test performance across two different measures (Mertler, 2003). For example, suppose you want to compare

Figure 6.3 Sample Individual Student Test Score Report


Norm-referenced score information appears in the upper portion
Grade-equivalent (GE) scores
National percentile rank (NPR) scores

THE IOWA TESTS

PERFORMANCE PROFILE FOR MARTIN GRANT

Iowa Tests of Basic Skills® (ITBS®)

Student:	Grant, Martin	**Student ID:**	
Class:	Prescott	**Form/Level:**	A/ 8
Building:	Lockwood Elementary	**Test Date:**	09/2003
District:	Port Charles CSD	**Norms:**	Fall 2000
		Order No.:	000000000
		Page:	10 **Grade:** 2

	Scores				PERCENTILE RANK
Tests	SS	GE	NS	NPR	Low 1 25 50 75 High 99
Vocabulary	162	2.4	6	62	
Reading Comprehension	160	2.3	5	59	
Reading Total	161	2.4	6	60	
Word Analysis	159	2.3	5	54	
Listening	159	2.3	5	57	
Spelling	172	3.0	8	90	
Language Total	170	2.9	7	81	
Math Concepts	169	2.9	7	81	
Math Problems	170	2.9	7	78	
*Math Computation	167	2.8	8	91	
Mathematics Total	170	2.9	7	81	
CORE TOTAL	167	2.7	6	75	
Social Studies	168	2.8	6	74	
Science	175	3.2	7	84	
Sources of Information	173	3.1	7	85	
COMPOSITE	167	2.7	7	77	

In the upper left part of this report, scores are printed for the tests, totals, and, if available, the composite. Several types of scores are reported, including the national percentile rank (NPR), which is the percent of students in this grade in the nation with a lower score on that test, total, or composite.

The graph of NPRs to the right of the scores provides a visual display of the student's performance on each test relative to the other tests. The horizontal bands help to show that each score is affected by a certain amount of error. Depending on the tests taken, the shaded region on the graph represents the width of the band for the Core Total, the Composite, or the Survey Total, the score that best describes the student's overall achievement across tests. Bands that lie completely outside the shaded region indicate scores that are probably higher or lower than this score. Bands for various test scores also can be compared with one another. Those that overlap each other represent test performances that are probably fairly similar. When two bands do not overlap, those scores indicate performances that are probably different.

The lower part of the report provides detailed information about skills in each test. The number of items for each skill, the number attempted, the percent correct for the student, and the percent correct for students in this grade in the nation are reported. The difference between the student's percent correct and the percent correct for students in the nation is displayed as a horizontal bar. These bars permit identification of skills that stand out as high or low when compared with students in the nation.

* = Math Computation not included in Totals or Composite

SS = Standard Score, GE = Grade Equivalent, NS = National Stanine, NPR = National Percentile Rank

Tests and Skills	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Differences* -20 0 +20
Vocabulary						
Vocabulary	32	32	56	51	5	
Reading Comprehension						
Sentence Comprehension	7	7	71	68	3	
Story Comprehension	31	28	68	59	9	
Word Analysis						
Phono. Awareness and Decoding	16	16	56	54	2	
Identify & Analyze Word Parts	22	22	55	56	-1	
Listening						
Literal Comprehension	14	14	57	61	-4	
Inferential Comprehension	17	17	65	57	8	
Spelling						
Vowels	10	10	90	63	27	+
Consonants	9	9	89	56	33	+
Vowel/Consonant Combinations ¤	2	2	100	59	41	
Affixes ¤	2	2	0	45	-45	
Language Total						
Spelling in Context	11	11	91	65	26	+
Capitalization in Context	7	7	86	59	27	+
Punctuation in Context	7	7	57	52	5	
Usage and Expression	17	17	71	57	14	
Math Concepts						
Number Properties & Operations	15	15	80	58	22	+
Algebraic Concepts	6	6	50	50	0	
Geometry	5	5	100	76	24	+
Measurement	5	5	40	49	-9	

Tests and Skills	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Differences* -20 0 +20
Math Problems						
Problem Solving	19	19	58	49	9	
Single-step	9	9	89	60	29	+
Multiple-step	4	4	75	45	30	+
Approaches and Procedures	6	6	0	36	-36	-
Data Interpretation	11	11	82	62	20	
Read Amounts ¤	2	2	50	59	-9	
Compare Quantities	7	7	86	65	21	+
Relationships & Trends ¤	2	2	100	53	47	
Math Computation						
Add with Whole Numbers	15	15	73	49	24	+
Subtract with Whole Numbers	15	15	60	43	17	
Social Studies						
History	5	5	60	56	4	
Geography	7	7	71	59	12	
Economics	7	7	57	53	4	
Government and Society	12	12	75	61	14	
Science						
Scientific Inquiry	7	7	86	55	31	+
Life Science	11	11	55	64	-9	
Earth and Space Science	6	6	67	48	19	
Physical Sciences	7	7	100	64	36	+
Sources of Information						
Maps and Diagrams	14	14	71	45	26	+
Locate Information	6	6	67	47	20	
Interpret Information	8	8	75	44	31	+
Reference Materials	14	14	79	60	19	
Use Reference Materials	9	9	67	55	12	

Tests and Skills	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.	Differences* -20 0 +20
Sources of Information (cont)						
Search for Information	5	5	100	68	32	+

* A plus (+) or minus (-) sign in the difference graph indicates that the bar extends beyond +/- 20. No. Att. = Number Attempted %C = Percent Correct ¤ = 1 and 2 item skills are not graphed

Riverside Publishing A HOUGHTON MIFFLIN COMPANY

SOURCE: Copyright © 2001 by Riverside Publishing. Sample “Individual Performance Profile” from the Iowa Tests of Basic Skills (ITBS). Reproduced with permission of the publisher. All rights reserved.

students' performances on a standardized reading test and a standardized mathematics test. However, the reading test is composed of 65 items and the math test contains 34 items. The mean score on the reading test is 45 and on the math test is 24. Simply comparing raw scores would not tell you very much about a student's relative standing as compared to the norm group. If Katherine received a raw score of 40 (i.e., 40 out of 65) on the reading test and a raw score of 30 (i.e., 30 out of 45) on the math test, it would be incorrect to say that she performed better on the reading test, even though she answered more items correctly. (Remember, in this case we are examining test performance from a norm-referenced perspective.) You might notice on the score report that her score of 40 on the reading test is below the average while her score of 30 on mathematics is above average. This type of norm-referenced comparison is possible only through the use of standardized scores because scores from two different subtests are essentially put on the same score scale.

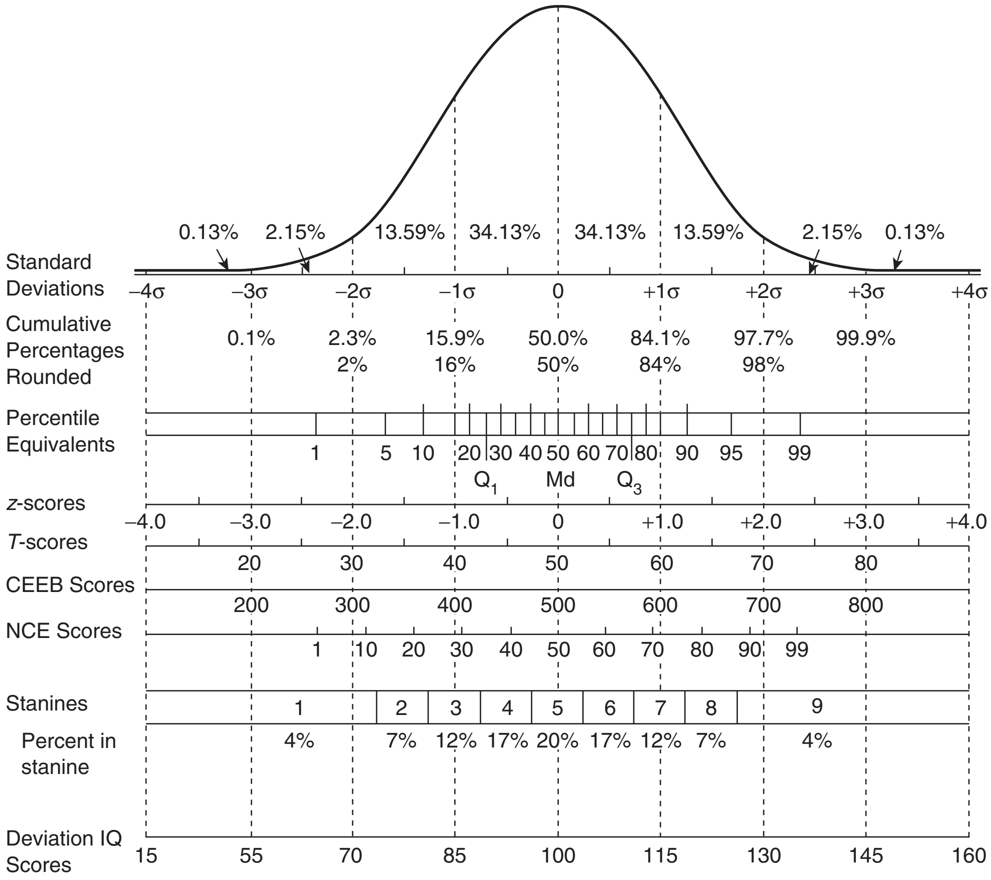
Standardized scores simply report performance on various scales in terms of how many standard deviations the score is away from the mean. There are several types of standardized scores. The main types of scores we will examine next include *z*-scores, *T*-scores, stanines, normal curve equivalent (NCE) scores, and deviation IQ scores. These various types of standard scores and their relation to the normal distribution are depicted in Figure 6.4. As you can see in this figure, these scores are essentially analogous to one another; they are simply being reported on different scales.

Linear Standard Scores

A *linear standard score* tells how far a raw score is located from the mean of the norm group with the distance being expressed in standard deviation units (Nitko, 2004). Generally speaking, a distribution of linear standard scores will have the same shape as the distribution of raw scores from which the standard scores were derived. This is not the case for percentile ranks, grade- or age-equivalent scores, or nonlinear standard scores (which we will examine shortly). These types of scores are often used to make two distributions (e.g., scores from a science test and those from a mathematics test) more comparable by placing them on the same numerical scale (Nitko, 2004). These standard scores are called linear because if you were to plot in a graph each raw score against its corresponding linear standard score and then connect the resulting points, it would always form a straight line (Nitko, 2004).

Z-scores. This type of norm-referenced, linear standard score is typically referred to as the most basic standard score (Gredler, 1999). *Z-scores* exist on a continuum, where more than 99% of the scores range from –3.00 to +3.00. The sign indicates whether the raw score is above or below the mean; the numerical value indicates how many standard deviations it is located away from the mean. A student's *z*-score is calculated in the following manner:

Figure 6.4 Comparison of Various Types of Standard Scores and Their Relation to the Normal Distribution



(1) The mean of the set of scores is subtracted from the student's raw score.

(2) The resulting value is then divided by the standard deviation for the set of scores.

Assume that the administration of a standardized test resulted in a mean score equal to 75 and a standard deviation equal to 8. A student whose raw score is 75 would receive a *z*-score equal to zero (i.e., her score is zero standard deviation units away from the mean). Another student whose raw score is 91 receives a *z*-score of +2.00 (i.e., 2 standard deviation units above the mean). Finally, a student who earns a raw score of 63 would receive a *z*-score of –1.50 (i.e., 1.5 standard deviation units below the mean).

One distinct disadvantage of *z*-scores is that by definition half the students will receive scores below the mean (see Figure 6.4). In other words, they will receive *z*-scores with negative values. It is very difficult to explain to students—and to parents—how a student could receive a score equal to –2.50 on a standardized test (Nitko, 2004). Receiving negative scores on an academic achievement test can also have adverse effects on a student's level of motivation. Understanding the proper interpretation requires knowledge of the mean, standard deviation, and norm-referencing in general.

T-scores. One way that this problematic characteristic of half of the students receiving negative scores can be overcome is through the use of *T-scores*. A *T*-score—also sometimes referred to as an *SS-score*—provides the location of a raw score in a distribution that has a mean of 50 and a standard deviation of 10 (Chase, 1999; Gredler, 1999). In addition to eliminating the possibility of having negative scores, the fractional portion of the score is also removed through the use of *T*-scores. Using the *z*-score scale as our guide, more than 99% of the *T*-scores on a standardized test will range from 20 (3 standard deviations below the mean) to 80 (3 standard deviations above the mean). A student's *T*-score is calculated in the following manner:

(1) A *z*-score is first calculated and then multiplied by 10 (this becomes the value for a standard deviation on the new scale).

(2) The resulting value is added to 50 (the new value for the mean) to obtain the *T*-score.

If we examine the hypothetical example from our previous discussion of *z*-scores, the first student's *z*-score of zero would equate to a *T*-score of 50; the second student (*z*-score = +2.00) would have a *T*-score of 70; and the third (*z*-score = −1.50) would have a *T*-score of 35. Furthermore, a student who obtains a *z*-score equal to +0.80 would have an equivalent *T*-score equal to 58. Thus, you can see how both negative and fractional scores have been eliminated. This comparison of *z*-scores and *T*-scores can be seen in Figure 6.4.

Although *T*-scores offer an improvement over *z*-scores, they too can be misinterpreted. Since they range from approximately 20 to 80 (i.e., the mean plus or minus 3 standard deviations), they are often confused for percentages. A *T*-score of 60 (i.e., one standard deviation above the mean) can be misinterpreted as meaning that a student answered 60% of the items correctly. Once again, this is essentially a criterion-referenced interpretation; *T*-scores provide norm-referenced information.

Normalized Standard Scores

Test publishers will also transform raw scores to a new set of scores that is distributed normally (or very close to a normal distribution) regardless of the shape of the distribution of the original set of raw scores. This type of transformation actually changes the shape of the distribution by making it conform to a normal distribution. Once the shape has been altered, various types of standard scores can be derived. Each of these types of scores will then have appropriate normal, bell-shaped curve interpretations (Nitko, 2004). These derived scores are collectively known as *normalized standard scores*, also sometimes called area transformations as opposed to linear transformations. Area transformations are so-called because the goal of the transformation is to obtain the same area beneath a curve representing the same distribution of scores as is found in a normal distribution. Following are discussions of several commonly used types of normalized standard scores.

Stanines. Stanines comprise a very common type of score scale on which to report norm-referenced performance, but do so by representing a band of scores as opposed in precise score values (Chase, 1999). A stanine (short for *sta*ndard *nine*) provides the location of a raw score in a specific segment of the normal distribution (Nitko, 2004). Furthermore, stanines range in value from 1 (i.e., the extreme low end) to 9 (i.e., the extreme high end), where the mean is equal to 5 and the standard deviation is equal to 2 (see Figure 6.4). Each band actually spans one half of a standard deviation (Chase, 1999).

All individuals falling in a specific interval are assigned the stanine number of that interval (Nitko, 2004). For example, individuals with percentile ranks ranging from 40 to 59 fall into stanine 5; those with percentile ranks from 60 to 76 would be assigned to stanine 6; and so on. This relationship between stanines and percentile ranks can be seen in Figure 6.4. Stanines can typically be interpreted in the following manner: stanine scores of 1, 2, and 3 indicate below average performance; scores of 4, 5, and 6 indicate average performance; and scores of 7, 8, and 9 indicate above average performance (Airasian, 2005).

The main disadvantage of stanines is that they represent more coarse groupings of scores, especially when compared to percentile ranks (Nitko, 2004). However, a stanine is likely a more accurate estimate of the student's achievement because it represents a band or range within which the student's test performance truly belongs (Gredler, 1999), as opposed to a precise estimate of the student's performance. An individual's stanine score is calculated in the following manner:

(1) A *z*-score is first calculated and then multiplied by 2 (this is the value for a standard deviation on the new scale).

(2) The resulting value is then added to 5 (the new value for the mean) to obtain the stanine score.

SAT/GRE Scores. The scores resulting from both the Scholastic Assessment Test (SAT) and Graduate Record Examination (GRE) are reported on yet a different type of scale, although the scores convey the same basic information. The *SAT/GRE* *scores* (also known as CEEB scores, for the College Entrance Examination Board, which originally developed them) are reported on a scale that has a mean of 500 and a standard deviation of 100 (see Figure 6.4). Once again, possible scores on the SAT and GRE range from a low of 200 (i.e., 3 standard deviations below the mean) to a high of 800 (i.e., 3 standard deviations above the mean). A student's SAT or GRE score is calculated in the following manner:

(1) A *z*-score is first calculated and then multiplied by 100 (this becomes the value for a standard deviation on the new scale).

(2) The resulting value is added to 500 (the new value for the mean) to obtain the SAT or GRE score.

Normal Curve Equivalent Scores. These normalized standard scores, also known as normal curve equivalent scores (NCE scores), have a mean of 50 and a standard deviation of 21.06. Similar to percentile ranks, NCE scores range from 1 to 99. The somewhat odd value for the standard deviation has been established so that NCE scores will precisely match percentile ranks at 3 specific points: 1, 50, and 99 (Chase, 1999; Oosterhof, 2001), as can be seen in Figure 6.4. The basic advantage of NCE scores is that they represent equal units across the entire continuum (i.e., 1 to 99), unlike percentile ranks (Chase, 1999; Oosterhof, 2001). NCE scores are calculated in a similar fashion to the scores previously discussed:

(1) A *z*-score is calculated and multiplied by 21.06 (the new value for a standard deviation).

(2) The value of 50 (the new value for the mean) is added to the resulting value in order to obtain the NCE score.

Deviation IQ Scores. A final type of standardized score, used primarily with assessments of mental ability, is a deviation IQ score (Nitko, 2004). *Deviation IQ scores* provide the location of a raw score in a normal distribution having a mean of 100 and a standard deviation equal to 15 or 16 (depending on the specific test). For a test with a standard deviation set at 15, an individual's deviation IQ score is calculated in the following manner:

(1) A *z*-score is first calculated and then multiplied by 15.

(2) The value of 100 is added to the resulting value to obtain the deviation IQ score.

Other "Unique" Scales. The basic advantage of all of these standard scores is that raw scores can be converted directly to scores related to the normal curve and to percentile ranks (McMillan, 2004). However, this advantage also tends to make things difficult for those of us who try to interpret these various test scores from the standpoint that there are so many of them to consider. In addition, many test publishers use their own unique standard score scales. Once you understand the nature of those scores, you should be able, with relative ease, to interpret those scores. You may have to examine the technical manual or norms books to find out what the publisher has used as the mean and standard deviation for a given type of scaled score (McMillan, 2004).

Shown in Figure 6.5 is a sample Group Report from the Stanford Achievement Test (SAT10). Note the inclusion of several norm-referenced scores for this group including a unique scaled score, national percentile ranks, national stanines, normal curve equivalent scores, and grade-equivalent scores.

A Final Note About Interpreting Norm-Referenced Scores

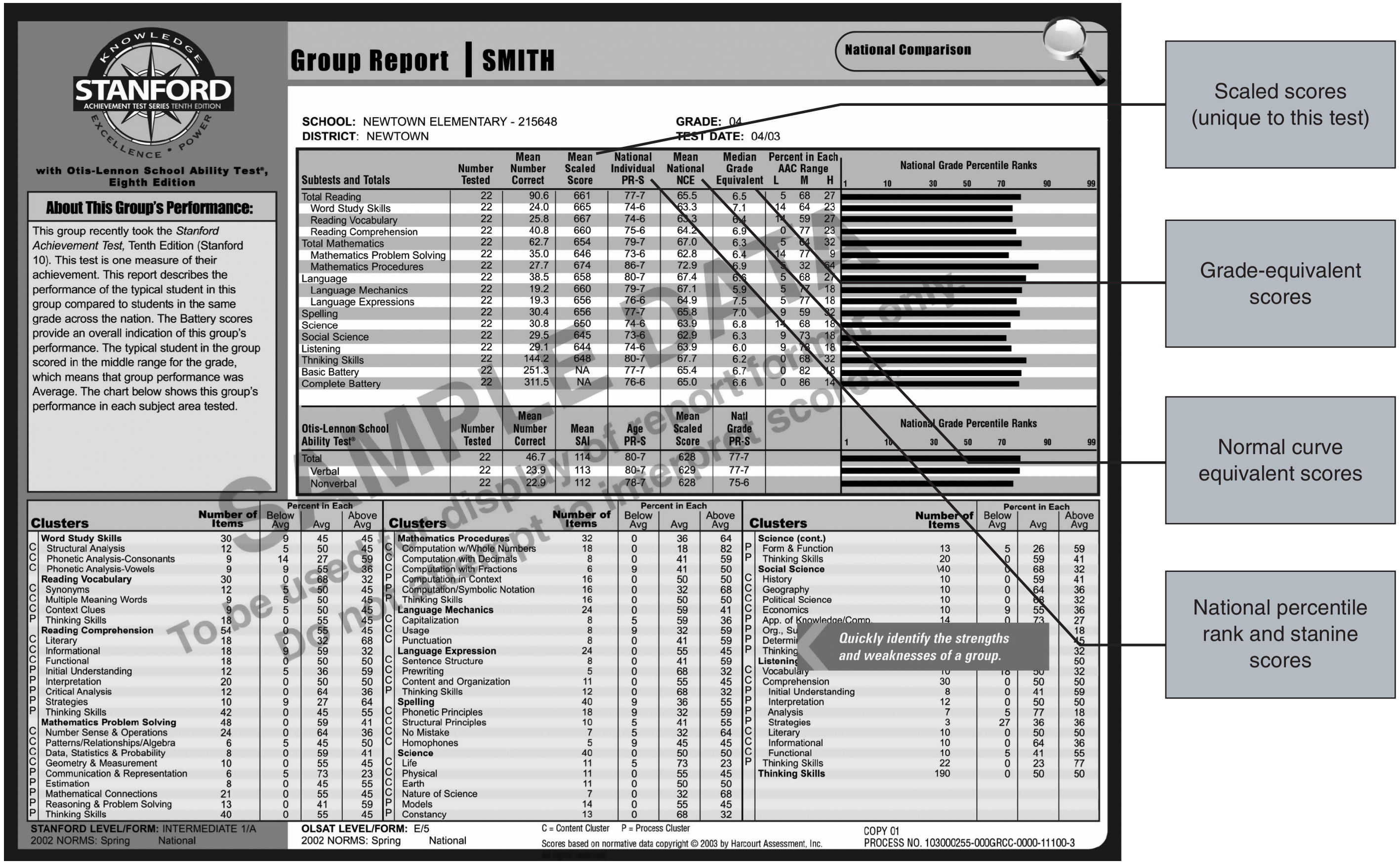
Figure 6.4 shows the relative correspondence between the normal distribution and the various standard score scales we have discussed. From this figure, it should be somewhat clear that all norm-referenced scores provide essentially identical information concerning the location of an individual raw score within a distribution; they simply do so on different scales. In this figure, it is also important to notice the unequal nature of percentile ranks, as well as the 1st and 9th stanines, which represent much larger bands than the other stanines. It really does not matter which specific norm-referenced score you choose to interpret, for they all provide the same information about a particular student's test performance.

Two final sample test reports are provided in Figures 6.6 and 6.7. The ITBS List of Student Scores (see Figure 6.6) provides norm-referenced information for each student in a class. Specifically, a scaled score, grade-equivalent score, stanine, and percentile rank are listed for each subtest taken by the student. Similar information is provided in the List Report of Student Scores from the Gates-MacGinite Reading Tests (GMRT), as shown in Figure 6.7.

STANDARD ERROR OF MEASUREMENT AND CONFIDENCE INTERVALS

When interpreting student performance on norm-referenced measures, it is important to remember that no educational assessment is perfect. Error exists in all test scores (Airasian, 2005). A test is given at a specific moment in time, and a variety of factors can affect—both positively and negatively—students' test scores. For example, these factors might include that a student was ill on the day of the test or recently experienced a traumatic event, such as a death in the family. These types of events would likely result in lowered student performance. In contrast, a student might be exceptionally good at guessing, which would result in performance

Figure 6.5 Sample Group Test Score Report



KNOWLEDGE
STANFORD
ACHIEVEMENT TEST SERIES TENTH EDITION
EXCELLENCE • POWER
National Comparison
National Grade Percentile Ranks
1
10
30
50
70
90
99

with Otis-Lennon School Ability Test®, Eighth Edition

Group Report | SMITH

About This Group's Performance:

This group recently took the *Stanford Achievement Test*, Tenth Edition (Stanford 10). This test is one measure of their achievement. This report describes the performance of the typical student in this group compared to students in the same grade across the nation. The Battery scores provide an overall indication of this group's performance. The typical student in the group scored in the middle range for the grade, which means that group performance was Average. The chart below shows this group's performance in each subject area tested.

SCHOOL: NEWTOWN ELEMENTARY - 215648
DISTRICT: NEWTOWN

GRADE: 04
TEST DATE: 04/03

Subtests and Totals	Number Tested	Mean Number Correct	Mean Scaled Score	National Individual PR-S	Mean National NCE	Median Grade Equivalent	Percent in Each AAC Range L	M	H
Total Reading	22	90.6	661	77-7	65.5	6.5	5	68	27
Word Study Skills	22	24.0	665	74-6	63.3	7.1	14	64	23
Reading Vocabulary	22	25.8	667	74-6	63.3	6.4	14	59	27
Reading Comprehension	22	40.8	660	75-6	64.2	6.9	0	77	23
Total Mathematics	22	62.7	654	79-7	67.0	6.3	5	64	32
Mathematics Problem Solving	22	35.0	646	73-6	62.8	6.4	14	77	9
Mathematics Procedures	22	27.7	674	86-7	72.9	6.9	5	32	64
Language	22	38.5	658	80-7	67.4	6.6	5	68	27
Language Mechanics	22	19.2	660	79-7	67.1	5.9	5	77	18
Language Expressions	22	19.3	656	76-6	64.9	7.5	5	77	18
Spelling	22	30.4	656	77-7	65.8	7.0	9	59	32
Science	22	30.8	650	74-6	63.9	6.8	14	68	18
Social Science	22	29.5	645	73-6	62.9	6.3	9	73	18
Listening	22	29.1	644	74-6	63.9	6.0	9	73	18
Thniking Skills	22	144.2	648	80-7	67.7	6.2	0	68	32
Basic Battery	22	251.3	NA	77-7	65.4	6.7	0	82	18
Complete Battery	22	311.5	NA	76-6	65.0	6.6	0	86	14

Otis-Lennon School Ability Test®	Number Tested	Mean Number Correct	Mean SAI	Age PR-S	Mean Scaled Score	Natl Grade PR-S
Total	22	46.7	114	80-7	628	77-7
Verbal	22	23.9	113	80-7	629	77-7
Nonverbal	22	22.9	112	78-7	628	75-6

	Clusters	Number of Items	Percent in Each Below Avg	Avg	Above Avg
	Word Study Skills	30	9	45	45
C	Structural Analysis	12	5	50	45
C	Phonetic Analysis-Consonants	9	14	27	59
C	Phonetic Analysis-Vowels	9	9	55	36
	Reading Vocabulary	30	0	68	32
C	Synonyms	12	5	50	45
C	Multiple Meaning Words	9	5	50	45
C	Context Clues	9	5	50	45
P	Thinking Skills	18	0	55	45
	Reading Comprehension	54	0	55	45
C	Literary	18	0	32	68
C	Informational	18	9	59	32
C	Functional	18	0	50	50
P	Initial Understanding	12	5	36	59
P	Interpretation	20	0	50	50
P	Critical Analysis	12	0	64	36
P	Strategies	10	9	27	64
P	Thinking Skills	42	0	45	55
	Mathematics Problem Solving	48	0	59	41
C	Number Sense & Operations	24	0	64	36
C	Patterns/Relationships/Algebra	6	5	45	50
C	Data, Statistics & Probability	8	0	59	41
C	Geometry & Measurement	10	0	55	45
P	Communication & Representation	6	5	73	23
P	Estimation	8	0	45	55
P	Mathematical Connections	21	0	55	45
P	Reasoning & Problem Solving	13	0	41	59
P	Thinking Skills	40	0	55	45

	Clusters	Number of Items	Percent in Each Below Avg	Avg	Above Avg
	Mathematics Procedures	32	0	36	64
C	Computation w/Whole Numbers	18	0	18	82
C	Computation with Decimals	8	0	41	59
C	Computation with Fractions	6	9	41	50
P	Computation in Context	16	0	50	50
P	Computation/Symbolic Notation	16	0	32	68
P	Thinking Skills	16	0	50	50
	Language Mechanics	24	0	59	41
C	Capitalization	8	5	59	36
C	Usage	8	9	32	59
C	Punctuation	8	0	41	59
	Language Expression	24	0	55	45
C	Sentence Structure	8	0	41	59
C	Prewriting	5	0	68	32
C	Content and Organization	11	0	55	45
P	Thinking Skills	12	0	68	32
	Spelling	40	9	36	55
C	Phonetic Principles	18	9	32	59
C	Structural Principles	10	5	41	55
C	No Mistake	7	5	32	64
C	Homophones	5	9	45	45
	Science	40	0	50	50
C	Life	11	5	73	23
C	Physical	11	0	55	45
C	Earth	11	0	50	50
C	Nature of Science	7	0	32	68
P	Models	14	0	55	45
P	Constancy	13	0	68	32

	Clusters	Number of Items	Percent in Each Below Avg	Avg	Above Avg
	Science (cont.)				
P	Form & Function	13	5	26	59
P	Thinking Skills	20	0	59	41
	Social Science	\40	0	68	32
C	History	10	0	59	41
C	Geography	10	0	64	36
C	Political Science	10	0	68	32
C	Economics	10	9	55	36
P	App. of Knowledge/Comp.	14	0	73	27
P	Org., Su	[illegible]	[illegible]	[illegible]	18
P	Determi	[illegible]	[illegible]	[illegible]	45
P	Thinking	[illegible]	[illegible]	[illegible]	32
	Listening	[illegible]	[illegible]	[illegible]	50
C	Vocabulary	10	18	50	32
C	Comprehension	30	0	50	50
P	Initial Understanding	8	0	41	59
P	Interpretation	12	0	50	50
P	Analysis	7	5	77	18
P	Strategies	3	27	36	36
C	Literary	10	0	50	50
C	Informational	10	0	64	36
C	Functional	10	5	41	55
P	Thinking Skills	22	0	23	77
	Thinking Skills	190	0	50	50

Quickly identify the strengths and weaknesses of a group.

STANFORD LEVEL/FORM: INTERMEDIATE 1/A
2002 NORMS: Spring National

OLSAT LEVEL/FORM: E/5
2002 NORMS: Spring National

C = Content Cluster P = Process Cluster

Scores based on normative data copyright © 2003 by Harcourt Assessment, Inc.

COPY 01
PROCESS NO. 103000255-000GRCC-0000-11100-3

Scaled scores (unique to this test)

Grade-equivalent scores

Normal curve equivalent scores

National percentile rank and stanine scores

SOURCE: Copyright © 2003 by Harcourt Assessment, Inc. Sample "Group Report" from the Stanford Achievement Test (SAT10). Reproduced with permission of the publisher. All rights reserved.

above the true ability or achievement level of the student. Norm-referenced test scores, which factor in this measurement error, are also often included on test reports. The concepts of *standard error of measurement (SEM)* and *confidence intervals* are important in understanding how to interpret these scores.

A SEM, also known simply as a *standard error,* is the average amount of measurement error across students in the norm group. It is basically interpreted as the standard deviation of all errors in measurement. If the standard error is both added to and subtracted from the score a student receives on a standardized test, a range of student performance can be defined. This range serves as an estimate within which the student's true performance most likely lies. In other words,

True Score Range = Scaled Score +/− (i.e., plus and minus) Standard Error

This true score range is known as a *confidence interval* or confidence band. The purpose of confidence intervals is to establish a range of scores that we are *reasonably* confident includes the student's true ability or achievement score (Gredler, 1999). Recall from Figure 6.2 that the mean plus and minus one standard deviation contains approximately 68% of the individuals in a normal distribution of test scores. A somewhat related interpretation can be made for confidence intervals. For example, assume that the standard error for a given standardized test is calculated to be equal to 3.5—in other words, this is the average of all errors in measurement across all students in the norm group. Further assume that a student receives a score of 64 on the test. The resulting confidence interval (based on the addition and subtraction of one standard error; i.e., 64 +/− 3.5) for that student would be 60.5 to 67.5. This confidence interval is appropriately interpreted in the following manner: If it were possible to test the student repeatedly under ideal conditions, 68% of this student's scores would fall within this interval (Gredler, 1999). In other words, 68% of the student's possible scores would be located between 60.5 and 67.5. A student's obtained score plus and minus one standard error is sometimes referred to as the 68% confidence interval. An alternative interpretation is to say that *we* are 68% confident that the student's true ability score lies between 60.5 and 67.5.

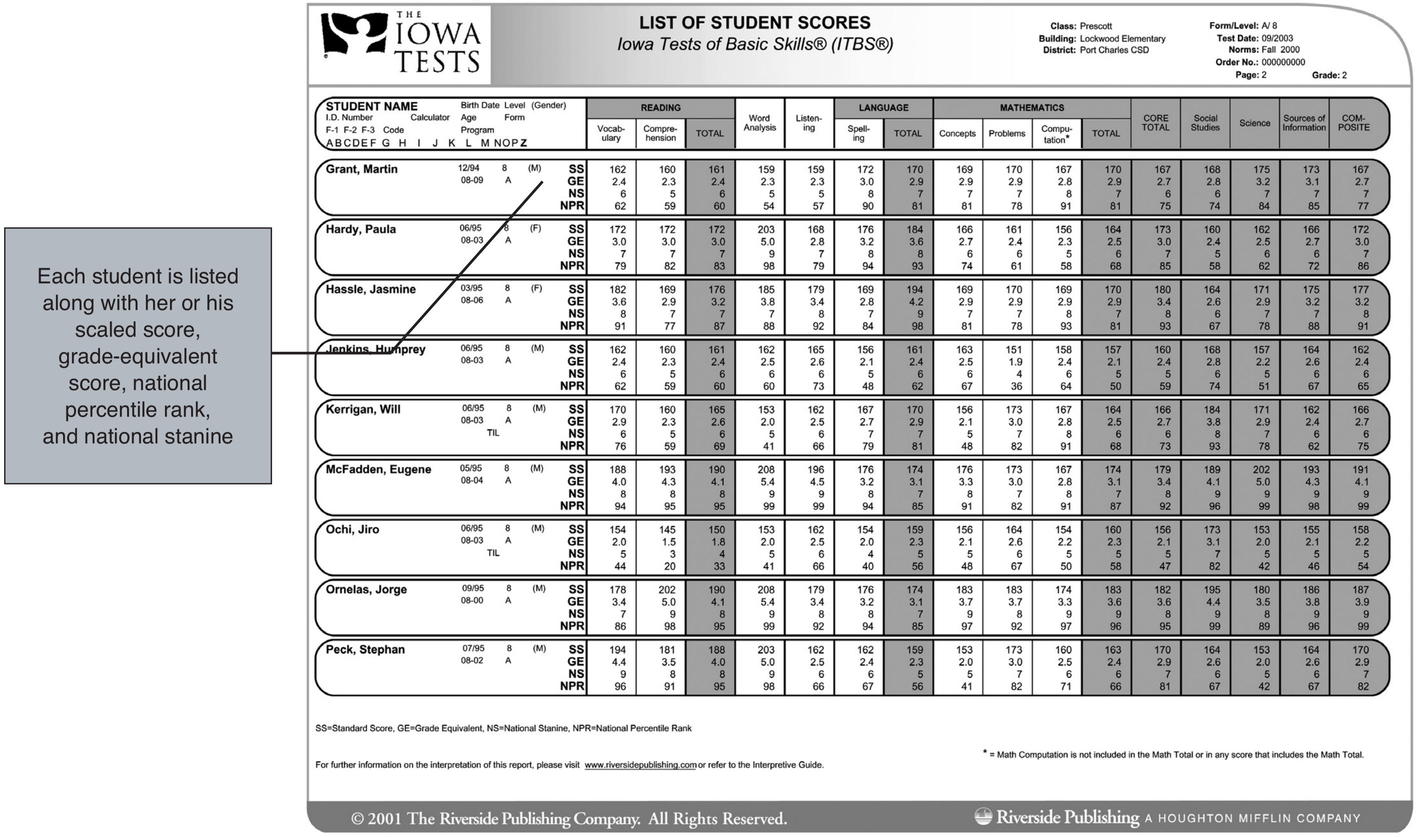
Since the interpretation of standard errors and the resulting confidence intervals are based on a normal distribution, we can generalize our example to provide various statements regarding the precision of the student's test scores. Again, using Figure 6.2 as a reference, we could conclude the following:

(1) We can be approximately 68% confident that the student's true scores lie in the range of 60.5 to 67.5 (i.e., within one standard error, or 64 +/− 3.5).

(2) We can be approximately 96% confident that the student's true scores lie in the range of 57 to 71 [i.e., within two standard errors, or 64 +/− (2)(3.5)].

(3) We can be approximately 99% confident that the student's true scores lie in the range of 53.5 to 74.5 [i.e., within three standard errors, or 64 +/− (3)(3.5)].

Figure 6.6 Sample Group Test Score Report



Each student is listed along with her or his scaled score, grade-equivalent score, national percentile rank, and national stanine

THE IOWA TESTS

LIST OF STUDENT SCORES
Iowa Tests of Basic Skills® (ITBS®)

Class: Prescott
Building: Lockwood Elementary
District: Port Charles CSD

Form/Level: A/ 8
Test Date: 09/2003
Norms: Fall 2000
Order No.: 000000000
Page: 2 **Grade:** 2

STUDENT NAME Birth Date Level (Gender) I.D. Number Calculator Age Form F-1 F-2 F-3 Code Program ABCDEF G H I J K L M NOPZ		READING			Word Analysis	Listen-ing	LANGUAGE		MATHEMATICS				CORE TOTAL	Social Studies	Science	Sources of Information	COM-POSITE
		Vocab-ulary	Compre-hension	TOTAL			Spell-ing	TOTAL	Concepts	Problems	Compu-tation*	TOTAL					
Grant, Martin 12/94 8 (M)	**SS**	162	160	161	159	159	172	170	169	170	167	170	167	168	175	173	167
08-09 A	**GE**	2.4	2.3	2.4	2.3	2.3	3.0	2.9	2.9	2.9	2.8	2.9	2.7	2.8	3.2	3.1	2.7
	NS	6	5	6	5	5	8	7	7	7	8	7	6	6	7	7	7
	NPR	62	59	60	54	57	90	81	81	78	91	81	75	74	84	85	77
Hardy, Paula 06/95 8 (F)	**SS**	172	172	172	203	168	176	184	166	161	156	164	173	160	162	166	172
08-03 A	**GE**	3.0	3.0	3.0	5.0	2.8	3.2	3.6	2.7	2.4	2.3	2.5	3.0	2.4	2.5	2.7	3.0
	NS	7	7	7	9	7	8	8	6	6	5	6	7	5	6	6	7
	NPR	79	82	83	98	79	94	93	74	61	58	68	85	58	62	72	86
Hassle, Jasmine 03/95 8 (F)	**SS**	182	169	176	185	179	169	194	169	170	169	170	180	164	171	175	177
08-06 A	**GE**	3.6	2.9	3.2	3.8	3.4	2.8	4.2	2.9	2.9	2.9	2.9	3.4	2.6	2.9	3.2	3.2
	NS	8	7	7	7	8	7	9	7	7	8	7	8	6	7	7	8
	NPR	91	77	87	88	92	84	98	81	78	93	81	93	67	78	88	91
Jenkins, Humprey 06/95 8 (M)	**SS**	162	160	161	162	165	156	161	163	151	158	157	160	168	157	164	162
08-03 A	**GE**	2.4	2.3	2.4	2.5	2.6	2.1	2.4	2.5	1.9	2.4	2.1	2.4	2.8	2.2	2.6	2.4
	NS	6	5	6	6	6	5	6	6	4	6	5	5	6	5	6	6
	NPR	62	59	60	60	73	48	62	67	36	64	50	59	74	51	67	65
Kerrigan, Will 06/95 8 (M)	**SS**	170	160	165	153	162	167	170	156	173	167	164	166	184	171	162	166
08-03 A	**GE**	2.9	2.3	2.6	2.0	2.5	2.7	2.9	2.1	3.0	2.8	2.5	2.7	3.8	2.9	2.4	2.7
TIL	**NS**	6	5	6	5	6	7	7	5	7	8	6	6	8	7	6	6
	NPR	76	59	69	41	66	79	81	48	82	91	68	73	93	78	62	75
McFadden, Eugene 05/95 8 (M)	**SS**	188	193	190	208	196	176	174	176	173	167	174	179	189	202	193	191
08-04 A	**GE**	4.0	4.3	4.1	5.4	4.5	3.2	3.1	3.3	3.0	2.8	3.1	3.4	4.1	5.0	4.3	4.1
	NS	8	8	8	9	9	8	7	8	7	8	7	8	9	9	9	9
	NPR	94	95	95	99	99	94	85	91	82	91	87	92	96	99	98	99
Ochi, Jiro 06/95 8 (M)	**SS**	154	145	150	153	162	154	159	156	164	154	160	156	173	153	155	158
08-03 A	**GE**	2.0	1.5	1.8	2.0	2.5	2.0	2.3	2.1	2.6	2.2	2.3	2.1	3.1	2.0	2.1	2.2
TIL	**NS**	5	3	4	5	6	4	5	5	6	5	5	5	7	5	5	5
	NPR	44	20	33	41	66	40	56	48	67	50	58	47	82	42	46	54
Ornelas, Jorge 09/95 8 (M)	**SS**	178	202	190	208	179	176	174	183	183	174	183	182	195	180	186	187
08-00 A	**GE**	3.4	5.0	4.1	5.4	3.4	3.2	3.1	3.7	3.7	3.3	3.6	3.6	4.4	3.5	3.8	3.9
	NS	7	9	8	9	8	8	7	9	8	9	9	8	9	8	9	9
	NPR	86	98	95	99	92	94	85	97	92	97	96	95	99	89	96	99
Peck, Stephan 07/95 8 (M)	**SS**	194	181	188	203	162	162	159	153	173	160	163	170	164	153	164	170
08-02 A	**GE**	4.4	3.5	4.0	5.0	2.5	2.4	2.3	2.0	3.0	2.5	2.4	2.9	2.6	2.0	2.6	2.9
	NS	9	8	8	9	6	6	5	5	7	6	6	7	6	5	6	7
	NPR	96	91	95	98	66	67	56	41	82	71	66	81	67	42	67	82

SS=Standard Score, GE=Grade Equivalent, NS=National Stanine, NPR=National Percentile Rank

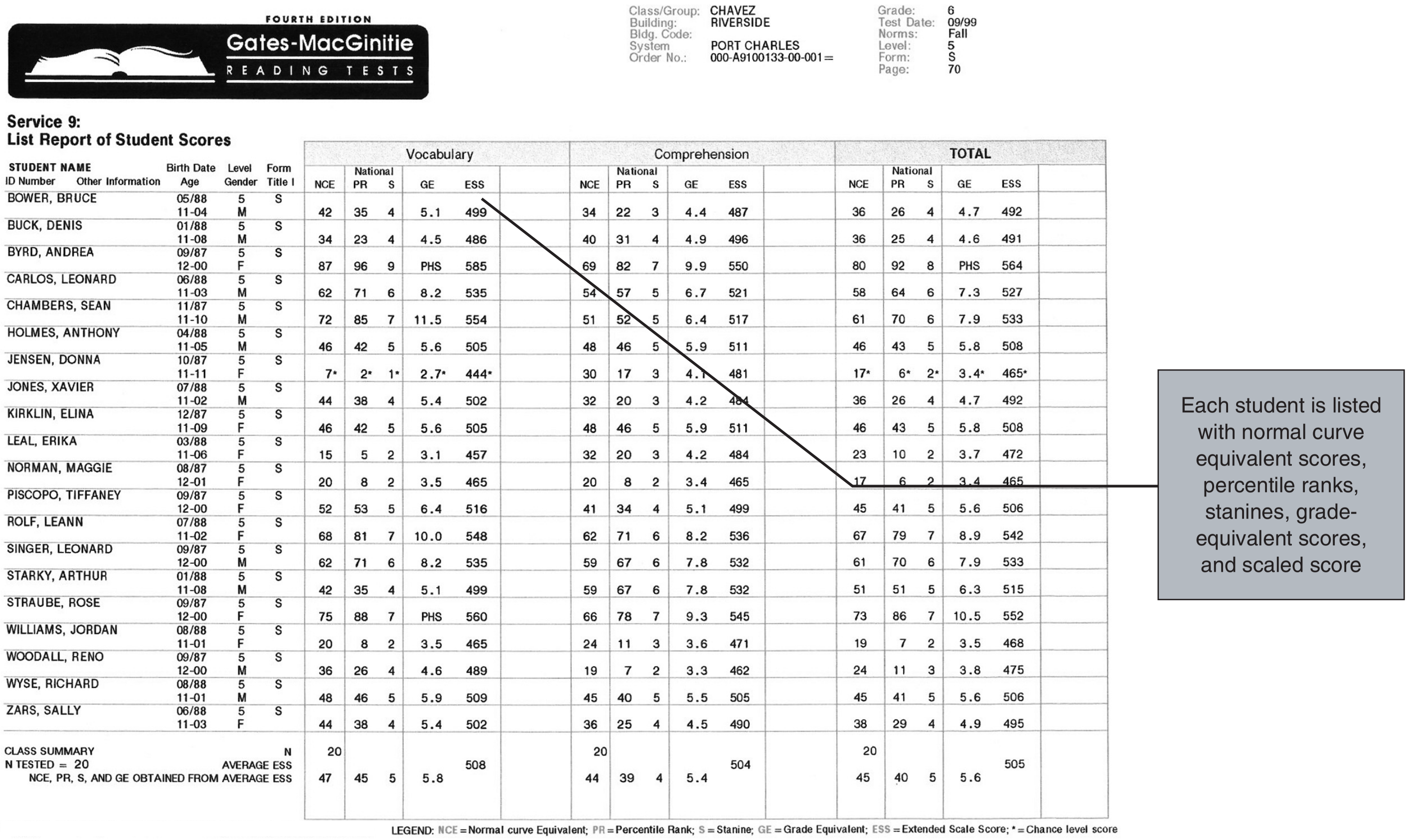
* = Math Computation is not included in the Math Total or in any score that includes the Math Total.

For further information on the interpretation of this report, please visit www.riversidepublishing.com or refer to the Interpretive Guide.

Riverside Publishing A HOUGHTON MIFFLIN COMPANY

SOURCE: Copyright © 2001 by Riverside Publishing. Sample "List of Student Scores" from the Iowa Tests of Basic Skills (ITBS). Reproduced with permission of the publisher. All rights reserved.

Figure 6.7 Sample Group Test Score Report



FOURTH EDITION Gates-MacGinitie READING TESTS

Class/Group:	CHAVEZ
Building:	RIVERSIDE
Bldg. Code:	
System	PORT CHARLES
Order No.:	000-A9100133-00-001=

Grade:	6
Test Date:	09/99
Norms:	Fall
Level:	5
Form:	S
Page:	70

Service 9:
List Report of Student Scores

STUDENT NAME ID Number Other Information	Birth Date Age	Level Gender	Form Title I	Vocabulary					Comprehension					TOTAL				
				NCE	National PR	S	GE	ESS	NCE	National PR	S	GE	ESS	NCE	National PR	S	GE	ESS
BOWER, BRUCE	05/88 11-04	5 M	S	42	35	4	5.1	499	34	22	3	4.4	487	36	26	4	4.7	492
BUCK, DENIS	01/88 11-08	5 M	S	34	23	4	4.5	486	40	31	4	4.9	496	36	25	4	4.6	491
BYRD, ANDREA	09/87 12-00	5 F	S	87	96	9	PHS	585	69	82	7	9.9	550	80	92	8	PHS	564
CARLOS, LEONARD	06/88 11-03	5 M	S	62	71	6	8.2	535	54	57	5	6.7	521	58	64	6	7.3	527
CHAMBERS, SEAN	11/87 11-10	5 M	S	72	85	7	11.5	554	51	52	5	6.4	517	61	70	6	7.9	533
HOLMES, ANTHONY	04/88 11-05	5 M	S	46	42	5	5.6	505	48	46	5	5.9	511	46	43	5	5.8	508
JENSEN, DONNA	10/87 11-11	5 F	S	7*	2*	1*	2.7*	444*	30	17	3	4.1	481	17*	6*	2*	3.4*	465*
JONES, XAVIER	07/88 11-02	5 M	S	44	38	4	5.4	502	32	20	3	4.2	484	36	26	4	4.7	492
KIRKLIN, ELINA	12/87 11-09	5 F	S	46	42	5	5.6	505	48	46	5	5.9	511	46	43	5	5.8	508
LEAL, ERIKA	03/88 11-06	5 F	S	15	5	2	3.1	457	32	20	3	4.2	484	23	10	2	3.7	472
NORMAN, MAGGIE	08/87 12-01	5 F	S	20	8	2	3.5	465	20	8	2	3.4	465	17	6	2	3.4	465
PISCOPO, TIFFANEY	09/87 12-00	5 F	S	52	53	5	6.4	516	41	34	4	5.1	499	45	41	5	5.6	506
ROLF, LEANN	07/88 11-02	5 F	S	68	81	7	10.0	548	62	71	6	8.2	536	67	79	7	8.9	542
SINGER, LEONARD	09/87 12-00	5 M	S	62	71	6	8.2	535	59	67	6	7.8	532	61	70	6	7.9	533
STARKY, ARTHUR	01/88 11-08	5 M	S	42	35	4	5.1	499	59	67	6	7.8	532	51	51	5	6.3	515
STRAUBE, ROSE	09/87 12-00	5 F	S	75	88	7	PHS	560	66	78	7	9.3	545	73	86	7	10.5	552
WILLIAMS, JORDAN	08/88 11-01	5 F	S	20	8	2	3.5	465	24	11	3	3.6	471	19	7	2	3.5	468
WOODALL, RENO	09/87 12-00	5 M	S	36	26	4	4.6	489	19	7	2	3.3	462	24	11	3	3.8	475
WYSE, RICHARD	08/88 11-01	5 M	S	48	46	5	5.9	509	45	40	5	5.5	505	45	41	5	5.6	506
ZARS, SALLY	06/88 11-03	5 F	S	44	38	4	5.4	502	36	25	4	4.5	490	38	29	4	4.9	495
CLASS SUMMARY			N	20					20					20				
N TESTED = 20			AVERAGE ESS					508					504					505
NCE, PR, S, AND GE OBTAINED FROM AVERAGE ESS				47	45	5	5.8		44	39	4	5.4		45	40	5	5.6	

LEGEND: NCE = Normal curve Equivalent; PR = Percentile Rank; S = Stanine; GE = Grade Equivalent; ESS = Extended Scale Score; * = Chance level score

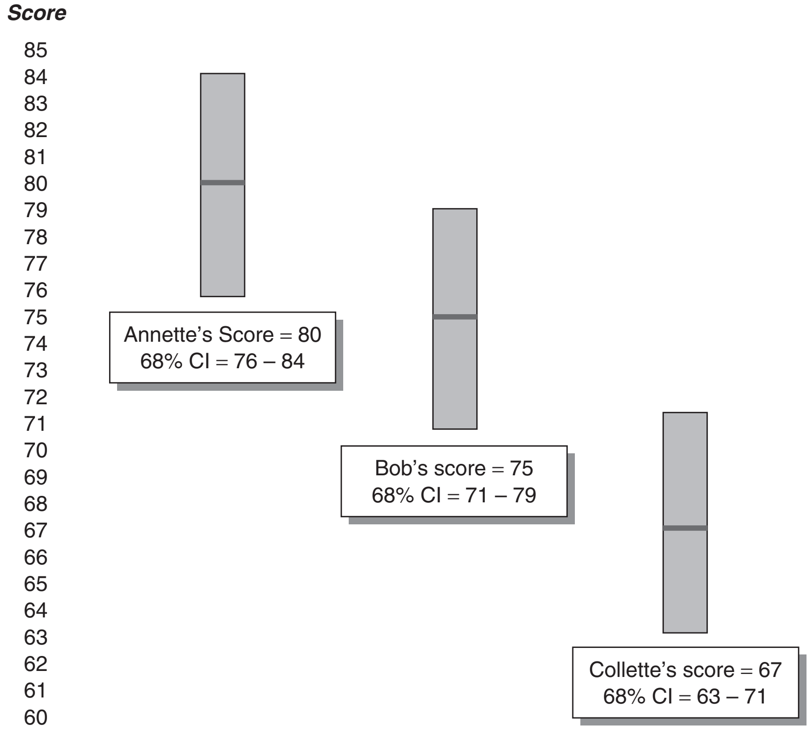
Each student is listed with normal curve equivalent scores, percentile ranks, stanines, grade-equivalent scores, and scaled score

SOURCE: Copyright © 2000 by Riverside Publishing Co. Sample "List Report of Student Scores" from the Gates-MacGinitie Reading Tests (GMRT). Reproduced with permission of the publisher. All rights reserved.

Although we would typically like to be as confident as possible when interpreting test results, confidence and precision have an inverse relationship. In other words, notice that as confidence increases (a good thing), precision decreases (a not-so-good thing). It is not very informative to say that we are 99% sure that a student's true achievement spans a more-than-20-unit range on the scale. The 68% confidence interval is typically seen as a meaningful compromise between confidence and precision.

On norm-referenced test reports, confidence intervals are typically presented around a student's obtained percentile rank scores. These are often referred to as *national percentile bands*. Figure 6.8 shows a hypothetical example of three students' scores, including their respective confidence bands. For each student, notice that the obtained score is located in the middle of the band, although this will not always be the case since, as you know, percentile ranks do not represent equal units. It is further important to note that Annette's performance is clearly better than that of Collette. However, Bob and Annette could actually be performing at nearly the same level, or Bob's performance could even be above that of

Figure 6.8 Examples of Percentile Rank Confidence Bands (Standard Error = 4)



Annette because their bands overlap. In other words, when the bands overlap, there is no real difference between estimates of the true achievement levels for students. This is true both across students and across subtests for a given student.

This interpretation can be extended to comparisons of the relative performances on various subtest scores for an individual student. For example, the national percentile rank bands for the ITBS are provided for an individual student, as shown in Figure 6.3. When examining the national percentile bands for an individual student, it is important to examine the overlap of subtest bands. If the bands for two subtests do not overlap, there is a significant difference in the performance in those areas. In this example, there is a significant difference in test performance between the Listening and Spelling subtests since the bands to not overlap.

Summary

Norm-referenced scores provide test performance information about individuals or groups of students as compared to a representative norm group. Norm groups must be selected such that they are current and representative of the larger population, whether it be a national population or a smaller, more specific one. Often, scores related to the performance of the norm group are based on the normal distribution, which has three consistent characteristics. Within the normal distribution, most individuals are average, with their performance falling roughly near the center of the distribution. Since the distribution is consistent, a fixed percentage of cases or scores can be defined within the distribution.

Norm-referenced scores are based on mathematical transformations of raw scores. Percentile ranks are defined as the percentage of the norm group that scored below a particular raw score. They do not represent equal units and are often misinterpreted as a percentage of items answered correctly (which is a criterion-referenced interpretation). Developmental scales, such as grade-equivalent and age-equivalent scores, seek to identify a student's development across various grade or age levels. Although they can be useful, they are often misinterpreted as standards that all students should be expected to achieve.

Standardized scores are obtained by transforming scores to fit some type of distribution. They allow for comparisons of test performance across two or more different measures. Linear standard scores, including *z*-scores and *T*-scores, maintain the same distributional shape as their corresponding raw scores. However, since they are on the same scale, they allow for comparisons between two distributions of test performance. Normalized standard scores are those that are transformed to convert their raw score distributions to a normal distribution. This permits interpretations based on knowledge of the characteristics of a normal distribution.

These types of scores include stanines, SAT/GRE scores, normal curve equivalent scores, deviation IQ scores, as well as others that may be unique to specific tests. It is important to note that norm-referenced scores all provide the same information about a particular student's test performance.

Also important to remember when interpreting norm-referenced test scores is that because no test is a perfect measure, all scores contain error. This error may serve to increase or decrease a student's true achievement or ability performance. The standard error of measure is the average amount of error across all students in the norm group. This value can be used to estimate the range within which a given student's true test performance probably falls. Furthermore, levels of confidence are attached to these ranges of scores, thus allowing an extension of the interpretation of test performance. Finally, these confidence levels permit the comparison across students or across subtests for a particular student.

Activities for Application and Reflection

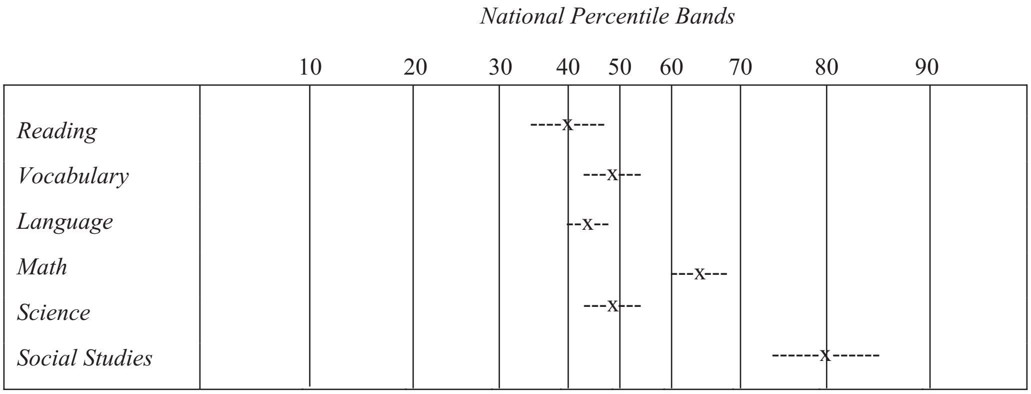
1. Obtain an actual student test report from any test you may use in your school or district that includes some norm-referenced score information. Closely examine the norm-referenced scores and briefly summarize the student's performance. What would you describe as the student's strengths and weaknesses relative to those students in the norm group?

2. If you could use only one type of norm-referenced score as presented in this module to interpret a student's test performance, which one would you choose? Why did you choose that particular score?

3. Paulette receives both norm-referenced and criterion-referenced scores for her performance on a standardized test in science. The norm-referenced scores indicate that she scores at the 85th percentile in "Earth and Space Science," but the criterion-referenced information indicates that she is deficient (i.e., failed to meet the performance criteria) in this area. Is this possible? Why or why not?

4. Using the information provided in Figure 6.4, approximate the values on the scales below that correspond to a *z*-score of + .50:

   a. percentile rank = ?
   b. *T*-score = ?
   c. stanine = ?
   d. NCE score = ?



5. Refer to the diagram above, which shows the 68% confidence intervals for a set of test scores:

    a. Which test has the largest standard error of measure? How do you know?
    b. Which pair(s) of test scores are significantly different from each other?
    c. The following interpretation of the percentile rank is incorrect; rewrite it so that it is accurate.
    The student correctly answered almost 80% of the social studies items.
    d. The following interpretation is incorrect; rewrite it so that it is accurate.
    The student's true reading achievement is at the 40th percentile.

6. Carefully read the interview transcript for Amy Kenyon, first- and second-grade teacher, paying particular attention to her discussion of using stanine scores for the identification of students for special services. Discuss the appropriateness (e.g., the pros and cons) of using stanine scores for this purpose.