CHAPTER

3 Giving Meaning to Scores

THE NATURE OF A SCORE

Quadra Quickly got a score of 44 on her spelling test. What does this score mean, and how should we interpret it?

Standing alone, the number has no meaning at all and is completely uninterpretable. At the most superficial level, we do not even know whether this number represents a perfect score of 44 out of 44 or a very low percentage of the possible score, such as 44 out of 100. Even if we do know that the score is 44 out of 80, or 55%, what then?

Consider the two 20-word spelling tests in Table 3–1. A score of 15 on Test A would have a vastly different meaning from the same score on Test B. A person who gets only 15 correct on Test A would not be outstanding in a second- or third-grade class. Have a few friends or classmates take Test B. You will probably find that not many of them can spell 15 of these words correctly. When this test was given to a class of graduate students, only 22% spelled 15 or more of the words correctly. A score of 15 on Test B is a good score among graduate students of education or psychology.

As it stands, then, knowing that Quadra spelled 44 words correctly, or even that she spelled 55% correctly, has no direct meaning or significance. The score has meaning only

Table 3–1
Two 20-Word Spelling Tests

Test A		Test B	
bar	feet	baroque	feasible
cat	act	catarrh	accommodation
form	rate	formaldehyde	inaugurate
jar	inch	jardiniere	insignia
nap	rent	naphtha	deterrent
dish	lip	discernible	eucalyptus
fat	air	fatiguing	questionnaire
sack	rim	sacrilegious	rhythm
rich	must	ricochet	ignoramus
sit	red	citrus	accrued

when we have some standard with which to compare it, some frame of reference within which to interpret it.

Frames of Reference

The way that we derive meaning from a test score depends on the context or frame of reference in which we wish to interpret it. This frame of reference may be described using three basic dimensions. First, there is what we might call a temporal dimension: Is the focus of our concern what a person can do now or what that person is likely to do at some later time? Are we interested in describing the current state or in forecasting the future?

A second dimension involves the contrast between what people *can* do and what they would *like* to do or would *normally* do. When we assess a person's capacity, we determine *maximum performance*, and when we ask about a person's preferences or habits, we assess *typical performance*. Maximum performance implies a set of tasks that can be judged for correctness; there is a "right" answer. With typical performance there is not a right answer, but we may ask whether one individual's responses are like those of most people or are unusual in some way.

A third dimension is the nature of the standard against which we compare a person's behavior. In some cases, the content of the test itself may provide the standard; in some cases, it is the person's own behavior in other situations or on other tests that provides the standard; and in still other instances, it is the person's behavior in comparison with the behavior of other people. Thus, a given measurement is interpreted as being either oriented in the present or oriented in the future; as measuring either maximum or typical performance; and as relating the person's performance to a standard defined by the test itself, to the person's own scores on this or other measures, or to the performance of other people.

Many instructional decisions in schools call for information about what a student or group of students can do now. Wakana Watanabe is making a good many mistakes in her oral reading. To develop an instructional strategy that will help her overcome this difficulty, we need to determine the cause of her problem. One question we might ask is whether she can match words with their initial consonant sounds. A brief test focused on this specific skill, perhaps presented by the

Table 3–2
A Focused Test

Test on Capitalizing Proper Nouns

Directions: Read the paragraph. The punctuation is correct, and the words that begin a sentence have been capitalized. No other words have been capitalized. Some need to be. Draw a line under *each word* that should begin with a capital.

We saw mary yesterday. She said she had gone to chicago, illinois, to see her aunt helen. Her aunt took her for a drive along the shore of lake michigan. On the way they passed the conrad hilton hotel, where mary's uncle joseph works. Mary said she had enjoyed the trip, but she was glad to be back home with her own friends.

teacher to Wakana individually while the other students work on other tasks, can help to determine whether a deficiency in this particular skill is part of her problem. Here the test itself defines the standard against which we will compare Wakana's performance.

We might also want to know how many children in Wakana's class have mastery of the rule on capitalizing proper nouns. A focused test such as the one in Table 3–2 can provide evidence to guide a decision on whether further teaching of this skill is needed. At a broader level, we may ask whether the current program in mathematics in the Centerville school district is producing satisfactory achievement. Administration of a survey mathematics test with national or regional norms can permit a comparison of Centerville's students with students in the rest of the country, and this comparison can be combined with other information about Centerville's students and its schools to make a decision on whether progress is satisfactory.

Whenever we ask questions about how much a person can do, we also face the issue of the purpose of our evaluation. There are two fundamental purposes for evaluating capacity in an educational context. One is to reach a summary statement of the person's accomplishments to date, such as teachers do at the end of each marking period. Evaluation for this purpose is called **summative evaluation.** It provides a summary of student achievement. By contrast, teachers and counselors are often interested in using tests to determine their students' strengths and weaknesses, the areas where they are doing well and those where they are doing poorly. Assessment for this purpose, to guide future instruction, is called **formative evaluation.** Test results are used to inform or to shape the course of instruction.

The type of maximum performance test that describes what a person *has learned to do* is called an **achievement test.** The oral reading test given to Wakana, the capitalization test in Table 3–2, and the mathematics test given to the students in Centerville are illustrations of sharply contrasting types of achievement tests. The test on initial consonant sounds is concerned with mastery of one specific skill by one student, and no question is raised as to whether Wakana's skill in this area is better or worse than that of any other student. The only question is, can she perform this task well enough so that we can rule out deficiency in this skill as a cause of her difficulty with oral reading?

Similarly, Wakana's teacher is concerned with the level of mastery, *within* this class, of a specific skill in English usage. Tests concerned with level of mastery of such defined skills are often called **domain-referenced** or **criterion-referenced tests,** because the focus is solely on reaching a standard of performance on a specific skill called for by the test exercises. The test itself and the domain of content it represents provide the standard. Many, perhaps most, assessments needed for instructional decisions are of this sort.

We may contrast these tests with the mathematics survey test given to appraise mathematics achievement in Centerville. Here, the concern is whether Centerville's students are showing

satisfactory achievement *when compared with the students in other towns and school systems like Centerville.* Performance is evaluated not in relation to the set of tasks per se, but in relation to the performance of some more general reference group. A test used in this way is spoken of as a **norm-referenced test,** because the quality of the performance is defined by comparison with the behavior of others. A norm-referenced test may appropriately be used in many situations calling for curricular, guidance, or research decisions. Occasionally throughout this book, we will compare and contrast criterion-referenced and norm-referenced achievement tests with respect to their construction, desired characteristics, and use.

Some decisions that we need to make require information on what a person *can learn to do.* Will Helen be able to master the techniques of computer programming? How readily will Rahim assimilate calculus? Selection and placement decisions typically involve predictions about future learning or performance, based on the present characteristics of the individual. A test that is used in this way as a predictor of future learning is called an **aptitude test.** Aptitude tests are usually norm referenced.

In some situations, our decision calls for an estimate of what a person *is likely to do.* The selection of bus drivers, police officers, and candidates for many other jobs is best made with an eye to aspects of the person's personality or temperament. We would not want to select someone with a high level of aggression to drive a large vehicle on confined city streets. Nor would we want people who have difficulty controlling their tempers carrying firearms and serving as keepers of the peace. A measure of typical performance can serve as a useful aid in such situations, and these measures usually are also norm referenced.

Note that some of the most effective predictors of future learning or behavior are measures of past learning or behavior. Thus, for both computer programming and calculus, an effective predictor might be a test measuring competence in high school algebra. Such a test would measure previously learned knowledge and skills, but we would be using that achievement measure to predict future learning. Any test, whatever it is called, assesses a person's present characteristics. We cannot directly measure a person's hypothetical "native" or "inborn" qualities. All we can measure is what that person is able and willing to do in the here and now. That information can then be used to evaluate past learning, as when an algebra test is used to decide whether Roxanne should get an A in her algebra course, or to predict future learning, as when a counselor must decide whether Roxanne has a reasonable probability of successfully completing calculus. The distinction between an aptitude and an achievement test often lies more in the purpose for which the test results are used than in the nature or content of the test itself.

Domains in Criterion- and Norm-Referenced Tests

It is important to realize that all achievement tests (in fact, all tests) relate to a more or less well-specified domain of content. The mathematics survey test covers a fairly broad array of topics, while the test on the rules for capitalization is restricted to a narrowly defined set of behaviors. Thus, it is not really appropriate to differentiate between criterion-referenced and norm-referenced tests by saying that the former derive their meaning from a precisely specified domain, while the latter do not. A well-constructed, norm-referenced achievement test will represent a very carefully defined domain, but the domain is generally more diverse than that of a criterion-referenced test, and it has only a small number of items covering a given topic or instructional objective. The criterion-referenced achievement test will represent a narrowly and precisely defined domain and will therefore cover its referent content more thoroughly than will a norm-referenced test of the same length.

There is a second dimension to using information from an achievement test. In addition to the traditional distinction between criterion-referenced and norm-referenced tests on the breadth of the domain they cover, the second dimension relates to the way that the level, or altitude, of

performance is represented or used in reaching decisions. A test score from either type of test gets its content meaning from the domain of content that the test represents, but the kind of inference that a teacher or counselor draws from the score can be either absolute or relative. The teacher makes a judgment on the basis of the test score. If the judgment is that when a student or group of students have gained a particular level of proficiency with respect to the content, the test represents they have mastered the material, then the judgment is an absolute, *mastery–nonmastery* one. The decision reached is either that the students have mastered the material or that they have not; degree of mastery is not an issue. Decisions of this type are called **mastery decisions.** The usual definition of a criterion-referenced test is a test that covers a narrow domain and is used for mastery decisions.

By contrast, teachers can also use tests to judge relative achievement of objectives. Relative mastery involves estimating the percentage of the domain that students have mastered. For example, the teacher may decide that students have mastered an objective relating to spelling when they can spell correctly 19 out of 20 words from the domain. But the same teacher might use the information that the average student got a score of 14 on the spelling test to indicate that the students had achieved about 70% mastery of the domain. We refer to decisions of this kind as **relative achievement decisions,** but the frame of reference is still the domain of content without regard to the performance of anyone other than the current examinees.

The typical norm-referenced test uses neither of these ways to represent the level of performance. Rather, level is referenced to a larger group called a **norm group,** or norm sample. A normative interpretation of a score could lead to the conclusion that the individual was performing at a very high level compared with an appropriate reference group, but the same performance might fall far below mastery from the criterion-referenced perspective. Conversely, a ninth-grader who has achieved mastery of multiplication facts at the level of 95% accuracy ordinarily would not show a high level of performance when compared with other ninth-graders.

CRITERION-REFERENCED EVALUATION

We can approach the problem of a frame of reference for interpreting test results from the two rather different points of view mentioned earlier. One, criterion-referenced evaluation, focuses on the tasks themselves, while the other, norm-referenced testing, focuses on the performance of typical people. Consider the 20 spelling words in Test A of Table 3–1. If we knew that these had been chosen from the words taught in a third-grade spelling program and if we had agreed on some grounds (at this point unspecified) that 80% correct represented an acceptable standard for performance in spelling when words are presented by dictation, with illustrative sentences, then we could interpret Ellen's score of 18 correct on the test as indicating that she had reached the criterion of mastery of the words taught in third-grade spelling and Peter's score of 12 correct as indicating that he had not. Here, we have test content selected from a narrowly defined domain and we have a mastery test interpretation. The test is criterion referenced in that (1) the tasks are drawn from and related to a specific instructional domain, (2) the form of presentation of the tasks and the response to them is set in accordance with the defined objective, and (3) a level of performance acceptable for mastery, with which the performance of each student is compared, is defined in advance. That is, what we call criterion-referenced tests relate to a carefully defined domain of content, they focus on achievement of specific behavioral objectives, and the results are often (but not necessarily) used for mastery judgments.

The "mastery" frame of reference is an appropriate one for some types of educational decisions. For example, decisions on what materials and methods should be used for additional instruction in spelling with Ellen and Peter might revolve around the question of whether they had reached the specified criterion of mastery of the third-grade spelling words. More crucially, in a sequential subject such as mathematics, the decision of whether to begin a unit involving borrowing in subtraction might depend on whether students had reached a criterion of mastery on a test of two-place subtraction that did not require borrowing.

Although the two topics of domain referencing of test content and mastery–nonmastery decisions about achievement historically have been linked, it is important to realize that they are quite different and independent ideas that have come to be treated together. It is also important to realize that both exist in a sociopolitical context that invests them with normative meaning. What, for example, should a third-grader be expected to know about multiplication or spelling? The answer to this question depends on what is expected of second- and fourth-graders, and these expectations put norm-referenced boundaries on what is taught in the third grade. Professional judgment and many years of experience combine to define a reasonable domain of content and a reasonable level of performance. A test is then constructed to represent this content at this level.

Given a test that is designed to represent a particular domain of content, the scores from that test may be interpreted strictly with respect to that content, or they may be interpreted in a normative framework by comparing one person's performance with that of others. Domain-referenced interpretation means that the degree of achievement is assessed relative to the test itself and the instructional objectives that gave rise to the test. The evaluation may result in a dichotomous judgment that the person has mastered the material and is ready for further instruction, for certification or licensure, or for whatever decision is the object of the measurement. Or, the evaluation may result in a judgment of degree of mastery. The latter approximates what teachers do when they assign grades, while the former is similar to a pass/fail decision or a decision to begin new material. Many uses of tests constructed under the mandate of the No Child Left Behind Act (see Chapter 7) involve pass/fail decisions regarding mastery of the content expected of students at the completion of various points of their education, but the content definition is quite broad.

For the group of tests that are typically called criterion referenced, the standard, then, is provided by the definition of the specific objectives that the test is designed to measure. When the type of decision to be made is a mastery decision, this description of the content, together with the level of performance that the teacher, school, or school system has agreed on as representing an acceptable level of mastery of that objective, provides an absolute standard. Thus, the illustrative domain-referenced test of capitalization of proper nouns in Table 3–2 is presumed to provide a representative sample of tasks calling for this specific competence. If we accept the sample of tasks as representative and if we agree that 80% accuracy in performing this task is the minimum acceptable performance, then a score of 10 out of 13 words correctly underlined defines the standard in an absolute sense.

Even the dichotomous or mastery judgment is made in a sociopolitical, hence normative, context. The teacher or school has to decide what constitutes mastery, and there are some not-so-subtle social pressures that affect such decisions. Most teachers define the level of achievement necessary for mastery in such a way that an "appropriate" minimum number of students are identified as masters. In practice, this means that over a period of time the teacher develops a fairly accurate idea of how typical students will perform on his or her tests covering a course of instruction. The tests, grading practices, or passing standards are adjusted so that, in the long run, the right number of students pass, which makes the setting of passing standards basically a normative decision! (See Shepard, 1984, for a discussion of setting standards in criterion-referenced testing and Jaeger, 1989 or Cizek and Bunch, 2006, for a discussion of standard-setting methods

generally. A major area of controversy in education today is where the standards should be set for testing competence under No Child Left Behind. Many states are finding that unacceptably large numbers of students do not meet the standards set in their legislation.)

In the usual classroom test used for summative evaluation, such a standard operates indirectly and imperfectly, partly through the teacher's choice of tasks to make up the test and partly through his or her standards for evaluating the responses. Thus, to make up their tests, teachers pick tasks that they consider appropriate to represent the learnings of their students. No conscientious teacher would give spelling Test A in Table 3–1 to an ordinary high school group or Test B to third-graders. When the responses vary in quality, as in essay examinations, teachers set standards for grading that correspond to what they consider is reasonable to expect from students like theirs. We would expect quite different answers to the question "What were the causes of the War of 1812?" from a ninth-grader and from a college history major.

However, the inner standard of the individual teacher tends to be subjective and unstable. Furthermore, it provides no basis for comparing different classes or different areas of ability. Such a yardstick can give no answers to such questions as "Are the children in School A better in reading than those in School B?" "Is Jennifer better in reading than in mathematics?" "Is Jacob doing as well in algebra as most ninth-graders?" We need some broader, more uniform objective and stable standard of reference if we are to be able to interpret those psychological and educational measurements that undertake to appraise some trait or to survey competence in some broad area of the school curriculum. Most of this chapter is devoted to describing and evaluating several normative reference frames that have been used to give a standard meaning to test scores.

NORM-REFERENCED EVALUATION

The most commonly used frame of reference for interpreting test performance is based not on a somewhat arbitrary standard defined by a particular selection of content and interpreted as representing mastery of that content domain, but rather on the performance of other people. This represents a norm-referenced interpretation. Thus, the scores of Charlotta Cowen (47) and Gail Galaraga (71) on the 80-item spelling test from Table 2–1 can be viewed in relation to the performance of a large reference group of typical sixth-graders or of students in different school grades. Their performance is viewed not in terms of mastery versus nonmastery or in terms of relative mastery of the subject matter, but instead as above average, average, or below average compared to the reference group; we need ways to refine that scale of relative performance so that all positions on the trait can be expressed in quantitative terms.

In seeking a scale to represent the amount of the trait a person possesses, we would like to report results in units that have the following properties:

1. Uniform meaning from test to test, so that a basis of comparison is provided through which we can compare different tests—for example, different reading tests, a reading test with an arithmetic test, or an achievement test with a scholastic aptitude test.
2. Units of uniform size, so that a change of 10 points on one part of the scale signifies the same thing as a change of 10 points on any other part of the scale.
3. A true-zero point of *just none of* the quality in question, so that we can legitimately think of scores as representing *twice as much as* or *two thirds as much as*.

The different types of norm-referenced scales that have been developed for tests represent marked progress toward the first two of these objectives and thus satisfy the requirements for an in-

terval scale. The third, which is the mark of a ratio scale, can probably never be reached for the traits with which we are concerned in psychological and educational measurement. We can put five 1-lb loaves of bread on one side of a pair of scales, and they will balance the contents of one 5-lb bag of flour placed on the other side. "No weight" is *truly* "no weight," and units of weight can be added so that 2 lb is twice 1 lb. But we do not have that type of zero point or that type of adding in the case of educational and psychological measurement. If you put together two below-average students, you will not get a genius, and a pair of bad spellers cannot jointly win a spelling bee. In some cases, this deficit is the result of the particular way we have chosen to measure the trait, but for many psychological and educational traits, the deficit is a result of how we conceptualize the trait itself.

Basically, a raw point score on a test is given normative meaning only by referring it to some type of group or groups called *norm groups*. A score on the typical test is not high or low or good or bad in any absolute sense; it is higher or lower or better or worse than other scores. We can relate one person's score to a normative framework in two general ways. One way is to compare the person with a graded series of groups to see which one he or she matches. Each group in the series usually represents a particular school grade or a particular chronological age. A variant on this approach is to prepare a graded set of work samples such as samples of handwriting or responses to an essay question. Each person's product is then compared to the standard set of samples and given the score of the sample it most closely matches.

The second way to set a normative standard is to find where in a particular group the person falls in terms of the percentage of the group surpassed or in terms of position relative to the group's mean and standard deviation. These two approaches produce four main patterns for interpreting the score of an individual, which are shown schematically in Table 3–3. We next consider each in turn, evaluating its advantages and disadvantages. At the end of the chapter we examine a third way to give quantitative meaning to scores, a method based on the probability that the examinee will respond in a particular way. This method has been given the label *item response theory* or IRT.

Grade Norms

For any trait that shows a progressive and relatively uniform increase from one school grade to the next, we can prepare a set of **grade norms** or **grade equivalents.** The norm for any grade, in this sense, is the average score obtained by individuals in that grade. Because school participation and the related cognitive growth are both more or less continuous, grade norms typically are expressed with one decimal place. The whole number gives the grade, and the

Table 3–3
Main Types of Norms for Educational and Psychological Tests

Type of Norm	Type of Comparison	Type of Group
Grade norms	Individual matched to group whose performance he or she equals	Successive grade groups
Age norms	Same as above	Successive age groups
Percentile norms	Percentage of group surpassed by individual	Single age or grade group to which individual belongs
Standard score norms	Number of standard deviations individual falls above or below average of group	Same as above

decimal is the month within the grade. Thus, a grade equivalent of 5.4 is read as performance corresponding to that of the average child in the fourth month of the fifth-grade.

In simplest outline, the process of establishing grade norms involves giving the test to a representative sample of pupils in each of a number of consecutive grades, calculating the average score at each level, and then establishing grade equivalents for the in-between scores. Thus, a reading comprehension test, such as that from the Iowa Tests of Basic Skills (ITBS)–Form J, Level 9, might be given in November to pupils in Grades 2, 3, 4, and 5, with the following results:

Grade Level	Average Raw Score
2.3	13
3.3	22
4.3	31
5.3	37

The testing establishes grade equivalents for raw scores of 13, 22, 31, and 37. However, grade equivalents are also needed for the in-between scores. These are usually determined arithmetically by interpolation, although sometimes intermediate points may be established by actually testing at other times during the school year. After interpolation, we have the following table*:

Raw Score	Grade Equivalent	Raw Score	Grade Equivalent
10	1.9	24	3.5
11	2.0	25	3.6
12	2.2	26	3.7
13	2.3	27	3.8
14	2.5	28	3.9
15	2.6	29	4.0
16	2.8	30	4.1
17	2.9	31	4.3
18	3.0	32	4.4
19	3.1	33	4.5
20	3.2	34	4.7
21	3.2	35	4.9
22	3.3	36	5.1
23	3.4	37	5.3

*Note: The most recent forms of this test series calculate Developmental Standard Scores (see following section) first and derive grade equivalent scores from these.

Because raw scores on this particular test can range from 0 to 49, some way is needed to establish grade equivalents for the more extreme scores. Establishing such grade equivalents is often done by equating scores on the level of the test on which we are working with scores from lower and higher levels of the same test series, forms that have been given to earlier and later grades. In this way, grade equivalents can be extended down as low as the first month of kindergarten (denoted K.1) and up as high as the end of the first year in college (denoted 13.9), and a complete table to translate raw scores to grade equivalents can be prepared. (The reading test of this particular edition of the ITBS actually is a multilevel test that uses six overlapping sets of passages and items in a single booklet. In this way, some of the same items are used for three different levels of the test, and the projection of grade equivalents is simplified and made more accurate.)

If Jennifer got a raw score of 28 on this test, it would give her a grade equivalent of 3.9, and this score could be translated as "performing as well on this test as the average child who has completed 9 months of third grade." Such an interpretation has the advantage of connecting the test score to familiar milestones of educational development. However, this seductively simple interpretation of a child's performance has a number of drawbacks as well.

A first major question about grade norms is whether we can think of them as providing precisely or even approximately equal units. In what sense is the growth in ability in paragraph reading from grade 3.2 to 4.2 equal to the growth from grade 6.2 to 7.2? Grounds for assuming equality are clearly tenuous. When the skill is one that has been taught throughout the school years, there may be some reason to expect a year's learning at one level to be about equal to a year's learning at some other. And there is evidence that during elementary school (and possibly middle school or junior high), grade-equivalent units are near enough to equal to be serviceable. However, even in this range and for areas where instruction has been continuous, the equality is only approximate. If, on the other hand, we are concerned with a subject like Spanish, in which instruction typically does not begin until secondary school, or in something like biology, for which instruction is concentrated in a single grade, grade equivalents become completely meaningless. In addition, instruction in many skills, such as the basic skills in reading and in arithmetic computation, tapers off and largely stops by high school, so grade units have little or no meaning at this level. For this reason many achievement batteries show a grade equivalent of 10.0+ or 11.0+ as representing the whole upper range of scores. When grade equivalents such as 12.5 are reported, these do not really represent the average performance of students tested in the middle of the 12th grade, but rather they are an artificial and fictitious extrapolation of the score scale, used to provide some converted score to be reported for the most capable eighth- and ninth-graders.

A further note of caution must be introduced with respect to the interpretation of grade norms. Consider a bright and educationally advanced child in the third grade. Suppose we find that on a standardized mathematics test this child gets a score with the grade equivalent of 5.9. This score does *not* mean that this child has a mastery of the mathematics taught in the fifth grade. The *score* is as high as that earned by the average child at the end of fifth grade, but this higher score almost certainly has been obtained in large part by superior mastery of third-grade work. The average child falls well short of a perfect score on the topics that have been taught at his or her own grade level. The able child can get a number of additional points (and consequently a higher grade equivalent) merely by complete mastery of this "at-grade" material. *This warning is worth remembering.* The fact that a third-grade child has a grade equivalent of 5.9 does not mean that the child is ready to move ahead into sixth-grade work. The grade equivalent is only the reflection of a score and does not tell in what way that score was obtained.

Reference to the content of the questions the child answered correctly would be needed to reach a judgment that the child had sufficient mastery of fifth-grade material to be able to move into the sixth grade. For this reason, grade equivalents should not be used to make mastery decisions.

Finally, there is reason to question the comparability of grade equivalents from one school subject to another. Does being a year ahead (or behind) one's grade level in language usage represent the same amount of advancement (or retardation) as the same deviation in arithmetic concepts? A good deal of evidence exists, which we consider later in this chapter, that it does not. Growth in different school subjects proceeds at different rates, depending on in-school emphasis and out-of-school learning. For this reason, the glib comparison of a pupil's grade equivalent in different school subjects can result in quite misleading conclusions.

To summarize, grade norms, which relate the performance of an individual to that of the average child at each grade level, are useful primarily in providing a framework for interpreting the academic accomplishment of children in elementary school. For this purpose, they are relatively convenient and popular, even though we cannot place great confidence in the equality of grade units or their exact equivalence from one subject to another.

Grade norms are relatively easy to determine because they are based on the administrative groups already established in the school organization. In the directly academic areas of achievement, the concept of grade level is perhaps more meaningful than is age level, for it is in relation to grade placement that a child's performance is likely to be interpreted and acted on. Outside the school setting, grade norms have little meaning.

Developmental Standard Scores

We have noted several problems with grade equivalents as normative representations of a child's performance, particularly that there is an implicit assumption that the amount of growth in the ability being tested is equal from one year to the next. Because this assumption clearly is violated for many abilities, test publishers have developed a type of score scale that is anchored to school grades but provides a better approximation to an equal interval scale, the *Developmental Standard Score Scale*.

Developmental standard scores (DSSs or SSs) are based on normalized score distributions within each grade (see the discussion of normalizing transformations later in this chapter). Scale values for two grades are chosen arbitrarily to define the scale metric, and the within-grade means and standard deviations are then used to locate other grade equivalents on this scale. For example, the Iowa Tests of Basic Skills authors have chosen to fix a scale value of 200 as equivalent to the median performance of fourth-graders and a value of 250 for eighth-graders tested in the spring. The relationship between grade equivalents and DSSs reported in the test manual is as follows:

Grade	K	1	2	3	4	5	6	7	8	9	10	11	12
DSS	130	150	168	185	200	214	227	239	250	260	268	275	280

One fact is quite clear from comparing grade equivalents and DSSs: Equal changes in grade equivalents do not correspond to equal changes in DSS. The DSS scale is constructed to have equal intervals (a 10-unit change has the same meaning everywhere on the scale). The comparison shows that there is a bigger change from year to year during the early years of school than there is in later years, 18 points from first to second grade, 10 points from eighth to ninth.

The main drawback of DSSs is that, unlike grade equivalents, they have no inherent meaning. The values chosen for the anchor points are quite arbitrary. Meaning is given only by their relationship to the grade-equivalent scale. It would be appropriate, for example, to say that a student who received a DSS of 255 was performing at the level of students in about December of their ninth-grade year. Because of their complexity and lack of obvious meaning, developmental standard scores are hard to interpret correctly and should be used with caution, even though they are reported by many test publishers. Test publishers who provide DSSs always offer normative information in other formats as well. For example, the process for reporting norm-referenced scores for the ITBS determines DSSs first because of their interval-scale properties and then provides tables for converting the DSSs to the other types of normative scores described in this chapter (with the exception of age norms).

Age Norms

If a trait is one that may be expected to show continuous and relatively uniform growth with age, it may be appropriate to convert the raw score into an **age score,** or **age equivalent,** as a type of common score scale. During childhood we can observe continuous growth in height and weight, in various indices of anatomical maturity, and in a wide range of perceptual, motor, and cognitive performances. It makes a crude type of sense to describe an 8-year-old as being as tall as the average 10-year-old and having the strength of grip of the average 9-year-old, as well as the speaking vocabulary of the average 6-year-old. In the early development of intelligence and aptitude tests, raw scores were typically converted into age equivalents, and the term *mental age* was added to the vocabulary of the mental tester and the general public alike, with occasionally unfortunate consequences.

An age equivalent is, of course, the average score earned by individuals of a given age and is obtained by testing representative samples of 8-year-olds, 9-year-olds, 10-year-olds, and so forth. In this respect, it parallels the grade equivalent described earlier. And, as in the case of grade equivalents, a major issue is whether we can reasonably think of a year's growth as representing a standard and uniform unit. Is growth from age 5 to age 6 equal to growth from age 10 to age 11? And is growth in any one year equivalent to growth in any other year on our scale? As we move up the age scale, we soon reach a point where we see that the year's growth unit is clearly not appropriate. There comes a point, some time in the teens or early 20s, when growth in almost any trait that we can measure slows down and finally stops. In Figure 3–1, which illustrates the normal growth of mean height for girls, the slowdown takes place quite abruptly around age 14. A year's change in height after age 14 seems clearly to be much less than a year's change earlier on the scale. At about age 14 or 15, the concept of height-age ceases to have any meaning. The same problem of a flattening growth curve is found, varying only in the age at which it occurs—and in abruptness, for any trait that we can measure. (Of course, individuals mature at different rates, so a given individual might stop growing at an earlier age or continue growing until a later age. This illustrates the problem of using the mean, or any other measure of central tendency to represent the score of an individual.)

The problem introduced by the flattening growth curve is most apparent when we consider the individual who falls far above average. What age equivalent shall we assign to a girl who is 5 ft 10 in. (70 in.) tall? The average woman *never* gets that tall at any age. If we are to assign any age value, we must invent some hypothetical extension of our growth curve, such as the dashed line in Figure 3–1. This line assumes that growth after age 14 continues at about the same rate that was typical up to age 14. On this extrapolated curve, the height of 5 ft 10 in. would be assigned a

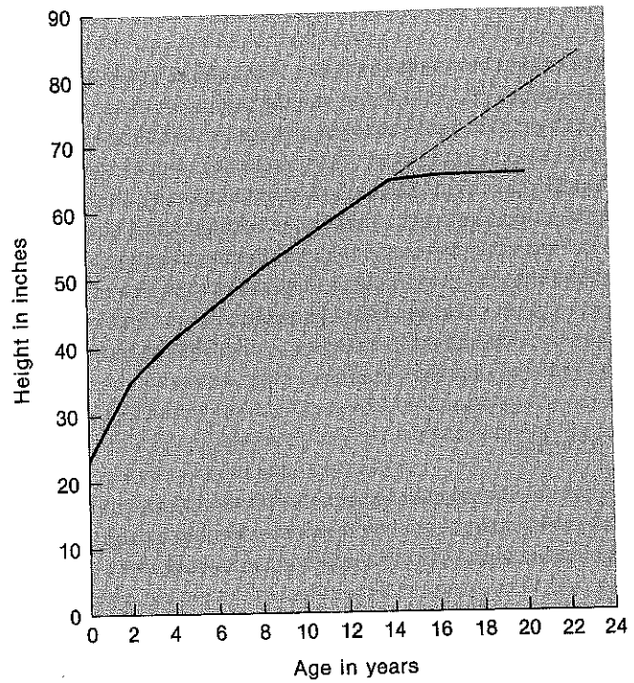


Figure 3–1
Girls' age norms for height.

height-age of about 16 years and 6 months. But this is a completely artificial and arbitrary age equivalent. It does *not* correspond to the average height of $16\frac{1}{2}$-year olds. It does not correspond to the average height at *any* age. It merely signifies "taller than average." Unfortunately, there is no cue to be gotten from these extrapolated age equivalents that suggests their arbitrary nature. The problem is even more severe here than it is with extrapolated grade equivalents.

Age norms, which are based on the characteristics of the average person at each age level, provide a readily comprehended framework for interpreting the status of a particular individual. However, the equality of age units is open to serious question, and as one goes up to adolescence and adulthood, age ceases to have any meaning as a unit in which to express level of performance. Age norms are most appropriate for infancy and childhood and for characteristics that grow as a part of the general development of the individual, such as height, weight, or dentition. General mental development, such as the cognitive characteristics embodied in the concept of mental age, shows a sufficiently universal pattern to be a useful normative indicator of status, but, in general, age norms should not be used for cognitive characteristics beyond the elementary school years, because the patterns of growth of these functions depend too heavily on formal school experiences or have not been found to show the pattern of growth necessary for age norms to be appropriate.

Percentile Norms

We have just seen that in the case of age and grade norms, meaning is given to the individual's score by determining the age or grade group in which the person would be exactly average. But often such a comparison group is inappropriate or some other group would be more useful. For example, we are frequently concerned with the performance of people who are no longer in the

elementary grades where grade norms have meaning. Or, we may be interested in personality or attitude characteristics for which age or grade norms are wholly unusable. Or, the type of information that we seek may require that we specify the group of interest more narrowly than is practical for age or grade norms. For example, we may be interested in people who are all the same age or are all in the same grade.

Each individual belongs to many different groups. An individual who is 18 years old belongs to some of the following groups, but not to others: all 18-year-olds, 18-year-olds in the 12th grade, 18-year-olds applying to college, 18-year-olds not applying to college, 18-year-olds applying to Ivy League colleges, 18-year-olds attending public (or parochial) schools, and 18-year-olds attending school in California. For some purposes it is desirable or necessary to define the comparison group more narrowly than is possible with grade or age norms. One universally applicable system of norms is the percentile norm system.

The typical percentile norm, or **percentile rank,** uses the same information that we used to compute percentiles in Chapter 2, but the procedure is slightly different. *Percentile ranks are calculated to correspond to obtainable score values.* If a test has 10 items, it can yield 11 different raw scores, the whole numbers from 0 to 10. There are only 11 possible values that percentile ranks could assume for this test—one for each obtainable score—but it would still be possible to calculate any number of percentiles. For example, one could compute, using the procedures described in Chapter 2, the 67.4th percentile as well as the 67th and 68th. But only the 11 obtainable scores would have corresponding percentile ranks. The normative interpretation of test scores more often uses percentile ranks than percentiles, because test results come in a limited number of whole score units.

The procedure for determining percentile ranks starts with a frequency distribution such as the one shown in Table 3–4. We assume, as we did for percentiles, that (1) the underlying trait the test measures is continuous, (2) each observable score falls at the midpoint of an interval on this continuum, and (3) the people who obtained a given raw score are spread evenly throughout the interval. Because each raw score falls at the middle of an interval, half of the people in the

Table 3–4
Determining Percentile Ranks for a 10-Item Test

Raw Score	Frequency	Cumulative Frequency	Percentile Rank
10	1	60	99
9	3	59	96
8	5	56	89
7	12	51	75
6	15	39	52
5	9	24	32
4	7	15	19
3	4	8	10
2	2	4	5
1	1	2	2
0	1	1	1

interval are considered to be below the midpoint and half above. Even if only one person falls into a particular interval, we assume that half of that person falls above the midpoint of the interval and half falls below.

To find the percentile rank of a raw score, we count the number of people who are below that score and divide by the total number of people. The number of people below a raw score value includes all of the people who obtained lower scores plus half of the people who received the score in question (the latter group because they are assumed to be in the bottom half of the interval and, therefore, below the raw score). For example, to calculate the percentile rank of a raw score of 4 in Table 3–4, we would take the eight people who got scores below 4 and half of the seven people at 4. The result is $(8 + 3.5)/60 = 11.5/60 = 0.1917$. In reporting percentile ranks it is conventional to round the answer to two decimal places and multiply by 100 to remove the decimal point except at the extremes of the scale. The percentile rank that corresponds to a raw score of 4 is therefore 19.

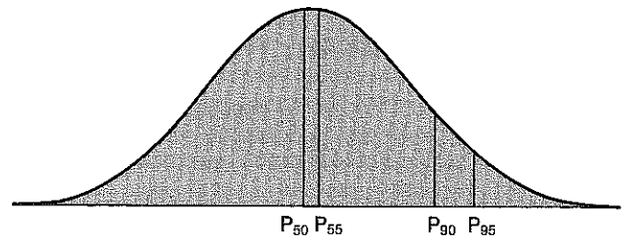
The major procedural difference between calculating percentiles, such as the median, and percentile ranks, such as those in Table 3–4, is where one starts. To calculate **percentiles,** we *specify a percent of interest,* such as the 25th or 60th, and determine the answer, a point on the continuous score scale, by the procedures described in Chapter 2. The values that correspond to these percentages need not be, and seldom are, whole points of score. When calculating **percentile ranks,** we start with *a point on the score scale,* an obtainable score value, and find as the answer the percentage of the group that falls below the chosen score.

Percentile ranks are very widely adaptable and applicable. They can be used wherever an appropriate normative group can be obtained to serve as a yardstick. They are appropriate for young and old and for educational, counseling, or industrial situations. To surpass 90% of a reference comparison group signifies a comparable degree of excellence whether the function being measured is how rapidly one can solve simultaneous equations or how far one can spit. Percentile ranks are widely used and their meaning is readily understood. Were it not for the two points we next consider, they would provide a framework very nearly ideal for interpreting test scores.

The first issue that faces us in the case of percentile ranks is specifying the norming group. On what type of group should the norms be based? Clearly, we will need different norm groups for different ages and grades in the population. A 9-year-old must be evaluated in terms of 9-year-old norms; a sixth-grader, in terms of sixth-grade norms; an applicant for a job as real estate agent, in terms of norms for real estate agent applicants. The appropriate norm group is in every case the relevant group to which the individual belongs and in terms of which his or her status is to be evaluated. It makes no sense, for example, to evaluate the performance of medical school applicants on a biology test by comparing their scores with norms based on high school seniors. If the test is to be used by a medical school, the user must find or develop norms for medical school applicants.

Hence, if percentile ranks are to be used, multiple sets of norms are usually needed. There must be norms appropriate for each distinct type of group or situation in which the test is to be used. This requirement is recognized by the better test publishers, and they provide norms not only for age and grade groups but also for special types of educational or occupational populations. However, there are limits to the number of distinct populations for which a test publisher can produce norms, so published percentile norms will often need to be supplemented by the test user, who can build norm groups particularly suited to local needs. Thus, a given school system will often find it valuable to develop local percentile norms for its own pupils. (Most test publishers will assist school districts with the development of local norms.) Such norms will permit scores for individual pupils to be interpreted in relation to the local group, a comparison that may be more significant for local decisions than is comparison with national, regional, or state

Figure 3–2
Normal curve, showing selected percentile points.



norms. Likewise, an employer who uses a test with a particular category of job applicant may well find it useful to accumulate results over a period of time and prepare norms for this particular group of people. These strictly local norms will greatly facilitate the evaluation of new applicants. Thus, the possibility of specifying many different norm groups for different uses of a test constitutes both a problem, in the sense of greater complexity, and a strength, in that more accurate comparisons can be made.

The second percentile rank issue relates to the question of equality of units. Can we think of five percentile points as representing the same amount of the trait throughout the percentile scale? Is the difference between the 50th and 55th percentile equivalent to the difference between the 90th and 95th? To answer this question, we must notice the way in which the test scores for a group of people usually pile up. We saw one histogram of scores in Figure 2–1 of Chapter 2. This picture is fairly representative of the way the scores fall in many situations. Cases pile up around the middle score values and tail off at either end. The ideal model of this type of score distribution, the normal curve, was also considered in connection with the standard deviation in Chapter 2 (see Table 2–5 and Figure 2–6) and is shown in Figure 3–2. The exact normal curve is an idealized mathematical model, but many types of test results distribute themselves in a manner that approximates a normal curve. You will notice the piling up of most cases in the middle, the tailing off at both ends, and the generally symmetrical pattern.

In Figure 3–2, four points have been marked: the 50th, 55th, 90th, and 95th percentiles. The baseline represents a trait that has been measured in a scale with equal units. The units could be items correct on a test or inches of height. Note that near the median, 5% of the cases (the 5% lying between the 50th and 55th percentiles) fall in a tall narrow pile. Toward the tail of the distribution, 5% of cases (the 5% between the 90th and 95th percentiles) make a relatively broad low bar. In the second instance, 5% of the cases spread out over a considerably wider range of the trait than in the first. The same number of percentile points corresponds to about three times as much of the score scale when we are around the 90th–95th percentiles as when we are near the median. The farther out in the tail we go, the more extreme the situation becomes.

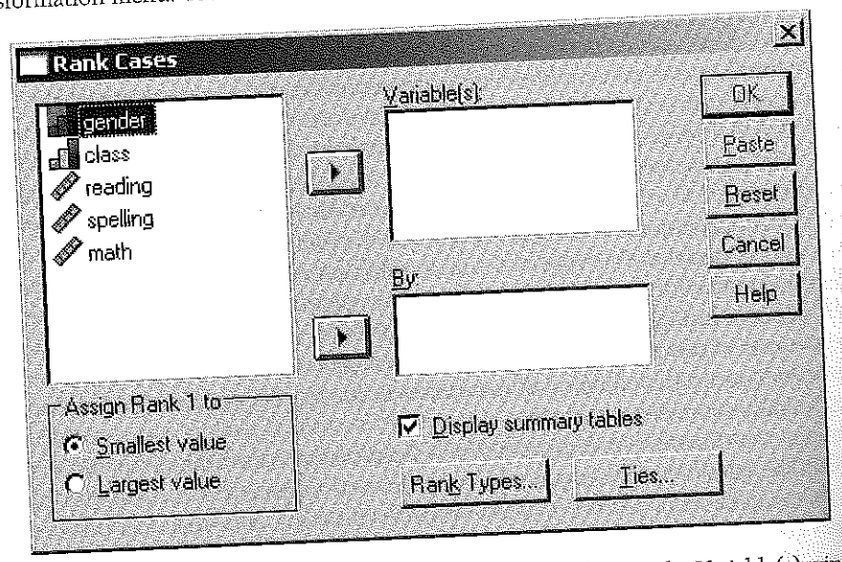
Thus, percentile units are typically and systematically unequal, relative to the raw score units. The difference between being first or second in a group of 100 is many times as great as the difference between being 50th and 51st. Equal percentile differences do not, in general, represent equal differences in amount of the trait in question. Any interpretation of percentile norms must take into account the fact that such a scale has been pulled out at both ends and squeezed in the middle. Mary, who falls at the 45th percentile in arithmetic and at the 55th in reading, shows a trivial difference in these two abilities, whereas Alice, with respective percentiles of 85 and 95, shows a larger difference—one that may be important for decision making.

MAKING THE COMPUTER DO IT

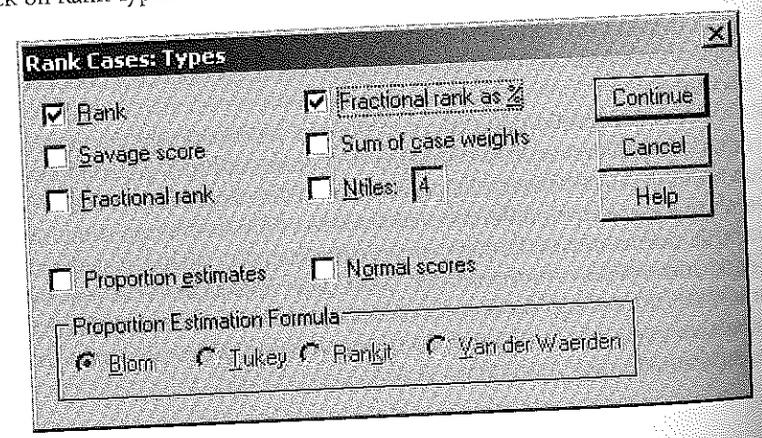
Percentile Ranks

Both SPSS and Excel claim to compute approximate percentile ranks for a set of raw scores, although the process is easier and more accurate with SPSS. Unfortunately, the programs define percentile ranks in different ways. SPSS determines the *rank* of each score in the set of data and divides the rank by the total number of cases. When there is an odd number of cases with a given rank, this introduces a small error due to the fact that the middle case is not divided in half, but the error is of little consequence. We mention it only to make you aware that you may get small differences when doing the computations by hand.

To obtain percentile ranks with SPSS, you must use the Rank Cases option on the Transformation menu. You will see a screen like this:

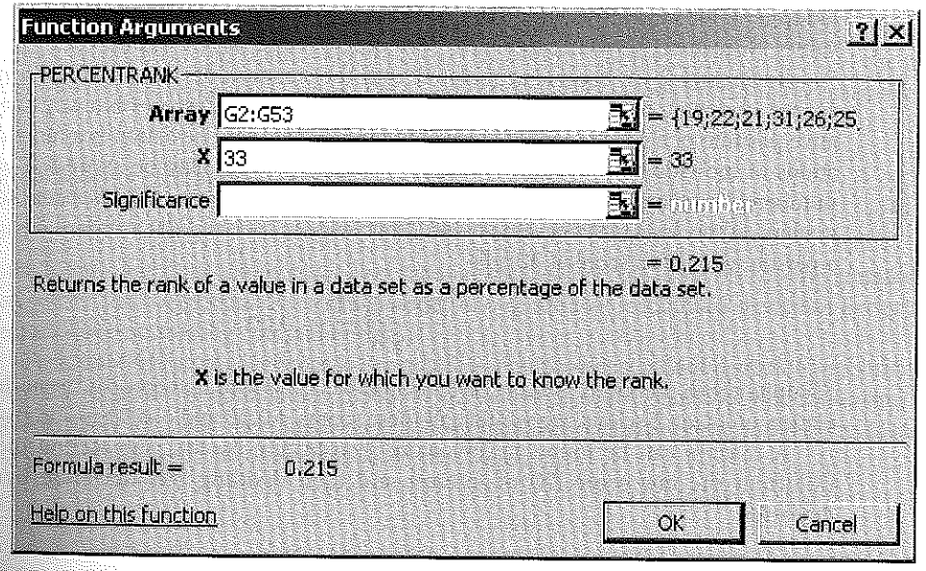


Transfer the variables for which you wish to get percentile ranks into the Variable(s) window, then click on Rank Types. You will see this screen:



Click on the box for "Fractional rank as %," click Continue, then click OK. The program will create two new variables in your data file: one giving the rank of each score, the other giving the percentile rank, but with the error noted above. The program creates two new variables for each variable analyzed. Each new variable will have the same name as the original variable, but preceded by an *r* for ranks and a *p* for percentile ranks. If your variable name has eight characters, the last character will be dropped. The new variables in the data file will have the labels "Rank of . . ." and "Fractional rank percent of . . ."

To obtain percentile ranks using Excel, you must use the "function" option (f_X or Σ). One of the functions in the statistics list is called PERCENTRANK. If you select this function, you will get a dialog box like the one shown below superimposed on your data. You must supply the portion of the data table to be included and the score value for which the percentile rank is to be computed. Here we have selected the 52 scores (rows 2–53) for the variable in column G (Math score) and requested the percentile rank of a score of 33. The program tells us that the percentile rank of this score is 0.215.



Unfortunately, the results provided by Excel are systematically in error by a substantial amount if you use the program according to the instructions. The values returned by the program are *approximately* equivalent to the ones returned by SPSS *for the score below the one you enter,* but this is not a reliable way to correct the error. For example, the correct percentile rank for a score of 33 in the data we have been using is 25.96 (13.5/52). The value returned by SPSS is 26.92(14/52). As we have seen, Excel produces a value of 0.215 for a percentile rank of 21.5. This is approximately the correct value for the score 32. Because Excel produces incorrect results that cannot easily be corrected, we cannot recommend using it to compute percentile ranks.

The fact that units on the percentile scale are systematically uneven means that this is an ordinal scale. Larger numbers mean more of the trait, but equal differences in percentile rank do not mean equal differences in the trait. Percentile ranks do not solve the equality of units problem that we encountered with age and grade equivalents.

One of the consequences of this inequality of units in the percentile scale is that percentiles cannot properly be treated with many of the procedures of mathematics. For example, we cannot add two percentile ranks together and get a meaningful result. The sum or average of the percentile ranks of two raw scores will not yield the same result as determining the percentile rank of the sum or average of the two raw scores directly. A separate table of percentile equivalents would be needed for every combination of raw scores that we might wish to use. Again, the better test publishers provide percentile rank conversion tables for all of the combinations of subtest scores that they recommend, as well as for the subtests themselves.

Standard Score Norms

Because the units of a score system based on percentile ranks are so clearly unequal, we are led to look for some other unit that does have the same meaning throughout its whole range of values. **Standard score scales** have been developed to serve this purpose.

In Chapter 2 we became acquainted with the standard deviation (*SD*) as a measure of the spread, or scatter, of a group of scores and standard scores or Z-scores as a way to express the relative position of a single score in a distribution. The standard deviation is a function of the deviations of individual scores away from the mean. Any score may be expressed in terms of the number of standard deviations it is away from the mean. The mean *mathematics* score for ninth-graders on the Tests of Achievement and Proficiency is 24.1 and the standard deviation is 9.8, so a person who gets a score of 30 falls

$$\frac{30 - 24.1}{9.8} = 0.60$$

SD units above the mean. A score of 15 would be 0.93 *SD* units *below* the mean. In standard deviation units, or Z-scores, we would call these scores +0.60 and −0.93, respectively.

A Z-score can be found in any score distribution by first subtracting the group mean (M) from the raw score (X) of interest and then dividing this deviation by the standard deviation:

$$Z = \frac{X - M}{SD}$$

If this is done for every score in the original distribution, the new distribution of Z-scores will have a mean of zero, and the standard deviation of the new distribution will be 1.0. About half of the Z-scores will be negative, indicating that the people with these scores fell below the mean, and about half will be positive. Most of the Z-scores (about 99%) will fall between −3.0 and +3.0.

Suppose we have given the Tests of Achievement and Proficiency—Form G during the fall to the pupils in a ninth-grade class, and two pupils have the following scores on mathematics and reading comprehension:

Pupil	Mathematics	Reading Comprehension
Hector	30	48
Jose	37	42

Let us see how we can use standard scores to compare performance of an individual on two tests or the performance of the two individuals on a single test.

The mean and standard deviation for the mathematics and reading comprehension tests are as follows:

	Mathematics	Reading Comprehension
Mean	22.7	33.8
SD	9.4	11.1

On mathematics, Hector is 7.3 points above the mean. His Z-score is 7.3/9.4 = +0.78. On reading comprehension, he is 14.2 points above the mean, or $Z = 14.2/11.1 = +1.28$. Hector is about one-half of a standard deviation better in reading comprehension relative to the norm group than in mathematics. For Jose, the corresponding calculations for mathematics give

$$(37 - 22.7)/9.4 = +1.52$$

and for reading comprehension give

$$(42 - 33.8)/11.1 = +0.74$$

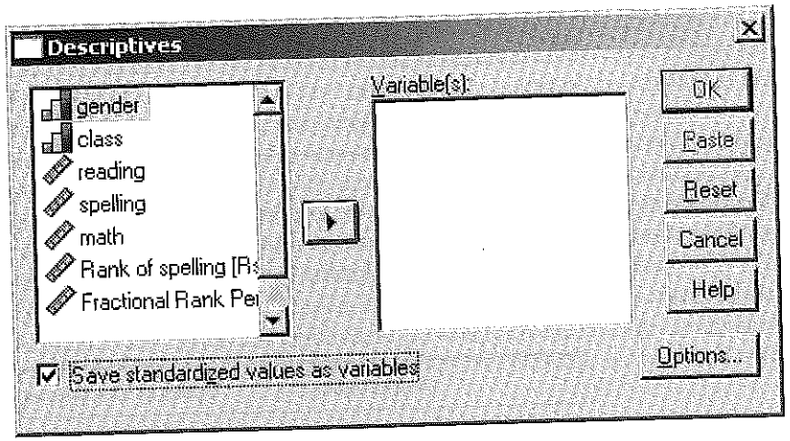
Thus, Hector has done about as well on mathematics as Jose has done on reading comprehension, while Jose's mathematics score is about one-quarter of a standard deviation better than Hector's score on reading comprehension.

Each pupil's level of excellence is expressed as a number of standard deviation units above or below the mean of the comparison group. The Z-scores provide a standard unit of measure having essentially the same meaning from one test to another. For aid in interpreting the degree of excellence represented by a standard score, see Table 2--5.

MAKING THE COMPUTER DO IT

Standard Scores

You can use either SPSS or Excel to compute Z-scores, but again the process is much easier using SPSS because the program will compute a new standard score variable for each person in the distribution and save the new variable in your data file. To create a new standard score variable with SPSS, simply select the Descriptives option from the Analysis menu. The following dialog box will appear:



Select the variables for which you wish standard scores, then click on the "Save standardized values as variables" box. A new variable with the original variable name prefixed by Z will be created in your data file. The only problem with this program is that it uses the population estimate of the standard deviation in computing the Z-scores rather than the sample value. This results in Z-scores that are slightly too small, but the error is consistent across individuals and variables and for reasonably large groups is not large enough to worry about.

Excel requires that you first compute the mean and standard deviation of the distribution. You must then use the function option (f_X or Σ) and select Standardize from the statistics list. You will then be prompted to enter a score, the mean, and the standard deviation. The Z-score for the selected score will be displayed. The advantage of Excel is that you can use the proper standard deviation, while the disadvantage is that you must compute each Z-score separately. We will describe an alternative way to obtain Z-scores with Excel in the next box.

Converted Standard Scores

All in all, Z-scores are quite satisfactory except for two matters of convenience: (1) They require use of plus and minus signs, which may be miscopied or overlooked, and (2) they get us involved with decimal points, which may be misplaced. Also, people do not generally like to think of themselves as negative or fractional quantities. We can get rid of the need to use decimal points by multiplying every Z-score by some convenient constant, such as 10, and we can get rid of minus signs by adding a convenient constant amount, such as 50. Then, for Hector's scores on the mathematics and reading comprehension tests, we would have

	Mathematics	Reading Comprehension
Mean of distribution of scores	22.7	33.8
SD of distribution	9.4	11.1
Hector's raw score	30	48
Hector's *Z*-score	+0.78	+1.28
Z-score × 10	8	13
Plus a constant amount (50)	58	63

(The convention is to round such converted scores to the nearest whole number, consistent with the objective of making them easy to use.) Because we have converted Hector's scores on the two tests to a common scale ($M = 50$, $SD = 10$), we can compare them directly and see that Hector is somewhat better in reading comprehension than he is in mathematics. However, as we discuss in more detail later, this comparison requires the tests to have been normed on comparable groups.

Converted standard scores are based on a simple equation that changes the size of the units and the location of the mean. In symbolic form, the equation for the above transformation is

$$C = 10(Z) + 50$$

where Z is the standard score defined earlier and C is the converted standard score. The general formula is

$$C = SD_A(Z) + M_A$$

where SD_A and M_A are any arbitrary standard deviation and any arbitrary mean, respectively, selected for convenience.

The use of 50 and 10 for the mean and the standard deviation, respectively, is an arbitrary decision. We could have used values other than 50 and 10 in setting up the conversion into convenient standard scores. The army has used a standard score scale with a mean of 100 and a standard deviation of 20 for reporting its test results. The College Entrance Examination Board has long used a scale with a mean of 500 and a standard deviation of 100 for reporting scores on the SAT, the Graduate Record Examination, and other tests produced under its auspices. The navy has used the 50 and 10 system; intelligence tests generally use a mean of 100 and a standard deviation of 15.

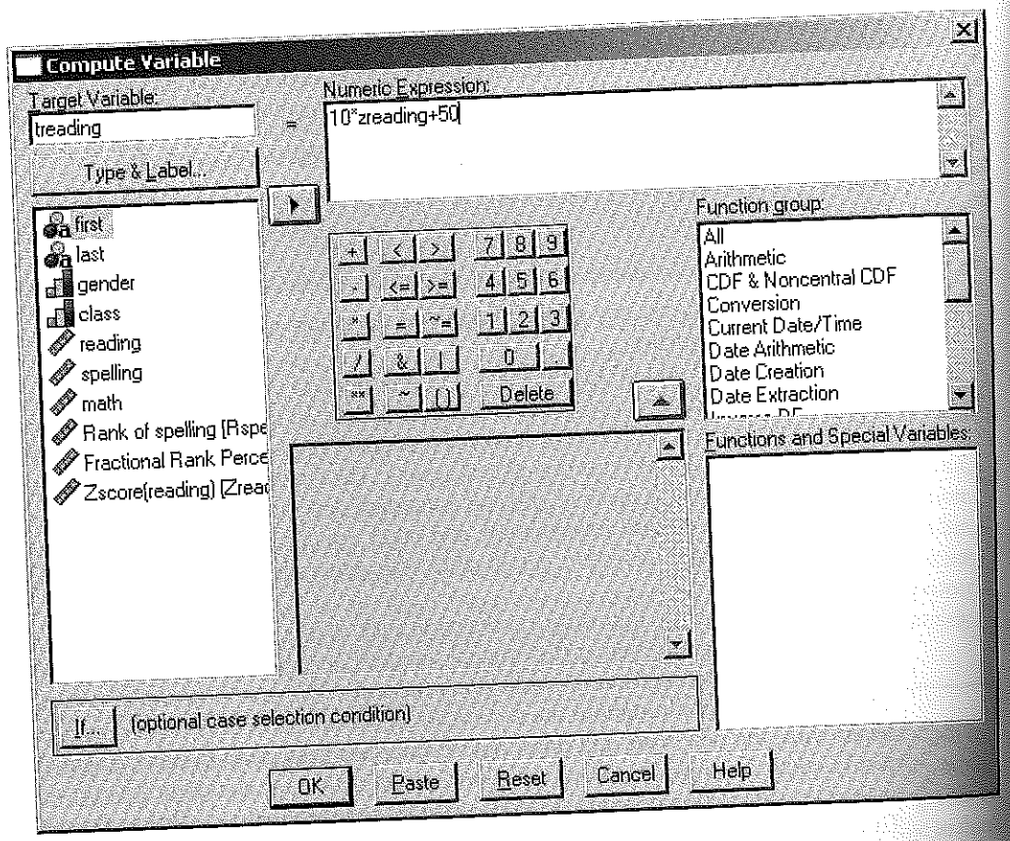
The scale of scores following a conversion such as this one is stretched out or squeezed together (depending on whether the original standard deviation is smaller or larger than the new one), but the stretching is uniform all along the scale. The *size* of the units is changed, but it is changed uniformly throughout the score scale. If the raw score scale represented equal units to begin with, the new scale still does, but nothing has been done to make unequal units more nearly equal. Because the above equation is an equation for a straight line ($y = ax + b$), this type of transformation of scores is called a **linear conversion,** or a **linear transformation.** (It is necessary here to add a note on terminology. We use the symbol *Z* to stand for standard scores in their original form and *C* to stand for *any* linear transformation of a *Z*-score. The symbol *T* is often used for the special case of a linear transformation using a mean of 50 and a standard deviation of 10.)

MAKING THE COMPUTER DO IT

Linear Transformations

It is quite simple to get either SPSS or Excel to perform any linear transformation of a set of standard scores. The easiest way to do it in SPSS is to use the Compute procedure in the Transform menu. You must first create the standard score variable as described in the preceding box. Then put the name you wish to use for the transformed variable in the Target Variable box and enter the transformation equation in the box labeled Numeric Expression. Click OK, and the new variable will be added to your data file. The box below shows

how to compute a new variable called "treading" (*T*-reading), which has an *SD* of 10 and a mean of 50, from the reading test scores. Note that we first created the *Z*-score variable "zreading" (*Z*-reading).



The procedure for transforming variables in Excel is quite similar. First, you must create the standard score variable, then compute the transformed variable. The easiest way to do this is to determine the mean and standard deviation of the variable to be transformed, then write a function that will compute the *Z*-scores as described earlier. For example, the mean and *SD* of the math scores for our 52 students in Table 2–1 are 38.17 and 8.84, respectively. The following screen shows the function for computing the *Z*-score for Quadra Quickly. Her math score is in cell G2, so the function "$= (G_2 - 38.17)/8.84$" placed in cell H2 computes her *Z*-score of −2.17. Scores for the other students can be found by highlighting the remaining cells in column H, clicking on the Edit menu, the Fill command, and the Down selection. (Ctrl+D will accomplish the same result.) This places the same function in all cells of column H. Once you have computed the *Z*-scores, you can use the same procedure to obtain any other linear transformation and put the *C*-scores in column I. For example, to convert the *Z*-scores in column H into scores using the scale commonly used for IQ scores, we would insert the function "= (H2*15) + 100" in cell I2. Again

using the Fill command with the Down selection, we would get the desired transformed standard scores for all pupils in column I.

Microsoft Excel - Statistics examples.xls

H2 =(G2-38.17)/8.84

	A	B	C	D	E	F	G	H	I	J	K	L
1	First	Last	Gender	Class	Reading	Spelling	Math	Z-scores	IQ-scores			
2	Quadra	Quickly	2	1	21	44	19	-2.17	67.47			
3	Nathan	Natts	1	1	22	47	22	-1.83				
4	Wakana	Watanabe	2	1	25	53	21	-1.94				
5	Xenum	Xerxes	1	1	25	54	31	-0.81				
6	Jack	Johanson	1	1	26	56	26	-1.36				
7	Kleven	Klipsch	1	1	28	51	25	-1.19				
8	Nancy	Nowitts	2	2	28	44	44	0.66				
9	Rahim	Roberts	1	1	29	64	43	0.55				
10	Larry	Lewis	1	2	29	40	34	-0.47				
11	Velma	Vauter	2	2	29	49	36	-0.25				
12	Herpo	Henry	1	1	30	51	34	-0.47				
13	Xena	Xerxes	2	2	30	57	37	-0.13				
14	Zephina	Zoro	2	2	30	47	38	-0.02				
15	Guido	Garcia	1	2	31	52	29	-1.04				
16	Rhonda	Rostropovich	2	2	31	50	31	-0.81				
17	Aaron	Andrews	1	1	32	64	43	0.55				
18	Petula	Peters	2	1	32	64	33	-0.58				
19	Yuan	Young	1	1	32	59	24	-1.60				
20	Bellinda	Brown	2	2	33	38	41	0.32				
21	Charlotta	Cowen	2	2	33	47	50	1.34				
22	Igor	Ivanovich	1	2	33	53	43	0.55				
23	Quincy	Quirn	1	2	33	48	33	-0.58				
24	Sally	Stebbens	2	2	33	51	32	-0.70				
25	William	Westerbeke	1	2	33	54	33	-0.58				
26	Thomas	Tank	1	1	35	65	38	-0.02				
27	Kaleen	Knowles	2	2	35	55	51	1.45				

Sheet4 / Sheet1 / Sheet2 / Sheet3

Ready NUM

Normalizing Transformations

Area Normalizing Transformation

Frequently, standard score scales are developed by combining the percentile ranks corresponding to the raw scores with a linear transformation of the *Z*-scores that are associated with those percentile ranks in the normal distribution, making the assumption that the trait being measured has a normal distribution. (This is called an **area conversion** of scores. Because the complete transformation cannot be expressed by a straight line, or linear equation, it is also called a **nonlinear transformation.**) Thus, in the mathematics test, we found that the percentile rank of a score of 33 for the data in Table 2–1 was 26. In the table of the normal distribution (provided in the Appendix), the *Z*-score below which 26% of the cases fall is −0.64. If the distribution of mathematics scores were exactly normal, this is what the *Z*-score of a raw score of 33 would be. Consequently, to create an area-normalized version of this raw score, we would *assign* a standard score of −0.64 to a raw score of 33. Expressing this result on a scale in which the standard deviation is to be 10 and the mean 50, we have

$$T = 10(-0.64) + 50 = -6 + 50 = 44$$

The complete process of preparing a normalized standard score scale by the area conversion method involves finding the percentile rank for each obtainable raw score. The *Z*-score below

which the specified percentage of the normal distribution falls is then substituted for the raw score, resulting in a set of *Z*-scores that yield a normal distribution for the group on which we have obtained our data. These *Z*-scores can then be subjected to a linear transformation using whatever mean and standard deviation are desired.

Normal Curve Equivalents

A second type of normalized standard score gaining popularity in education is the scale of **normal curve equivalents,** or the NCE scale. This scale is developed using the normalizing procedures described above and the same mean that the *T* scale uses, but the standard deviation is set at 21.06 rather than at 10. The reason for choosing this particular standard deviation is that it gives a scale in which a score of 1 corresponds to a percentile rank of 1 and a score of 99 corresponds to a percentile rank of 99. The relationship between NCEs and percentile ranks (PRs) is shown in the first two columns of Table 3–5. Most major publishers of educational achievement tests provide tables of NCE scores, thus allowing for comparison of relative performance on different tests. As these publishers note, however, the tests differ in content, so a common score scale does not imply that one test could be substituted for another. Also, the norm groups may not be comparable, so unless you know that two tests were normed on the same or comparable groups, NCEs from different tests should be compared with caution.

We have now identified two ways to develop standard score scales based on an arbitrary mean and standard deviation. In one, the *linear transformation method,* *Z*-scores are *computed* from the observed mean and standard deviation and the resulting *Z*-scores may be further transformed by first being multiplied by an arbitrary new standard deviation and then added to an arbitrary new mean. This method does not change the relative distances between scores and leaves the shape of the score distribution unchanged. In the other method, the *area or normalizing transformation,* percentile ranks are used to *assign* *Z*-scores to raw scores, based on the percentage of the normal distribution that falls below the *Z*-score. These assigned *Z*-scores are then transformed with an arbitrary standard deviation and mean to a desired scale. The resulting scores will form a normal distribution, regardless of the shape of the distribution of the raw scores.

Table 3–5
Relationship Between Normal Curve Equivalents, Percentile Ranks, and Stanines

NCE	PR	Stanine	PR
99	99	9	≥ 96
90	97	8	89–95
80	92	7	77–88
70	83	6	60–76
60	65	5	40–59
50	50	4	23–39
40	32	3	11–22
30	17	2	4–10
20	8	1	3 ≥
10	3		
1	1		

Normalized standard scores make sense whenever it seems likely that the group is a complete one that has not been curtailed by systematic selection at the upper or lower ends. Furthermore, they make sense whenever it seems likely that the original raw score scale does not represent a scale of equal units but the underlying trait could reasonably be assumed to have a normal distribution. Many test makers systematically plan to include in their tests several items of medium difficulty and few easy or hard items. The effect of this practice is to produce tests that spread out and make fine discriminations among the middle 80% or 90% of test takers, while making coarser discriminations at the extremes. That is, the raw score units in the middle of the distribution correspond to smaller true increments in the ability being measured than do raw score units at the extremes. The "true" distribution of ability is pulled out into a flat-topped distribution of scores. The operation of normalizing the distribution reverses this process.

Stanines

A type of normalized standard score that has become quite popular for educational tests is the **stanine** (a condensation of the phrase *standard nine-point scale)* score. The stanine scale has a mean of 5, and stanine units each represent half of a standard deviation on the basic trait dimension. Stanines tend to play down small differences in score and to express performance in broader categories, so that attention tends to be focused on differences that are large enough to matter. The relationship between the stanine scale and the percentile rank scale is shown in the last two columns of Table 3–5. Like the area-transformed scores discussed earlier, stanine scores are assigned based on percentile information.

The relationships between a number of the different standard score scales (after normalization) and the relationship of each to percentiles and to the normal distribution are shown in Figure 3–3. This figure presents the model of the normal curve, and beneath the normal curve are a scale of percentiles and several of the common standard score scales. This figure illustrates the equivalence of scores in the different systems. Thus, a College Entrance Examination Board (CEEB) standard score of 600 would represent the same level of excellence (in relation to some common reference group) as an Army standard score (or AGCT) of 120, a Navy standard score (or *T*-score) of 60, a stanine score of 7, a percentile rank of 84, an NCE of 71, or a Wechsler IQ of 115. The particular choice of score scale is arbitrary and a matter of convenience. It is unfortunate that all testing agencies have not been able to agree on a common score unit. However, the important thing is that the same score scale and comparable norming groups be used for all tests in a given organization, so that results from different tests may be directly comparable.

Earlier, we discussed the importance of identifying an appropriate norm group, to allow interpretation of a raw score using percentile norms. The same requirement applies with equal force when we wish to express a person's characteristics within a standard score framework. The conversion from raw to standard score must be based on a relevant group of which the individual with whom we are concerned can be considered a member. It makes no more sense to determine an engineering graduate student's standard score on norm data obtained from high school physics students than it does to express the same comparison in percentiles.

In summary, standard scores, like percentile ranks, base the interpretation of the individual's score on his or her performance in relation to a particular reference group. They differ from percentile ranks in that they are expressed in units that are presumed to be equal, hence they represent an interval scale. The basic unit is the standard deviation of the reference group, and the

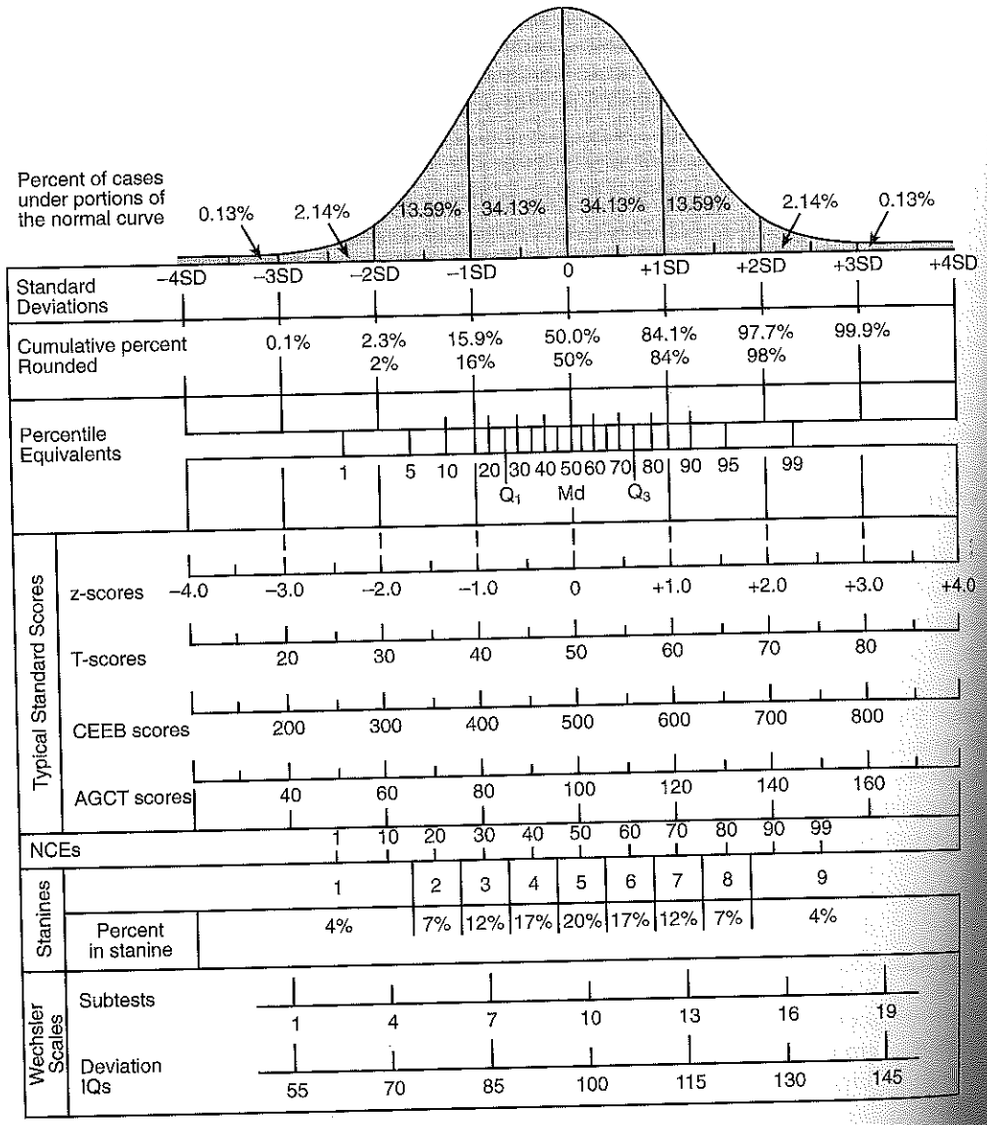


Figure 3–3
Various types of standard score scales in relation to percentiles and the normal curve.
Source: Sample items similar to those in the Differential Aptitude Tests. Copyright © 1972, 1982, 1990 by Psych Corp/Harcourt. Reproduced by permission. All rights reserved.

individual's score is expressed as a number of standard deviation units above or below the mean of the group. Standard score scales may be based on either a linear or an area (normalizing) conversion of the original scores. Different numerical standard score scales have been used by different testing agencies. Standard score scales share with percentile ranks the problem of defining an appropriate reference group.

INTERCHANGEABILITY OF DIFFERENT TYPES OF NORMS

Whichever type of normative scale is used, a table of norms will be prepared by the test publisher. This table will show the different possible raw scores on the test, together with the corresponding score equivalents in the system of norms being used. Many publishers provide tables giving more than one type of score equivalent. Table 3–6 gives an example, which shows the fall testing norms for the vocabulary test of the ITBS–Form A, Level 9 (Grade 3). Five types of norms are shown. The developmental standard scores (standard scores in this publisher's terminology) are based on a group tested early in the third grade. The NCE score scale assigns a mean of 50 and a standard deviation of 21.06 to an early third-grade group. Thus, a boy with a raw score of 21 can be characterized as follows:

1. Having a DSS of 191 (200 is the mean for fourth-graders tested in the spring)
2. Having a grade equivalent of 4.2
3. Falling at the 78th percentile in the third-grade group
4. Receiving an NCE of 66
5. Receiving a stanine of 7.

Table 3–6
Vocabulary Norms for the Iowa Tests of Basic Skills–Form A, Level 9, Grade 3 Fall Norms

Raw Score	Standard Score	Grade Equivalent	Percentile Rank	Normal Curve Equivalent	Stanine
0	121	K.2	1	1	1
1	124	K.4	1	1	1
2	128	K.6	1	1	1
3	132	K.9	2	7	1
4	136	1.1	4	13	1
5	141	1.4	6	17	2
6	147	1.7	9	22	2
7	152	1.9	13	26	3
8	157	2.1	19	32	3
9	161	2.4	24	35	4
10	164	2.6	29	38	4
11	167	2.7	34	41	4
12	170	2.9	39	44	4
13	172	3.0	43	46	5
14	174	3.1	47	48	5
15	177	3.3	54	52	5
16	179	3.5	58	54	5
17	181	3.6	61	56	6
18	183	3.7	65	58	6
19	185	3.8	68	60	6

(continued)

Table 3–6 (*Continued*)

Raw Score	Standard Score	Grade Equivalent	Percentile Rank	Normal Curve Equivalent	Stanine
20	188	4.0	73	63	6
21	191	4.2	78	66	7
22	194	4.4	82	69	7
23	197	4.6	86	73	7
24	200	4.8	89	76	8
25	204	5.1	92	80	8
26	209	5.5	95	85	8
27	216	6.0	97	90	9
28	226	6.7	99	99	9
29	240	7.9	99	99	9

From Table 3–6, it is easy to see that the different systems of norms are different ways of expressing the same thing. We can translate from one to another, moving back and forth. Thus, a child who receives an NCE of 66 in the third-grade group tested in October has a grade equivalent of 4.2. A grade equivalent of 4.0 corresponds to a percentile rank of 73 and a stanine of 6. The different systems of interpretation support one another for different purposes.

However, the different norm systems are not entirely consistent as we shift from one school subject or trait to another. This inconsistency occurs because some functions mature or change more rapidly from one year to the next, relative to the spread of scores at a given age or grade level. This can be seen most dramatically by comparing subjects like reading comprehension and mathematics. The phenomenon is illustrated by the pairs of scores shown in Table 3–7, based on the ITBS. It is assumed that the three boys were tested at the end of 5 months in the fifth grade (midyear norms). John received scores on both tests that were just average. His grade equivalent was 5.5, and he was close to the 50th percentile for

Table 3–7
Comparison of Developmental Standard Scores, Grade Equivalents, and Percentiles

	Reading Comprehension			Mathematics Computation		
Type of Score	John	Henry	Will	John	Henry	Will
DSS	210	223	235	210	223	226
Grade equivalent	5.5	6.5	7.5	5.5	6.5	6.7
Percentile rank	50	65	77	53	74	77

pupils tested after 5 months in the fifth grade. Henry shows superior performance, but how does he compare in the two subjects? From one point of view, he does equally well in both; he is just 1 full year ahead in grade equivalent. But in terms of percentiles he is better in mathematics than in reading, that is, at the 74th percentile in mathematics compared with the 65th percentile in reading. Will, on the other hand, falls at just the same percentile in both reading and mathematics. However, in his case the grade equivalent for reading is 7.5, and for mathematics, it is 6.7.

The discrepancies that appear in this example result from the differences in the variability of performance and rate of growth in reading and mathematics. Reading shows a *wide* spread within a single grade group, relative to the mean change from grade to grade. Some fifth-graders read better than the average eighth- or ninth-grader, so a reading grade equivalent of 8.0 or even 9.0 is not unheard of for fifth-graders. In fact, a grade equivalent of 9.0 corresponds to the 89th percentile for pupils at grade 5.5 in this particular test series. Ten percent of fifth-graders read as well as the average ninth-grader. By contrast, a fifth-grader almost never does as well in mathematics as an eighth- or ninth-grader—in part because the fifth-grader has not encountered or been taught many of the topics that will be presented in the sixth, seventh, and eighth grades and included in a test for those grade levels. All the basic skills that are involved in reading usually have been developed by fifth grade, so changes in reading performance result largely from greater mastery of those processes. With mathematics the case is quite different. Eighth-graders are not doing the same things better than fifth-graders do; eighth-graders are doing different things. For example, fifth-graders are likely to be working with whole numbers and relatively simple fractions, whereas eighth-graders will be studying decimals, complex fractions, and geometry. A fifth-grader might well be able to read and understand an eighth-grade history book, but very few could do eighth-grade mathematics. Thus, fifth-graders are more homogeneous with respect to mathematics than to reading skills.

The preceding point must always be kept in mind, particularly when comparing grade equivalents for different subjects. A bright child will often appear most advanced in reading and language and least so in mathematics and spelling, when the results are reported in grade equivalents. This difference may result, in whole or in part, simply from the differences in the growth functions for the subjects and need not imply a genuinely uneven pattern of progress for the child. For this reason most testing specialists are quite critical of grade equivalents and express a strong preference for percentile ranks or some type of standard score. However, because they *appear* to have a simple and direct meaning in the school context, grade equivalents continue to be popular with school personnel and are provided by most test publishers.

QUOTIENTS

In the early days of mental testing, after age norms had been used for a few years, it became apparent that there was a need to convert the age score into an index that would express rate of progress. The 8-year-old who had an age equivalent of $10\frac{1}{2}$ years was obviously better than average, but how much better? Some index was needed to take account of chronological age (actual time lived), as well as the age equivalent on the test (score level reached).

One response to the need was the expedient of dividing a person's test age equivalent by his or her chronological age to yield a quotient. This procedure was applied most extensively with tests of intelligence, where the age equivalent on the test was called a **mental age** and the

corresponding quotient was an **intelligence quotient** (IQ). In the 1920s it became common practice to multiply this fraction by 100 (to eliminate decimals), thus giving rise to the general form of the scale that is now so well known in education and psychology (see Chapter 12). However, quotients are subject to the same problems that beset the age equivalent scores from which they are computed, and when growth stops, the quotient starts to decline because chronological age continues to increase at a constant rate.

The notion of the IQ is deeply embedded in the history of psychological and educational testing and, in fact, in contemporary American language and culture. The expression *IQ test* has become part of our common speech. We are probably stuck with the term. But the way that the IQ is defined has changed. IQs have become, in almost every case, normalized standard scores with a mean of 100 and a standard deviation of 15, and we should think of them and use them in this way. These scores are sometimes referred to as *deviation intelligence quotients,* or deviation IQs, because they are basically standard scores expressed as a deviation above or below a mean of 100. The 1986 revision of the Stanford-Binet Intelligence Scale substituted the term *standard age score* for IQ to reflect more accurately the true nature of the scores, and many other tests have followed suit in dropping references to IQ.

Unfortunately, the score scale for reporting IQs does not have *exactly* the same meaning from test to test. Different tests include different tasks as measures of intelligence. Furthermore, tests may be normed at different points in time and use different sampling procedures. These differences in procedure also lead to some variation in the norms and, consequently, in the distribution of IQs they yield for any given school or community. A series of studies by Flynn (1984, 1998) also suggests that there has been a long-term rise in IQs worldwide, dating at least to the mid-1930s, which would mean that norms that are 15 to 20 years old are probably not appropriate for use today. Such a change in mean performance makes it difficult to compare results over time or between successive test forms. We discuss issues related to intelligence and tests used to measure it in Chapter 12.

PROFILES

The various types of normative frames of reference we have been considering provide a way of expressing scores from quite different tests in common units, so that the scores can be meaningfully compared. No direct way exists to compare a score of 30 words correctly spelled with a score of 20 arithmetic problems solved correctly. But, if both are expressed in terms of the grade level to which they correspond or in terms of the percentage of some defined common group that gets scores below that point, then a meaningful comparison is possible. A set of different test scores for an individual, expressed in a common unit of measure, is called a **score profile.** The separate scores may be presented for comparison in tabular form by listing the converted score values. A record showing such converted scores for several pupils is given in Figure 3–4. The comparison of different subareas of performance is made pictorially clearer by a graphic presentation of the profile. Two ways of plotting profiles are shown in Figures 3–5 and Figure 3–6.

Figures 3–4 and 3–5 show part of a class record form and an individual profile chart for the ITBS, respectively. The class record illustrates the form in which the data are reported back to the schools by the test publisher's computerized test scoring service. (The precise form that the reporting of results takes differs from one scoring service to another.) Four norm-referenced

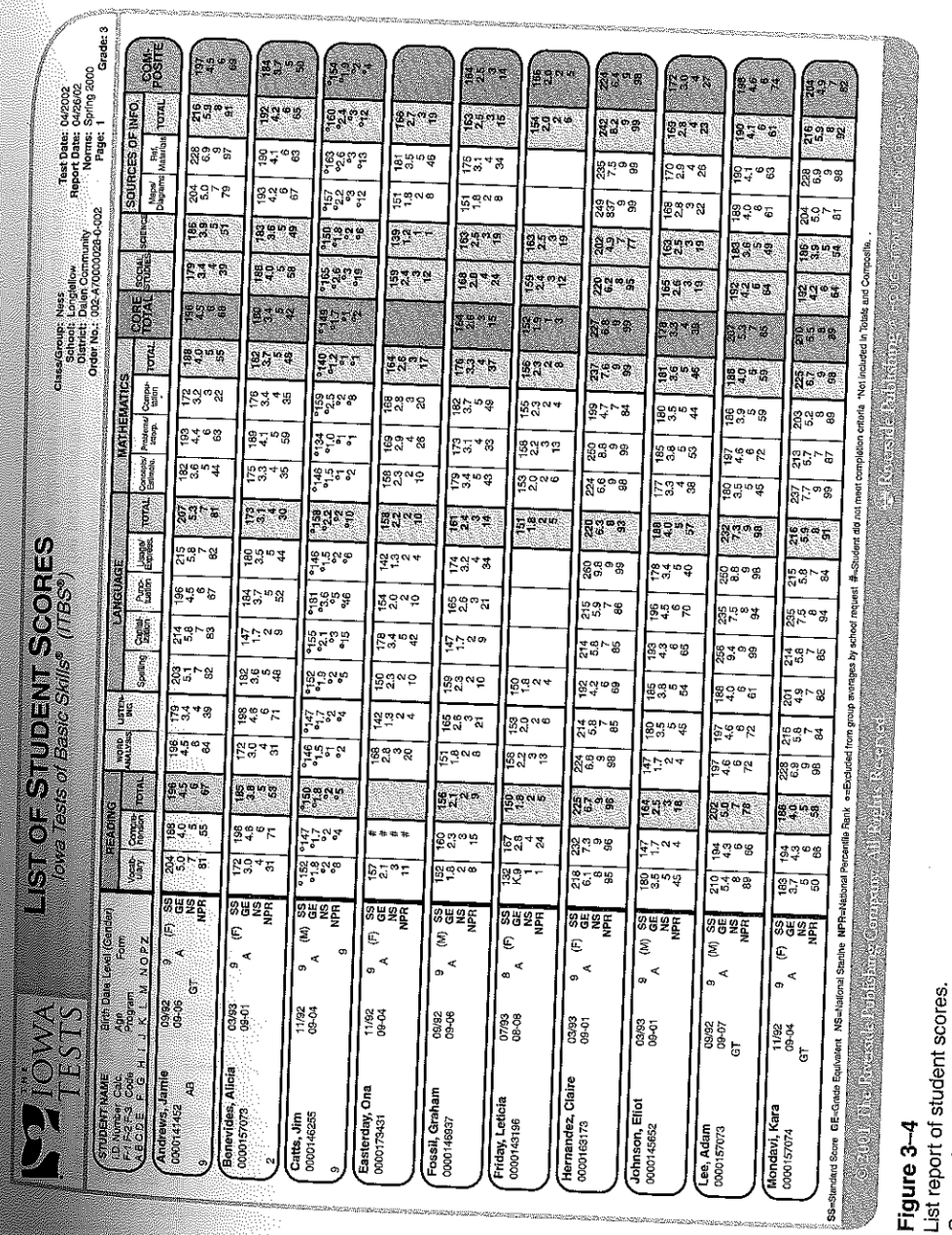


Figure 3–4
List report of student scores.

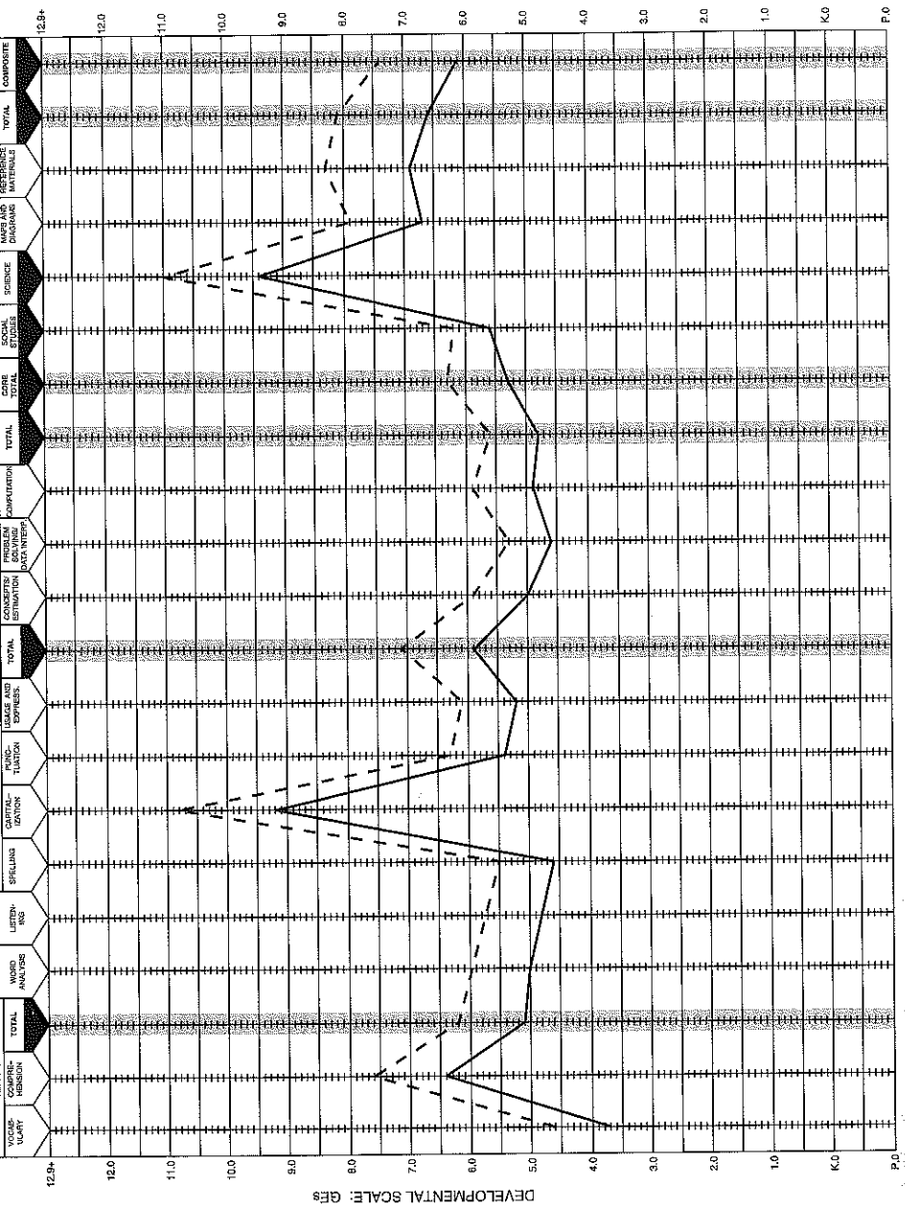


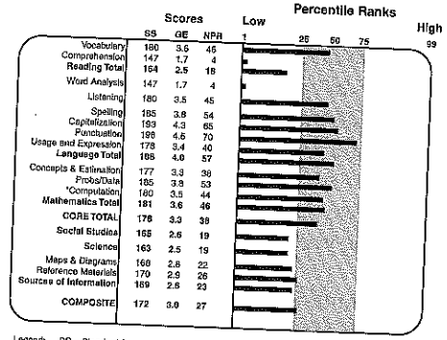
Figure 3–5
Student profile chart.
Source: Iowa Test of Basic Skills® (ITBS®). Copyright © 2001, 2006 by The Riverside Publishing Company. All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording or by any information storage or retrieval system without the proper written permission of The Riverside Publishing Company, unless such copying is expressly permitted by federal copyright law. Address inquiries to Contracts and Permissions Department, The Riverside Publishing Company, 3800 Golf Road, Suite 100, Rolling Meadows, Illinois 60008-4015.

THE IOWA TESTS

PROFILE NARRATIVE FOR ELIOT JOHNSON
Iowa Tests of Basic Skills® (ITBS®)

Student: Johnson, Eliot
Class/Group: Ness
Building: Longfellow
System: Dalen Community

Student ID: 0000145652
Form/Level: A/9
Test Date: 04/2002
Norms: Spring 2000
Order No.: 002-A70000028-0-002
Page: 1
Grade: 3



Legend: SS = Standard Score
GE = Grade Equivalent
NPR = National Percentile Rank
*Not Included in Totals or Composite
National Percentile Rank

Message from School:

This space may be left blank for the teacher to write a message or may be used for a predefined message that the school can provide.

Scores for Eliot Johnson:

Eliot was given the *Iowa Tests of Basic Skills* in April 2002. At the time of testing, he was in third grade at Longfellow in Dalen Community.

Eliot's Composite score is the score that best describes his overall achievement on the tests. Eliot's Composite national percentile rank of 27 means that he scored higher than 27 percent of third-grade students nationally. His overall achievement appears to be somewhat below average for third grade.

A student's ability to read is related to success in many areas of schoolwork. Eliot's Reading Comprehension score is low when compared with other students in third grade nationally.

A student's scores can be compared with each other to determine relative strengths and weaknesses. Spelling, Capitalization, Punctuation, and Math Problem Solving and Data Interpretation seem to be areas of relative strength for Eliot. Compared to Eliot's other test areas, Reading Comprehension and Word Analysis may need the most work.

Figure 3–6
Profile narrative report—parent copy.
Source: Iowa Test of Basic Skills® (ITBS®). Copyright © 2001, 2006 by The Riverside Publishing Company. All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying and recording or by any information storage or retrieval system without the proper written permission of The Riverside Publishing Company unless such copying is expressly permitted by federal copyright law. Address inquiries to Contracts and Permissions Department, The Riverside Publishing Company, 3800 Golf Road, Suite 100, Rolling Meadows, Illinois 60008-4015.

scores are reported for each pupil on each test (see Figure 3–4). The first row of the report for each student contains developmental standard scores (called *SSs* in this publisher's materials) for the 13 subtests and eight composites. The second row of scores are grade equivalents (GEs), and because the tests were given after the pupils had spent 8 months in the third grade, average performance for the country as a whole would be 3.8. The last two rows for each student contain stanines (NS) and percentile ranks (NPR) based on the spring 2000 national norm group. Looking at the scores for Eliot Johnson on the reading total score, we can see that the four score systems give an essentially equivalent picture of his performance. His reading total grade equivalent of 2.5 is below average, and this is also reflected in his stanine and NPR scores of 3 and 18, respectively. The standard scale score of 164 is also consistent with below-average performance in the spring of third grade. All four reference systems show him to be well below average in reading. Eliot's performance is noticeably better in language skills, where he comes in at grade level.

Figure 3–5 shows data for testings of a student in two successive years (fifth and sixth grades). The so-called "developmental scale" referred to toward the left is actually a scale of grade equivalents (GEs). Thus, this pupil had a vocabulary grade equivalent of 3.7 when she was tested the first time. By the next year her grade equivalent on this test was 4.6. Similar growth of approximately one GE is shown for each of the other subtests, although the level of performance in either year shows considerable variation from one subject to another.

The results show her scores generally to have been at or above the national average. An examination of her profile for the fifth-grade test indicates that she was strongest in capitalization, science, and reading comprehension skills and weakest in vocabulary and spelling. Some of the hazards of paying a great deal of attention to small ups and downs in a profile can be seen in a comparison of her performance on successive testings. Although the profile shows a relatively consistent pattern of highs and lows over the years, relative superiority changes somewhat from one year to the next.

Figure 3–6 shows a second type of profile chart for the ITBS. Here, the scores for Eliot Johnson (see Figure 3–4) are shown for each of the separate subtests of the battery. Note that in this case the different tests are represented by separate bars rather than by points connected by a line. The scale used in this case is a percentile rank scale, but in plotting percentile values, appropriate adjustments in the scale have been made to compensate for the inequality of percentile units. That is, the percentile points have been spaced in the same way that they are in a normal curve, being more widely spaced at the upper and lower extremes than in the middle range. This percentile scale corresponds to the scale called Percentile Equivalents in Figure 3–3. By this adjustment, the percentile values for an individual are plotted on an equal unit scale. A given linear distance can reasonably be thought to represent the same difference in amount of ability, whether it lies high in the scale, low in the scale, or near the middle of the scale. By the same token, the same distance can be considered equivalent from one test to another.

In the profile in Figure 3–6, the middle 50% is shaded to indicate a band of average performance for the norm group. The scores of this student have been plotted as bars that extend from the left side of the chart. For this type of norm, the average of the group constitutes the anchor point of the scale, and the individual scores can be referred to this base level. This type of figure brings out the individual's strengths and weaknesses quite clearly. Note also that the numerical values for this student's percentile ranks in the national norm group are given to the left of the profile. In addition, this particular test publisher's scoring service provides a narrative interpretation of the profile. Such an interpretation can also help draw the attention of teachers and parents to noteworthy features of the student's performance. Because this profile is intended to serve as a report to parents, there is also space for teacher comments.

The profile chart is a very effective way of representing an individual's scores, but profiles must be interpreted with caution. First, procedures for plotting profiles assume that the norms for the tests are comparable. For this to be true, age, grade, or percentile scores must be based on equivalent groups for all the tests. We usually find this to be the case for the subtests of a test battery. Norms for all the subtests are established at the same time, on the basis of testing the same group. This guarantee of comparability of norms for the different component tests is one of the most attractive features of an integrated test battery. If separately developed tests are plotted in a profile, we can usually only hope that the groups on which the norms were established were comparable and that the profile is an unbiased picture of relative achievement in different fields. When it is necessary to use tests from several different sources, one way to be sure of having equivalent norm groups is to develop local norms on a common population and to plot individual profiles in terms of those local norms.

A second problem in interpreting profiles is that of deciding how much attention to pay to the ups and downs in the profile. Not all the differences that appear in a profile are meaningful, either in a statistical or in a practical sense. We must decide which of the differences deserve some attention on our part and which do not. This problem arises because no test score is completely exact. No magic size exists at which a score difference suddenly becomes worthy of attention, and any rule of thumb is at best a rough guide. But, differences must be big enough so that we can be reasonably sure (1) that they would still be there if the person were tested again and (2) that they make a practical difference in terms of what they imply for performance, before we start to interpret them and base action on them. We will return to this topic during our discussion of reliability in Chapter 4.

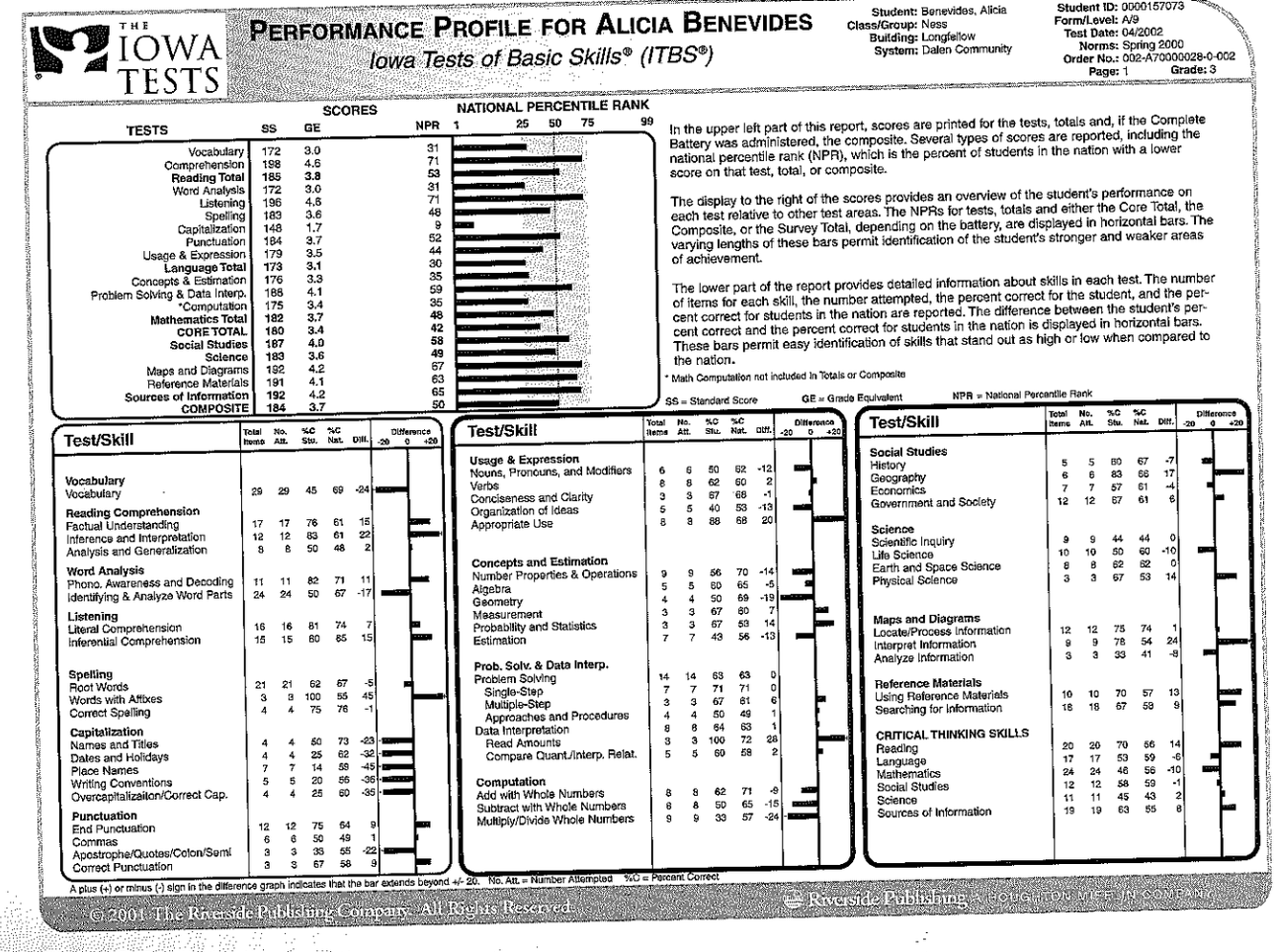
CRITERION-REFERENCED REPORTS

Interest in criterion-referenced interpretations of test scores has led test publishers to produce a profile of student performance based on specific item content. A well-designed test will include items that tap various aspects of skill or knowledge development. Modern test scoring and computer technology have made it possible to report a student's performance on subsets of items that are homogeneous with respect to a particular kind of content. An example of such a report for the ITBS is shown in Figure 3–7.

The report presented in Figure 3–7 lists each subtopic for each test of the ITBS, along with the number of items assessing that skill. The number of items the student attempted, the percentage of items correct for the student, the percentage correct for the nation, and the difference in percent correct for this student from the national norm group are also given. This report allows the teacher to identify specific strengths and weaknesses at a more fine-grained level than is possible with the ordinary norm-referenced report. For example, this student seems to have particular problems with the use of apostrophes, quotes, and colons, although her overall punctuation performance is average. Although each subskill is measured by too few items to yield a very reliable assessment, the information can be valuable to the classroom teacher in designing the instructional program for the individual student. Skills marked with a plus (+) represent areas of relative strength for the student, while those marked with a minus (–) are areas of relative weakness.

An even more detailed description of this student's performance can be provided in an individual item analysis such as that illustrated in Figure 3–8, which shows part of the class results for math concepts and estimation. Each column represents a student and each row corresponds to an item. The item numbers are given, organized by the skill they measure, and the student's response to the item is indicated if it was incorrect (a solid dot indicates the student got the item correct and an open dot indicates an omission). From the information on this chart the teacher can see that eight students got item 23 correct, one student omitted the item, seven chose alternative A, two chose alternative C, and one chose alternative D. By looking for commonly made errors such as alternative A, the teacher can diagnose particular skill areas where the students need extra work. For comparison purposes, percentages correct are given for each item for the national norm group and the school system, as well as this particular class.

Figure 3–8 gives student-by-student detail, but for examining the strengths and weaknesses of the class as a whole, information such as that provided in Figure 3–9 may be more useful. This report compares the performance of this class with that of the national norm group,



THE IOWA TESTS

PERFORMANCE PROFILE FOR ALICIA BENEVIDES

Iowa Tests of Basic Skills® (ITBS®)

Student: Benevides, Alicia
Class/Group: Ness
Building: Longfellow
System: Dalen Community

Student ID: 0000157073
Form/Level: A/9
Test Date: 04/2002
Norms: Spring 2000
Order No.: 002-A70000028-0-002
Page: 1 Grade: 3

TESTS	SS	GE	NPR
Vocabulary	172	3.0	31
Comprehension	198	4.6	71
Reading Total	185	3.8	53
Word Analysis	172	3.0	31
Listening	196	4.5	71
Spelling	183	3.6	48
Capitalization	148	1.7	9
Punctuation	184	3.7	52
Usage & Expression	179	3.5	44
Language Total	173	3.1	30
Concepts & Estimation	176	3.3	35
Problem Solving & Data Interp.	186	4.1	59
*Computation	175	3.4	35
Mathematics Total	182	3.7	48
CORE TOTAL	180	3.4	42
Social Studies	187	4.0	58
Science	183	3.6	49
Maps and Diagrams	192	4.2	67
Reference Materials	191	4.1	63
Sources of Information	192	4.2	65
COMPOSITE	184	3.7	50

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

The display to the right of the scores provides an overview of the student's performance on each test relative to other test areas. The NPRs for tests, totals and either the Core Total, the Composite, or the Survey Total, depending on the battery, are displayed in horizontal bars. The varying lengths of these bars permit identification of the student's stronger and weaker areas of achievement.

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 the nation are reported. The difference between the student's percent correct and the percent correct for students in the nation is displayed in horizontal bars. These bars permit easy identification of skills that stand out as high or low when compared to the nation.

* Math Computation not included in Totals or Composite

SS = Standard Score GE = Grade Equivalent NPR = National Percentile Rank

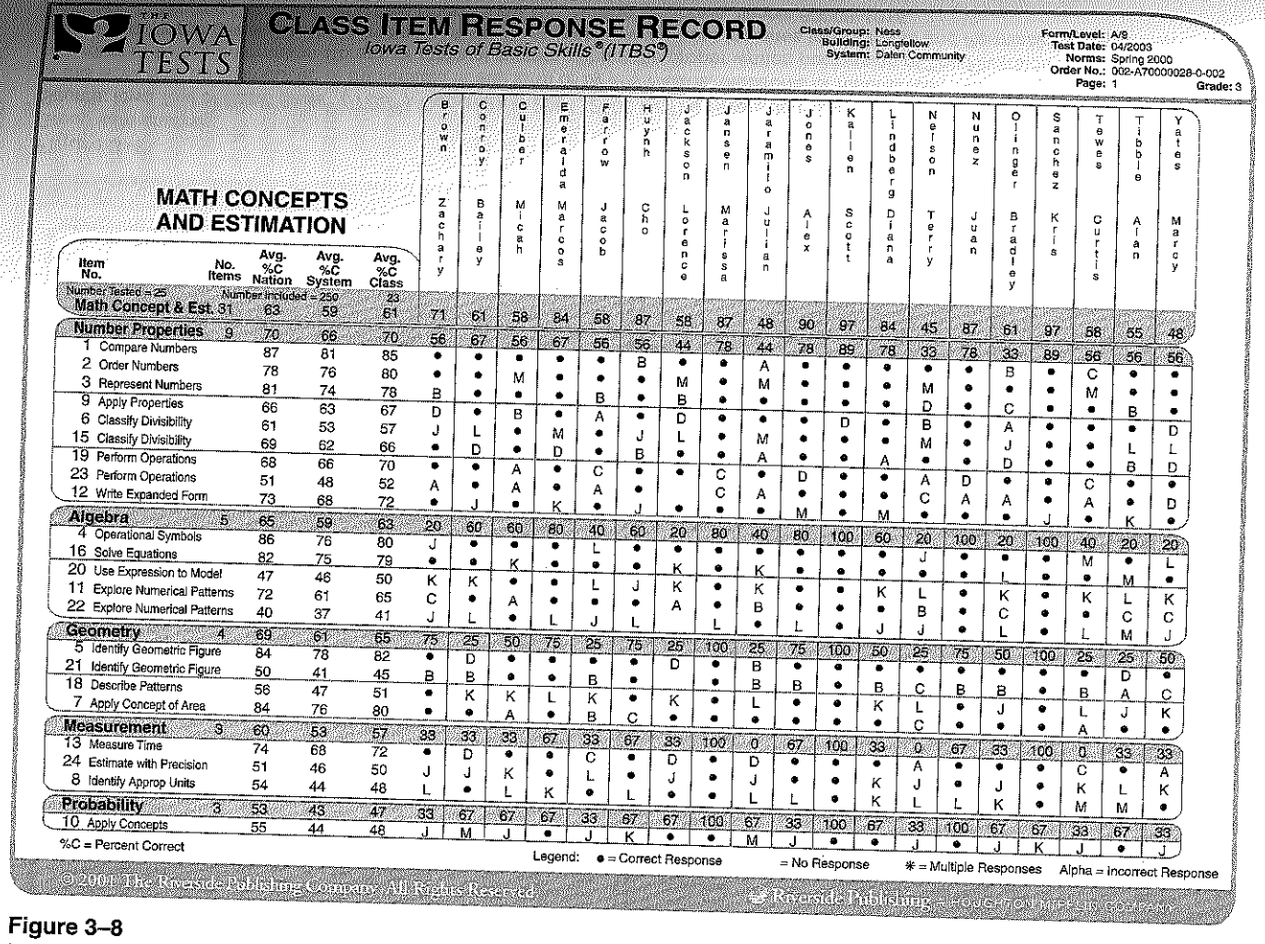
Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.
Vocabulary					
Vocabulary	29	29	45	69	-24
Reading Comprehension					
Factual Understanding	17	17	76	61	15
Inference and Interpretation	12	12	83	61	22
Analysis and Generalization	8	8	50	48	2
Word Analysis					
Phono. Awareness and Decoding	11	11	82	71	11
Identifying & Analyze Word Parts	24	24	50	67	-17
Listening					
Literal Comprehension	16	16	81	74	7
Inferential Comprehension	15	15	80	65	15
Spelling					
Root Words	21	21	62	67	-5
Words with Affixes	3	3	100	55	45
Correct Spelling	4	4	75	76	-1
Capitalization					
Names and Titles	4	4	50	73	-23
Dates and Holidays	4	4	25	62	-37
Place Names	7	7	14	58	-45
Writing Conventions	5	5	20	55	-35
Overcapitalization/Correct Cap.	4	4	25	60	-35
Punctuation					
End Punctuation	12	12	75	64	9
Commas	6	6	50	49	1
Apostrophe/Quotes/Colon/Semi	3	3	33	55	-22
Correct Punctuation	3	3	67	58	9

Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.
Usage & Expression					
Nouns, Pronouns, and Modifiers	6	6	50	62	-12
Verbs	8	8	62	60	2
Conciseness and Clarity	3	3	67	68	-1
Organization of Ideas	5	5	40	53	-13
Appropriate Use	8	8	88	68	20
Concepts and Estimation					
Number Properties & Operations	9	9	56	70	-14
Algebra	5	5	60	65	-5
Geometry	4	4	50	69	-19
Measurement	3	3	67	60	7
Probability and Statistics	3	3	67	53	14
Estimation	7	7	43	56	-13
Prob. Solv. & Data Interp.					
Problem Solving	14	14	63	63	0
Single-Step	7	7	71	71	0
Multiple-Step	3	3	67	61	6
Approaches and Procedures	4	4	50	49	1
Data Interpretation	8	8	64	63	1
Read Amounts	3	3	100	72	28
Compare Quant./Interp. Relat.	5	5	60	58	2
Computation					
Add with Whole Numbers	8	8	62	71	-9
Subtract with Whole Numbers	8	8	50	65	-15
Multiply/Divide Whole Numbers	9	9	33	57	-24

Test/Skill	Total Items	No. Att.	%C Stu.	%C Nat.	Diff.
Social Studies					
History	5	5	80	67	-7
Geography	6	6	83	66	17
Economics	7	7	57	61	-4
Government and Society	12	12	67	61	6
Science					
Scientific Inquiry	9	9	44	44	0
Life Science	10	10	50	60	-10
Earth and Space Science	8	8	62	62	0
Physical Science	3	3	67	53	14
Maps and Diagrams					
Locate/Process Information	12	12	75	74	1
Interpret Information	9	9	78	54	24
Analyze Information	3	3	33	41	-8
Reference Materials					
Using Reference Materials	10	10	70	57	13
Searching for Information	18	18	67	58	9
CRITICAL THINKING SKILLS					
Reading	20	20	70	56	14
Language	17	17	53	59	-6
Mathematics	24	24	46	56	-10
Social Studies	12	12	58	59	-1
Science	11	11	45	43	2
Sources of Information	19	19	63	55	8

A plus (+) or minus (-) sign in the difference graph indicates that the bar extends beyond +/- 20. No. Att. = Number Attempted %C = Percent Correct

Figure 3–7
Student criterion-referenced skills analysis.



THE IOWA TESTS

CLASS ITEM RESPONSE RECORD

Iowa Tests of Basic Skills® (ITBS®)

Class/Group: Ness
Building: Longfellow
System: Dalen Community

Form/Level: A/9
Test Date: 04/2003
Norms: Spring 2000
Order No.: 002-A70000028-0-002
Page: 1 Grade: 3

MATH CONCEPTS AND ESTIMATION

Item No.	No. Items	Avg. %C Nation	Avg. %C System	Avg. %C Class	Brown, Zachary	Conroy, Bailey	Cuber, Micah	Emerald, Marcos	Farrow, Jacob	Huynh, Cho	Jackson, Lorence	Jansen, Marissa	Jarmillo, Julian	Jones, Alex	Kalen, Scott	Lindberg, Diana	Nelson, Terry	Nunez, Juan	Olinger, Bradley	Sanchez, Kris	Tewes, Curtis	Tibble, Alan	Yates, Marcy
Number Tested = 25		Number Included = 250		23																			
Math Concept & Est.	31	63	59	61	71	61	58	84	58	87	58	87	48	90	97	84	45	87	61	97	58	55	48
Number Properties	9	70	66	70	56	67	56	67	56	56	44	78	44	78	89	78	33	78	33	89	56	56	56
1 Compare Numbers		87	81	85	•	•	•	•	•	B	•	•	A	•	•	•	•	•	B	•	C	•	•
2 Order Numbers		78	76	80	•	•	M	•	•	•	M	•	M	•	•	•	M	•	•	•	M	•	•
3 Represent Numbers		81	74	78	B	•	•	•	B	•	B	•	•	•	•	•	D	•	C	•	•	B	•
9 Apply Properties		66	63	67	D	•	B	•	A	•	D	•	•	D	•	D	B	•	A	•	•	•	D
6 Classify Divisibility		61	53	57	J	L	•	M	•	J	L	•	M	•	•	•	M	•	J	•	•	L	L
15 Classify Divisibility		69	52	66	•	D	•	D	•	B	•	•	A	•	•	A	•	•	D	•	•	B	D
19 Perform Operations		68	66	70	•	•	A	•	C	•	•	C	•	D	•	•	A	D	•	•	C	•	D
23 Perform Operations		51	48	52	A	•	A	•	A	•	•	C	A	•	•	•	C	A	A	•	A	•	•
12 Write Expanded Form		73	68	72	•	J	•	K	•	J	•	•	•	M	•	M	•	•	•	J	•	K	•
Algebra	5	65	59	63	20	60	60	80	40	60	20	80	40	80	100	60	20	100	20	100	40	20	20
4 Operational Symbols		86	76	80	J	•	•	•	L	•	•	•	K	•	•	•	J	•	I	•	M	•	L
16 Solve Equations		82	75	79	•	•	K	•	•	K	•	•	•	•	•	•	•	•	•	•	•	M	•
20 Use Expression to Model		47	46	50	K	K	•	•	L	J	K	•	K	•	•	K	L	•	K	•	K	L	K
11 Explore Numerical Patterns		72	61	65	C	•	A	•	•	•	A	•	B	•	•	•	B	•	C	•	•	C	C
22 Explore Numerical Patterns		40	37	41	J	L	•	L	J	L	•	L	•	L	•	J	J	•	L	•	L	M	J
Geometry	4	69	61	65	75	25	50	75	25	75	25	100	25	75	100	50	25	75	50	100	25	25	50
5 Identify Geometric Figure		84	78	82	•	D	•	•	•	•	D	•	B	•	•	•	•	•	•	•	•	D	•
21 Identify Geometric Figure		50	41	45	B	B	•	•	B	•	•	•	B	B	•	B	C	B	B	•	B	A	C
18 Describe Patterns		56	47	51	•	K	L	K	•	K	•	•	L	•	•	K	L	•	J	•	L	J	K
7 Apply Concept of Area		84	76	80	•	A	•	B	C	•	•	•	•	•	•	•	C	•	•	•	A	•	•
Measurement	3	60	53	57	33	33	33	67	33	67	33	100	0	67	100	33	0	67	33	100	0	33	33
13 Measure Time		74	68	72	•	D	•	•	C	•	D	•	D	•	•	•	A	•	•	•	C	•	A
24 Estimate with Precision		51	46	50	J	J	K	•	L	•	J	•	J	•	•	K	J	•	•	•	K	L	K
8 Identify Approp Units		54	44	48	L	•	L	K	•	L	•	•	L	L	K	K	L	L	K	•	M	M	•
Probability	3	53	43	47	33	67	67	67	33	67	67	100	67	33	100	67	33	100	67	67	33	67	33
10 Apply Concepts		55	44	48	J	M	J	•	J	K	•	•	M	J	•	•	J	•	J	K	J	•	J

%C = Percent Correct

Legend: • = Correct Response = No Response * = Multiple Responses Alpha = Incorrect Response

Figure 3–8
Individual item analysis.

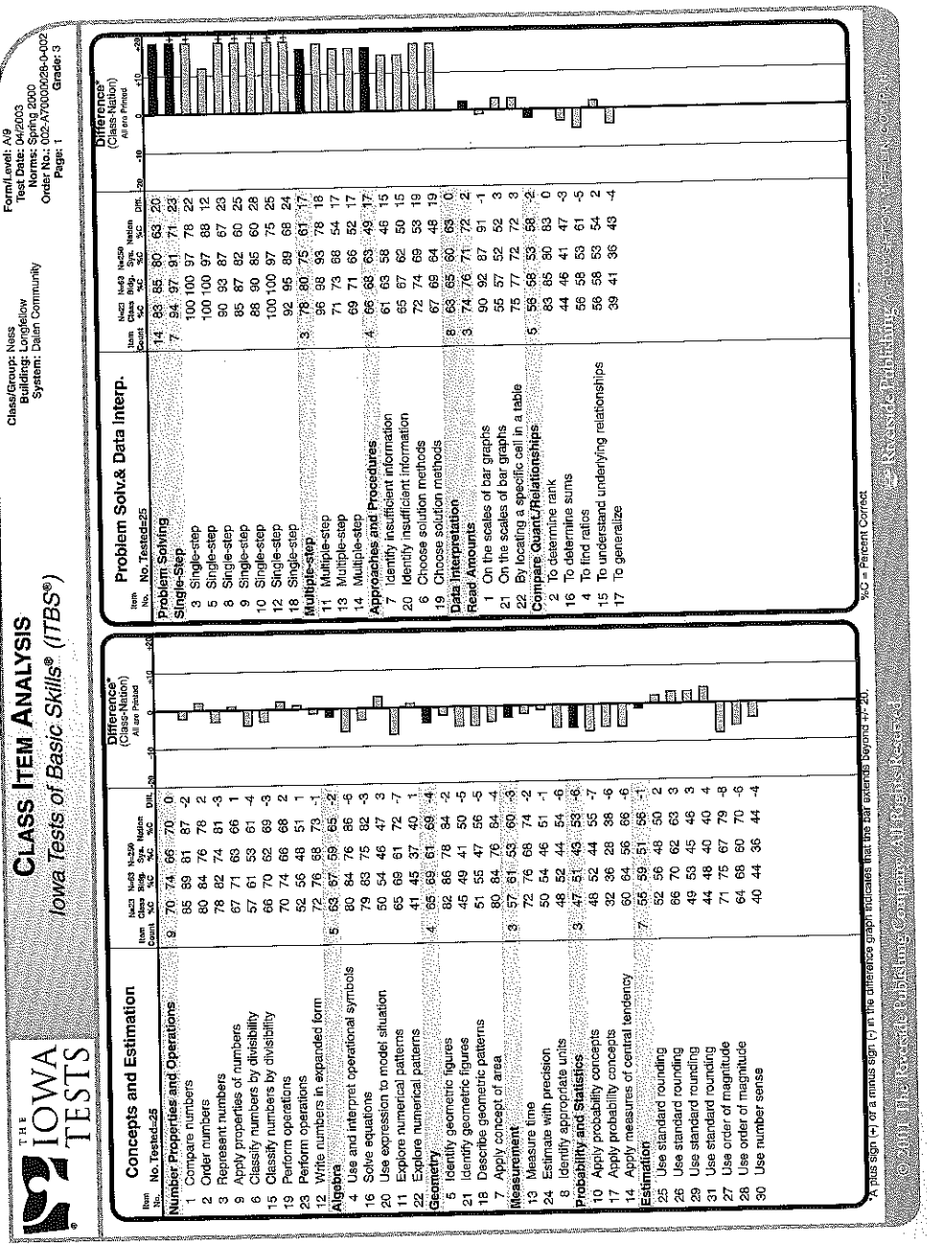


Figure 3–9
Group item analysis.

the building, and the school system. The results, shown item by item in terms of percent correct, are displayed both numerically and graphically. The two vertical lines indicate when the difference is less than 10% and, therefore, probably too small to be of interest. The results for this class show a broad pattern of performance above the norm group in problem solving and approaches and procedures with performance in other areas near the national norm. (Note that the complete table for the class would contain similar information about items covering other skills and knowledge areas.)

Figure 3–7 illustrates quite clearly the way content-based and norm-based frames of reference can coexist in the same test and can supplement each other in score interpretation. The report shows this student's performance, by content area, with reference to the number of items covering that content, the average performance of her class, and the average performance of the grade-equivalent national norm group. Additional reports are available that show, for example, the performance of the class on each item relative to national, system, and building norms (school performance) or that summarize the individual information in Figure 3–7 for the entire class. The publisher's catalog for this test lists more than 30 forms of reports that are available. However, it is important to keep in mind that criterion-referenced interpretations of standardized tests are based on very small numbers of items (one or two in some cases) for each content area or objective. Therefore, any conclusions based on such data must be tentative and should be confirmed using other sources of information.

NORMS FOR SCHOOL AVERAGES

Up to this point, we have asked how we can interpret an individual's standing on a test. Sometimes a question arises about the relative performance of a class, a school, a school district, or even the schools of a whole state. The current emphasis on accountability in education provides ample reason for educators to be concerned about evaluating the performance of students taken as groups. When evaluating the achievement of a school in relation to other schools, it is necessary to have norms for school averages.

It should be clear that the variation from school to school in average ability or achievement will be much less than the variation from pupil to pupil. No school average comes even close to reaching the level of its ablest student, and no average drops anywhere near the performance of the least able. Thus, a single pupil at the beginning of fifth grade who gets a reading grade equivalent of 6.2 might fall at the 75th percentile, whereas a school whose *average* reading grade equivalent of beginning fifth-graders is 6.2 might fall at about the 94th percentile of schools. The relationship between norms for individuals and groups is illustrated more fully in Table 3–8. The two distributions center at about the same point, but the greater variation among individuals quickly becomes apparent. On this test, an individual grade equivalent of 6.6 ranks at the 79th percentile, but a school in which the average performance is a grade equivalent of 6.6 is at the 92nd percentile. The same effect is found for performances that are below average.

When a school principal or an administrator in a central office is concerned with interpreting the average performance in a school, norms for school averages are the appropriate ones to use, and it is reasonable to expect the test publisher to provide them. The better test publishers will also provide item analyses and criterion-referenced reports at the level of the class, building, district, and state.

Table 3–8
Individual and School Average Norms for the Iowa Tests of Basic Skills–Form A, Fall Norms for Grade 5 Vocabulary Test

Grade Equivalent	Percentile Rank for Individual	Percentile Rank for School Averages
9.4	99	99
7.2	87	97
7.0	83	96
6.6	79	92
5.7	61	67
5.5	58	62
5.2	50	48
5.0	46	43
4.4	34	23
4.0	24	11
3.6	16	5
3.5	14	4
2.4	5	1
1.0	1	1

CAUTIONS IN USING NORMS

For a test that assesses standing on some trait or competence in some area of knowledge, norms provide a basis for interpreting the scores of an individual or a group. Converting the score for any test taken singly into an age or grade equivalent, percentile rank, or standard score permits an interpretation of the level at which the individual is functioning on that particular test. Bringing together the set of scores for an individual in a common unit of measure, and perhaps expressing these scores in a profile, brings out the relative level of performance of the individual in different areas.

The average performance for a class, a grade group in a school, or the children in the same grade throughout a school system may be reported similarly. We can then see the average level of performance within the group on some single function or the relative performance of the group in each of several areas. Norms provide a frame of reference within which the picture may be viewed and bring all parts of the picture into a common focus. Now, what does the picture mean, and what should we do about it?

Obviously, it is not possible, in a few pages, to provide a ready-made interpretation for each set of scores that may be obtained in a practical testing situation. However, we can lay out a few general guidelines and principles that may help to forestall some unwise interpretations of test results.

The most general point to keep in mind is that test results, presented in any normative scale, are a *description of what is,* not a *prescription of what should be* or a statement of what will be (although they may give an indication of probable future performance). The results make it possible to compare an individual or a class with other individuals and classes with respect to one or more aspects of accomplishment or personality, but they do not in any absolute sense tell us whether the individual is doing "well" or "poorly." They do not provide this information for several reasons.

Normative Scores Give Relative Rather Than Absolute Information. They tell whether an individual pupil's achievement is as high as that of other pupils or whether a class scores as high as other classes. But they do not tell us whether the basic concepts of the number system are being mastered or whether the pupils read well enough to comprehend the instructions for filling out an income tax return. Furthermore, they give us little guidance on how much improvement we might expect from *all* pupils if our educational system operated throughout at higher efficiency.

Remember that by the very nature of relative scores, there will be as many people below average as above. When "the norm" means the average of a reference group, it is a statistical necessity that about half of the group be, to a greater or lesser degree, below average. There has been an enormous amount of foolishness—both in single schools and in statewide legislation—about bringing all pupils "up to the grade norm." This might conceivably be done temporarily if we had a sudden and enormous improvement in educational effectiveness; however, the next time new norms were established for the test it would take a higher absolute level of performance to, say, read at the sixth-grade level. So we would be back again with half of the pupils falling at or below average. And if the effectiveness of the schools were to return to the former level, we would be faced with the unhappy prospect of more than half of the students testing "below grade level."

The relative nature of norms has been recognized in the criterion-referenced test movement. When a teacher or a school is concerned with appraising mastery of some *specific* instructional objective, it may be more useful to develop test exercises that appraise that objective, to agree on some standard as representing an acceptable level of mastery, and to determine which students do and which do not have mastery of that specific objective than it would be to know how the students from this school perform relative to those from other schools. In the context described, it is possible for all students to achieve mastery, but some will get there faster than others. Even in a criterion-referenced framework there will still be differences among individuals in their levels of accomplishment.

Output Must Be Evaluated Relative to Input. Test results typically give a picture of output—of the individual or of the group as it exists at the time of testing, after a period of exposure to educational effort. But what of the input? Where did the group start?

The notion of input is a complex and rather subtle one. Our conception of input should include not only earlier status on the particular ability being measured and individual potential for learning, as far as we are able to appraise this, but also the familial circumstances and environmental supports that make it easier for some children to learn than for others. Parental aspirations for the child, parental skills at teaching and guidance of learning, parental discipline and control, linguistic patterns, and cultural resources in the home are part of the input just as surely as are the biological characteristics of the young organism. Furthermore, peer group and community attitudes are an additional real, though possibly modifiable, part of the input as far as the prospects for learning for a given child are concerned. We must recognize that the adequate appraisal of input is no simple matter, and that, correspondingly, the appraisal of output as "satisfactory" or "unsatisfactory" is something we can do with only modest confidence.

Output Must Be Evaluated Relative to Objectives. The design, content, and norms for published standardized tests are based on their authors' perceptions of common national curricular objectives. The topics included, their relative emphasis, and the levels at which they are introduced reflect that perceived general national pattern. To the extent, then, that a given school system deviates in its objectives and curricular emphases from the national pattern, as interpreted by the test maker, its output at a given grade level can be expected to deviate from the national norms. If computational skills receive little emphasis, it is reasonable to find that computational facility will be underdeveloped. If map reading has been delayed beyond the grade level at which it is introduced into the test, it is reasonable to find that relative standing on that part of the test will suffer. Unevenness of the local profile, in relation to national norms, should always lead one to inquire whether the low spots represent failures of the local program to achieve its objectives or a planned deviation of emphasis from what is more typical of schools nationally. Low performance that results from conscious curricular decisions would be much less cause for alarm than a similar level of performance would be in an area of curricular emphasis. Which of these conditions obtained will no doubt influence what is done with the finding.

To the extent that individual states have uniform objectives for all districts within their boundaries, well-designed standardized tests measuring achievement of these objectives often are available through contract arrangements with test publishers. Several states now contract with organizations that specialize in test development to have tests constructed according to specifications provided by the state board of education. Such tests usually are intended to be used at particular points in the educational program, such as the transitions from elementary school to middle school, middle school to high school, and near the end of high school.

If these considerations and some of the caveats discussed in the next two chapters are borne in mind, the teacher, principal, superintendent, or school board will be able to interpret the reported test results with increased wisdom and restraint.

A THIRD FRAME OF REFERENCE: ITEM RESPONSE THEORY

Many of the recent developments in testing stem from what has come to be called **item response theory** (IRT), or **latent trait theory.** (We will use the terms more or less interchangeably.) The origins of this approach go back before 1910, and the basic logic of the theory was pretty well worked out by E. L. Thorndike and L. L. Thurstone in the 1920s (see R. M. Thorndike, 1999a), but the practical application of the theory has depended on the availability of computers. IRT itself has in turn shaped the ways in which computers are used in testing. Let us look at the set of interlocking developments that stem from the interactions of the theoretical models and the availability of computers to implement them. Our discussion will be in terms of cognitive abilities, but item response theory can be applied equally well to personality measures and measures of other psychological traits. Additional details are provided in Yen and Fitzpatrick (2006).

Latent trait theory assumes the existence of a relatively unified underlying trait, or characteristic, that determines an individual's ability to succeed with some particular type of cognitive task. Possible attributes might be *knowledge of word meanings, arithmetic reasoning,* or *spatial visualizing.* We can represent the trait as a linear scale (as shown in Figure 3–10) on which both tasks and people can be placed in an ordered sequence. The tasks in this example are words to be defined, and the trait is knowledge of word meanings. A given test may contain items measuring several such dimensions, but each item should be a relatively pure measure of only one trait.

Figure 3–10
Scale of word knowledge.

For the *tasks,* the scale can be thought of as a scale of difficulty (we could also think of this as the ability requirement of the item), so the words in the illustration go from very easy on the left to quite difficult on the right. *The difficulty of an item is defined as the ability level at which half of the examinees will get the item correct.* For any single item, the point on the scale where the probability of a correct response is 50% is called *b*. Thus, the five words to be defined have five different *b*-values or item difficulties.

For *people,* the scale can be thought of as a scale of ability. A person's ability level is defined by the tasks that that person *can just about do*—that is, the difficulty level at which the examinee would get half of the items correct. Thus, Joe can most likely define *borrow* because his ability exceeds the difficulty of the item, but he is unlikely to be able to define *incision* because its difficulty exceeds his ability. The likelihood that Billy can correctly define *borrow* is relatively low, but he could probably define *dig.* It is the joint ordering of the people and the tasks which define the scale. The term used to refer to a person's ability level is the Greek letter theta (θ). Billy has a relatively low value of θ, Joe and Mr. Jones have intermediate values, and Dr. Thompson's θ-value is quite high.

It is important to note that in this model, a person's ability level is subject to change over time. That is, if Billy is 6 years old, we can expect that his position on the scale of ability, his θ, will change as he matures. Conversely, we would expect the *b*-values, the relative difficulty of the words in a large and representative sample from the population, to remain nearly constant for long periods of time. The stability of the difficulty scale is what gives meaning to the ability scale.

The ability/difficulty scale is an arbitrary one, just as the Fahrenheit scale of temperature is. We could use a scale with different-sized units (for example, Celsius) or a different zero point (for example, Celsius or Kelvin). But, for a given scale, the units are presumably equal throughout the scale and the *relative* position of a person or task does not depend either on the size of the units or on the placement of the zero point. As an example of our temperature-scale analogy, consider that a summer day may be warmer than a winter day, and this fact does not depend on whether the summer temperature is expressed as 20°C or 68°F and the winter temperature is 0°C or 32°F. The relative positions of the two days are the same in either scale, and a spring day of 10°C or 50°F would be halfway between them on the scale. Likewise, a heat greater than 100°C (212°F) has the ability to cause water to boil (at standard atmospheric pressure) and a heat below this point does not.

The relationship between ability level and passing an item of a given difficulty is not an all-or-none matter but, instead, is a question of probability. The form of the relationship between ability and the probability of passing an item is shown in Figure 3–11. The graph in this figure is called the **item characteristic curve,** or **item trace line,** and there is one such curve for every item. The item characteristic curve shows that for an item of a given difficulty, the probability that a person will pass the item increases as their ability level goes up. Thus, if we include the word *incision* on a test given to a group of people at Joe's level of ability, about 25% would be

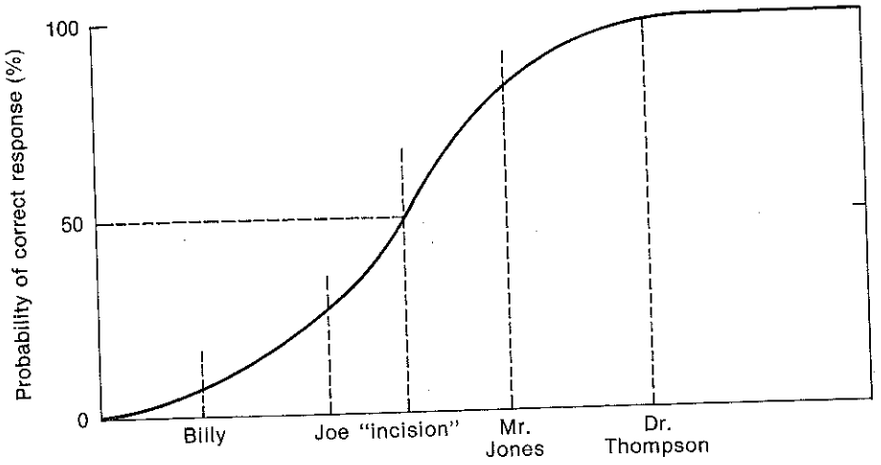


Figure 3–11
Item characteristic curve for the meaning of the word *incision*.

able to define the word, while among those at Mr. Jones' level, about 85% would be able to provide a correct definition. Turning things around, Joe could define about 25% of the words whose difficulty was the same as *incision*'s, while Mr. Jones could provide correct definitions for about 85% of such words.

As we see from Figure 3–11, the probability of a person's passing an item as a function of ability level is expressed by a curve that is quite flat at the two extremes but rises steeply around the level that matches the difficulty of the item. The test item differentiates most effectively between those whose abilities are somewhat above and those whose abilities are somewhat below the difficulty level of the task, but provides very little differentiation among those whose abilities are very high or very low. Dr. Thompson and other people at her ability level would pass almost all the items at the difficulty level represented by *incision*, and we would know that she had high verbal ability, but we would not know *how* high. We would need words in the difficulty range from *prologue* to *tenet* (see Figure 3–10) in order to locate Dr. Thompson's ability with any precision, because these words are sufficiently difficult that she would not get all of them correct.

The trace line of an item is a characteristic of the item that does not depend on the people taking the test, but our ability to reveal the entire curve depends on applying the item to a sufficiently heterogeneous group that people over the full range of ability are represented. If we gave items like *incision* only to people like Joe and Mr. Jones, we could only see that part of the curve that falls between them. Because this range covers the difficulty level of the item (the point on the ability dimension where 50% of the examinees would get it correct or the item's *b*-value), the result would not be too serious, but if only people like Billy or like Dr. Thompson were included, we would be able to tell very little about the item.

Each item has its own characteristic curve that is defined by its *difficulty level* (the 50% point, which is called its *b* parameter) and its *discrimination*. **Discrimination** is the ability of the item to separate those of higher ability from those of lower ability. A widely used index of item discrimination is the correlation between score on the single item and score on the entire set of items that measure the dimension. The ability of an item to allow for discrimination of

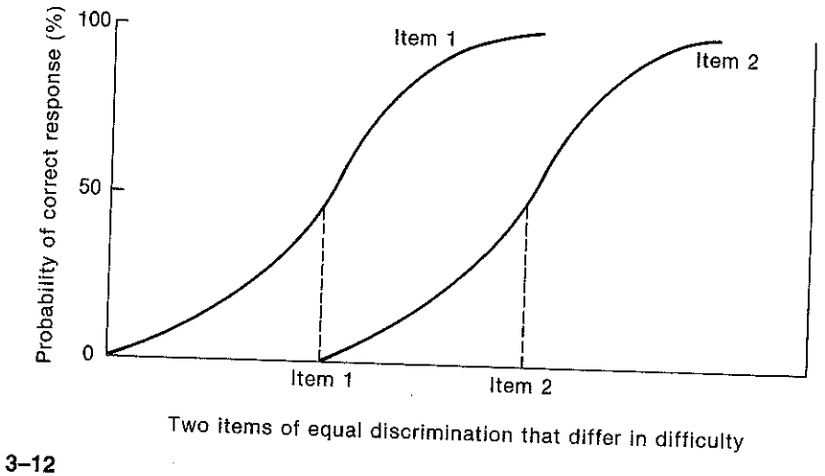


Figure 3–12
Item characteristic curves for items that differ in difficulty.

different levels of ability is a function of the rate at which the probability of getting the item correct changes with ability. Graphically, this rate of change can be seen as the *slope of the item characteristic curve*. The label given to the slope of the curve is *a*. *The a parameter is the slope of the curve at the 50% point*. Figure 3–12 shows curves representing two items that differ in difficulty but are equal in discrimination. The shapes of the two curves are the same, but their *b* parameters, the ability levels required to have a 50% chance of getting the item correct, are different.

Figure 3–13 shows two items that are of the same difficulty but differ in discrimination. The *a* parameter, or rate of change in the probability of getting the item correct, is much steeper for Item 1 than for Item 2. There is a higher correlation between item score and test score. More discriminating items are more sensitive to differences in the ability levels of examinees. To illustrate

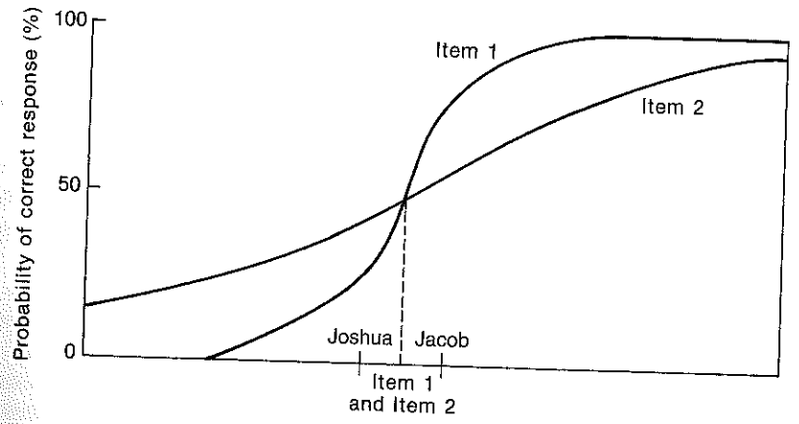


Figure 3–13
Item characteristic curves for two items of equal difficulty that differ in discrimination.

this point, look at the two people who are plotted on the ability continuum. On Item 1, Joshua has a probability of about .25 of getting the item correct and Jacob has a probability of about .75 of correctly answering the item. However, on Item 2 the probabilities are much closer together. Because the slope of the item characteristic curve, and hence the correlation of the item with ability level, is lower, the difference in probability of a correct response is only about 20% instead of 50%. The item does not differentiate between the two examinees as well. (However, Item 2 does provide some information over a wider range of ability levels.)

For items where examinees choose their answers from a set of given alternatives (true–false, multiple choice, and similar item forms called *select-response items*), the curve has a third feature, the probability of getting the item correct by chance, or guessing. This effect is seen when the curve flattens out at some probability greater than zero. The probability value at which the curve flattens out is called the *c* parameter or the guessing parameter. Figure 3–14 provides an example of such a curve in which the value of *c* is 20%. This is different from the situation for Item 2 in Figure 3–13. In the latter case, although the graph does not go down far enough for the trace line to reach zero, the curve is still descending. If people of extremely low ability had been included in the sample, we should find an ability level where examinees have zero chance of getting the item correct. This occurs when examinees must produce their own answers for the items, such as writing definitions of words rather than selecting the definition from a list of alternatives. Item trace lines for items where the examinees produce their own responses, such as short answer and definition items, can always reach zero; those for select-response items never can.

Computer Adaptive Testing (CAT)

The most efficient measurement using the IRT framework requires that we be able to give each person the items most appropriate for people at their ability level. It is this feature for which a computer is essential, and the technology for selecting which items to administer is known as **computer adaptive testing,** or CAT. The rapid development and widespread availability of computers has combined with item response theory to lead to the development of CAT. By adaptive testing, we mean the rapid adjustment of the difficulty level of the test tasks to the ability level of the person being tested. As we indicated earlier, tasks that are much too difficult or much too easy give little new information about the ability level of an examinee. For example, our knowing that an examinee who is an applicant for college admission could define the word *dig* would tell us essentially nothing about whether the applicant was a promising candidate. *Dig* is a word that *any* high school senior could be expected to know. We would also learn relatively little if the examinee failed on *fracedinous*. Probably not one high school senior in 100,000 would know the meaning of the word. (*Webster's Unabridged Dictionary* defines it as "producing heat by putrefaction [obsolete].") We gain useful information by testing with tasks on which we *do not* know in advance whether the examinee will succeed.

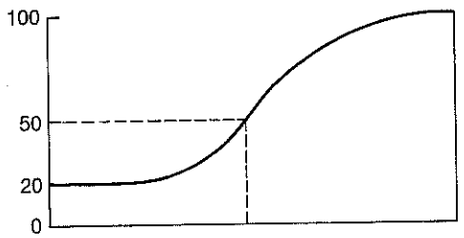


Figure 3–14
Effect of guessing on the item characteristic curve.

Ideally, in adaptive testing, we start at a difficulty level at which we are *most uncertain* whether the examinee can pass the item. This point would usually be at an item of about 50% difficulty for that age or grade group, because in the absence of any special information, this is our best guess of the person's ability level. Each correct response raises our estimate of the person's ability level somewhat and makes it appropriate to present next a somewhat more difficult item. As long as the examinee continues to get items correct, we continue to raise our estimate of the ability level and to raise the difficulty level of the next item presented. If our initial guess was accurate, and thus our starting level was close to the examinee's ability level, the examinee should quickly get an item incorrect, and we would lower our estimate slightly and drop back to a slightly easier item. With a sequence of passes and failures, we would soon zero in with items near the examinee's ability level, and the record of passes and failures would give a good final estimate of this ability.

Any given level of precision in the estimate can be achieved by well-designed adaptive testing, using perhaps half as many test items as would be required in a conventional test designed for use with a complete age or grade group. This is because in conventional tests, in order to include items that are appropriate for people of differing abilities, it is necessary to include items that are too easy or too hard for a given examinee. In conventional tests, only a relatively small proportion of the items are at the correct level for a particular person. We will discuss how precise our estimates are in Chapter 4.

A truly effective procedure for adaptive testing must entail searching the pool of items and finding one item that is of most appropriate difficulty, given our current estimate of the examinee's ability level. This searching is something that only a computer can do efficiently and rapidly enough for effective testing. The complete pool of items can be stored on a disk, coded by difficulty level and, if need be, by discrimination level and content category. The computer can be programmed to adjust its estimate of the examinee's ability level by an appropriate amount after each pass or failure and to seek out the yet unused item that best matches that new estimate. An ongoing record is maintained of the examinee's estimated ability level and of the precision of that estimate. The computer can be programmed to terminate testing either after a specified number of items has been administered or when a specified precision of estimate has been reached. This approach is almost universally used for testing programs that are administered by computer, such as the Scholastic Assessment Test (SAT) and Graduate Record Exams. See Wainer (1990) for a discussion of how to implement computer adaptive testing.

One feature of adaptive testing is that within any given group, no two individuals are likely to take exactly the same test, because each unique pattern of right and wrong answers produces a different test. A person with very consistent performance would take a short test because the process would focus in on the ability level quite quickly. On the other hand, a person showing variable performance, that is, passing some quite hard items while missing some relatively easy ones, would require a much longer test in order to reach the same level of precision in the ability estimate.

One of the interesting potential applications of item response theory and adaptive testing is the linking of grading procedures between classes or sections of a course. One of the motivating forces that led to the creation of standardized tests such as the SAT was the desire for comparable information for students from different schools. If some items that had been calibrated by IRT methods were included in locally developed tests in different schools or in different sections of a course where, for example, ability grouping had been used, the tests could be equated and the relative performance of pupils in different classes assessed on a common scale. The different levels of the Cognitive Abilities Test use item response theory to achieve a common scale, called a

universal scale score, for the different levels of the test. The American College Testing Program also uses IRT to create a common scale for several different tests that have no items in common but have been given to a common group of examinees.

If we have determined the difficulty values on a common scale for a sufficient pool of items, we can use items from that pool to do a number of useful and interesting things. We outline some of them here.

1. *Estimating ability level from any set of items.* If the difficulty scale values are available for a large pool of items measuring a common trait, an unbiased estimate (unbiased in a statistical sense, not in the social policy sense discussed in Chapter 5) of a person's ability level on that trait can be obtained from *any* set of items drawn from that pool. The precision of the estimate will depend on the number of items, increasing in precision as the number of items increases. It will also depend on how closely the difficulty of the items matches the ability level of the person, with accuracy increasing with the closeness of the match. But there will be no systematic error in the estimate in any case, assuming the probability of a correct response is not zero (or chance level for select-response items) or 1.0. In these two cases, the item gives no information about the person's ability.

2. *Preparing equivalent test forms.* By drawing from the pool sets of items having the same average difficulty value, the same spread of difficulty values, and the same average discrimination, we can prepare test forms that are equivalent in the sense that any given raw score signifies the same level of examinee ability, irrespective of the test form on which it is based. We could use the test forms to measure gains from instruction, giving one form before some instructional period, and another after. Or, in situations in which test security or examinee copying is likely to be a problem, different forms could be given to individuals in alternating seats. If, for some reason, one test were invalidated for a person, an alternate, equivalent form could be administered.

3. *Matrix sampling in testing.* At times, a researcher may wish to cover very completely some domain of content and may not be interested in making decisions about specific individuals but, instead, in assessing the performance level of a class, a school, or a school system. It is then not necessary that every test item be administered to every person. Each person can be given a fraction of the items, as long as each item has been given to some of the people. The investigator can then think of the class or school or school system as a composite "person." Upon determining the proportion of items passed at known difficulty scale values, the researcher can estimate an ability level for the group, either with respect to the complete domain or with respect to specific limited segments of that domain. This approach makes it possible to hold testing time within reasonable limits and yet to cover completely the domain of content in which the investigator is interested.

SUMMARY

A raw score, taken by itself, rarely has meaning. A score may be given meaning by a consideration of the domain of instructional content that the test items represent. The performance of individuals or groups can then be assessed either in terms of the percentage of the domain they have mastered or relative to a standard of performance set before the test is administered. These methods of giving meaning to a raw score are called criterion-referenced interpretations. They are appropriate for tests that focus on one or a small number of carefully defined objectives and for which standards of performance can be either empirically or logically derived.

Because many tests are designed to appraise several objectives, and because meaningful absolute standards of performance are not available for most tests, a raw score is generally given meaning by comparison with some reference group or groups. This method of giving a raw score meaning is called norm-referenced interpretation. The comparison may be with

1. A series of grade groups (grade norms)
2. A series of age groups (age norms)
3. A single group, in which performance is indicated by what percentage of that group the score surpassed (percentile norms)
4. A single group, in which performance is indicated by the number of standard deviations the score is above or below the group mean (standard score norms). (Norms of this type may be subjected to a linear conversion to eliminate decimal points and negative values or to nonlinear transformations to normalize the score distribution.)

Each alternative has certain advantages and certain limitations.

Quotients such as the IQ were developed to get a single index to express the degree to which individuals deviated from their age group. Because of their various limitations, quotients have been replaced by standard scores, and the term *IQ* is no longer technically appropriate.

If the norms available for a number of different tests are of the same kind and are based on comparable groups, all the tests can be expressed in comparable terms. The results can then be shown pictorially in the form of a profile. Profiles emphasize score differences within the individual. When profiles are used, we must take care not to overinterpret their minor ups and downs.

Norms represent a descriptive framework for interpreting the score of an individual, a class group, or some larger aggregation. However, before a judgment can be made on whether an individual or group is doing well or poorly, allowance must be made for ability level, cultural background, and curricular emphases. The norm is merely an average and not a bed of Procrustes into which everyone can be forced to fit. It describes the person's current performance, relative to some specified comparison group.

An alternative to traditional methods of giving meaning to test scores is provided by item response theory and computer adaptive testing. IRT determines the difficulty level of each item and places the items on a continuum of difficulty. Examinees are placed on the same continuum in terms of their ability level. CAT is then used to select the most appropriate next item for an examinee, based on his or her pattern of past successes and failures.

QUESTIONS AND EXERCISES

1. Why does the frame of reference used to interpret a test score make a difference?
2. Can the same test be interpreted in both a criterion-referenced and a norm-referenced manner? If so, how would the two interpretations differ?
3. A pupil in the sixth grade received a raw score of 25 on the Level 12 Reading Test (Form J) of the Iowa Tests of Basic Skills. What additional information would be needed to interpret this score?
4. Why do standardized tests designed for use with high school students almost never use age or grade norms?
5. What limitations would national norms have for use by a county school system in rural West Virginia? What might the local school system do about the limitations?
6. What assumptions lie behind developing and using age norms? Grade norms? Normalized standard scores?
7. In Figure 3–3, why are the standard scores evenly spaced, while the percentile scores are unevenly spaced?
8. State A gives a battery of achievement tests each May in the 4th, 8th, and 11th grades. The median grade level in each subject in each district

in the state is reported to the state board of education. Should these results be reported? If so, what else should be included in the report? In what ways might the board use the results to promote better education? What uses should the board avoid?

9. Ms. P takes pride in the fact that each year she has gotten at least 85% of her fourth-grade class "up to the norm" in each subject. How desirable is this as an educational objective? What limitations or dangers do you see in it?
10. School F has a policy of assigning transfer students to a grade on the basis of their average grade equivalent on an achievement battery. Thus, a boy with an average grade equivalent of 5.3 would be assigned to the fifth grade, no matter what his age or his grade in his previous school. What are the values and limitations of such a practice?
11. The superintendent of schools in Riverview, Iowa, noted that Springdale Elementary School fell consistently about a half grade below national norms on an achievement battery. He was distressed because this performance was the lowest of any school in the city. How justified is his dissatisfaction? Do you need other information to answer this question? If so, what?
12. The board of education in East Centerville noted that the fourth and fifth grades in their community fell substantially below national norms in mathematics, although they scored at or above average in all other subjects. They propose to study this situation further. What additional information do they need?
13. The third-grade teachers in Bigcity school district have prepared a 30-item test to assess mastery of the basic multiplication facts. What score should they accept as demonstrating "mastery" of these facts? How should such a score be determined?
14. What are the advantages of reporting test performance in terms of stanines? In terms of normal curve equivalents? What problems arise from using each of these forms of normative report?
15. Obtain the manual for some test, and study the information given about norms.
    a. How adequate is the norming population? Is sufficient information given for you to make a judgment?
    b. Calculate the chance score (i.e., the score to be expected from blind guessing) for the test, and note its grade equivalent. What limitations does this suggest for using the test?
    c. What limitations are there on the usefulness of the test at the upper end of its range?
    d. How many raw score points correspond to 1 full year on the grade-equivalent scale? Is this number of points of score the same throughout the range of the test?
16. Have you ever taken a computer adaptive test? If you have, how did you find the experience different from that of taking an ordinary paper-and-pencil test? If you have not taken a test using CAT, can you think of ways this testing format might affect your performance?

SUGGESTED READINGS

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp. 508–600). Washington, DC: American Council on Education.

Cizek, G. J., & Bunch, M. B. (2006). Standard setting: A guide to establishing and evaluating performance standards on tests. Thousand Oaks, CA: Sage.

Cohen, A. S., & Wollack, J. A. (2006). Test administration, security, scoring and reporting. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 355–386). Westport, CT: Praeger.

Flynn, J. R. (1998). WAIS-III and WISC-III gains in the United States from 1972 to 1995: How to compensate for obsolete norms. *Perceptual & Motor Skills, 86,* 1231–1239.

Hambleton, R. K., & Pitoniak, M. J. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 433–470). Westport, CT: Praeger.

Holland, P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 187–220). Westport, CT: Praeger.

Holland, P. W., & Rubin, D. B. (Eds.). (1982). *Test equating*. New York: Academic Press.

Hoover, H. D. (1984). The most appropriate scores for measuring educational development in the elementary schools: GE's. *Educational Measurement: Issues and Practices, 3,* 8–14.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 485–514). New York: Macmillan.

Kolen, M. J. (1988). Defining score scales in relation to measurement error. *Journal of Educational Measurement, 25,* 97–110.

Kolen, M. J. (2006). Scaling and norming. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 155–186). Westport, CT: Praeger.

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Michell, J. (1986). Measurement scales and statistics: A clash of paradigms. *Psychological Bulletin, 3,* 398–407.

Nitko, A. J. (1984). Defining "criterion-referenced test." In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 8–28). Baltimore: Johns Hopkins University Press.

Petersen, N. S., Kolen, M. J., & Hoover, H. D. (1989). Scaling, norming, and equating. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 221–262). New York: Macmillan.

Shepard, L. A. (1984). Setting performance standards. In R. A. Berk (Ed.), *A guide to criterion-referenced test construction* (pp. 169–198). Baltimore: Johns Hopkins University Press.

Thorndike, R. L. (1982). *Applied psychometrics*. Boston: Houghton Mifflin.

Thorndike, R. M. (1999a). IRT and intelligence testing: Past, present, and future. In S. E. Embretson and S. L. Hershberger (Eds.), *The new rules of measurement: What every psychologist and educator should know* (pp. 17–35). Mahwah, NJ: Erlbaum.

Wainer, H. (1990). *Computer adaptive testing: A primer*. Mahwah, NJ: Erlbaum.

Yen, W. M. (1986). The choice of scale for educational measurement: An IRT perspective. *Journal of Educational Measurement, 23,* 299–325.

Yen, W. M., & Fitzpatrick, A. R. (2006). Item response theory. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 111–154). Westport, CT: Praeger.