6

Understanding and Using Numerical Data

Even if you could avoid this chapter and its emphasis on data, numbers, and "dreaded statistics," you wouldn't want to. This is because it is essential to have a good understanding of the meaning of data and how data can be used to improve instruction. This chapter examines numerical data—scores that are obtained when student performance is measured. Here are six reasons why an understanding of numerical data is essential:

1. Numerical data can be used to efficiently summarize and describe a large number of scores.
2. Numerical data are important in determining student grades.
3. Numerical data are used in reporting standardized test scores; an accurate interpretation and the valid use of these scores depend on understanding the data.
4. Numerical data are used extensively in describing validity, reliability, norms, item statistics, and other characteristics of tests and surveys; adequate analysis and interpretation of the quality of tests and surveys depend on understanding numerical data.
5. With an increasing emphasis on school accountability, numerical data are used extensively in school report cards and other reports; teachers and administrators need to be able to understand and evaluate these data.
6. Most research uses numerical data; an understanding of the nature of data helps interpret and use research findings on topics of interest.

I will begin the discussion with the presentation of the most fundamental and important way data are summarized—looking at different types of scores.

TYPES OF SCORES

Several types of scores are commonly used in measurement. The *raw score* typically indicates the actual number of items a student has answered correctly. If 15 of 20 items are correct, then the raw score is 15. This is based on the simple frequency of items. A *frequency* is a count of items or of students. Frequency can refer to the number of students obtaining a specific score or range of scores. For example, if 10 students obtained a score of 75 and 10 students scored 78, the frequency of scores from 75 to 78 is 20.

Percentage indicates the number of items or students per hundred. Thus, if there are 25 items on a test and a student answered 20 of the items correctly, the student answered 80 percent of the items correctly. Similarly, if 15 students in a class of 30 students fail a test, 50 percent of the students in the class fail. Percentage is calculated with division and multiplication. Simply divide the number of items correct or students obtaining a specific score by the total number of items or students, and multiply by 100. For example, if 24 of 62 students obtain a passing score, it can be concluded that 39 percent of the students passed ($24/62 \times 100$). Closely related to percentage is *proportion*, which expresses the result as part of one. Thus, in the previous example, the proportion of students passing is .39.

A *percentile* score, or rank, is a measure of relative standing that indicates the percentage of scores in a distribution that are at or below a specified score. Hence, a score with a percentile rank of 70 is higher than 70 percent of the scores in the distribution. A percentile score is based on the number of items answered correctly, but it does not indicate the percentage of items answered correctly. In other words, a percentile score indicates the percentage of other scores that a student outscored. As we will see in Chapter 8, percentile scores are essential for interpreting norm-referenced tests.

Ranked (or rank-ordered) scores are those that are presented in order of magnitude (size) or frequency of scores. This indicates the relative position of each score or student. Consider the ranking of the following five scores from five students.

Score	*Rank*
90	1
85	2
83	3
81	4
70	5

Tied scores would yield the same rank for each score, determined by the average of the scores. Although ranking scores indicates relative position, it is a crude index of best to worst because the magnitude of the difference between the scores is not indicated. Ranked scores can under- or overestimate the degree of difference between scores. For example, if the difference between the grade point averages is small, such as a hundredth of a point separating the valedictorian and the other top four students in the class, it is reasonable to conclude that the difference between the students is, in a practical sense, meaningless. If the valedictorian scored a full three tenths higher than any other student, however, then the difference is significant. In both cases, a mere rank ordering, without consideration of the degree of difference, would have suggested that the differences between the rankings were the same.

Classroom assessments typically use raw score totals as points or percentages to grade students, give feedback, and inform parents of academic progress (more on how this is done in Chapter 8). Traditionally, standardized tests have used percentile and other derived scores. Standards-based tests may provide both kinds of scores. For example, a state accountability test may give both the number of items answered correctly and a derived score of, for example, something between 200 and 800. The raw score may be related to proficiency while the derived score is tied to percentile rank. These differences in scores can be tricky, especially when comparing scores from both or in trying to use large-scale standards-based scores to drive instruction that is then measured with classroom assessments. We'll discuss some of these derived scores in this chapter; others are presented in Chapter 7.

FREQUENCY DISTRIBUTIONS

When there are many student scores, it is difficult to understand and interpret the results as a whole without organizing the scores in a meaningful way. The *frequency distribution* is the most fundamental approach to organizing a set of data. This type of distribution simply indicates the number of students who obtained different scores. In a *simple frequency distribution,* the scores obtained are rank ordered from highest to lowest, and the number of students who obtained each score is tallied. Figure 6.1 shows how a group of scores can be represented by a simple frequency distribution. If there is a large number of scores or students, it may be best to use a *grouped frequency distribution.* In this type of distribution, score intervals are created, and the number of students whose scores are within each interval is indicated. This type of frequency distribution is also illustrated in Figure 6.1.

One disadvantage of grouped frequency distributions is that information about individual students may be lost. This problem is often encountered in summarizing a large number of scores. Although on the one hand, a single index or a few categories provide a more succinct summary, individual data are embedded within the group. To construct a group frequency distribution, determine the difference between the highest and lowest score and then divide that

		Simple Frequency Distribution		Grouped Frequency Distribution	
Student	*Score*	*Score*	*f*	*Interval*	*f*
Nina	98	98	1	92–98	3
Scott	94	94	2	86–91	3
Therease	94	92	1	80–85	5
Felix	88	88	1	74–79	5
Jim	86	86	2	70–73	4
Lex	86	85	1		
Jon	85	82	1		
Jan	82	80	3		
Hannah	80	79	1		
Karon	80	77	2		
Tyler	80	75	1		
Austin	79	74	1		
Tristen	77	72	2		
Megan	77	71	1		
Janine	75	70	1		
Freya	74				
Rosemary	72				
Frank	72				
Susan	71				
Benjamin	70				

Figure 6.1 Frequency Distributions of Test Scores

number by the number of intervals or categories desired. Usually, this is 5 to 10 intervals, although the actual number is determined somewhat arbitrarily. You will want to have a workable number of intervals and at the same time a sufficient number to reflect the variation in scores. In the end, you want intervals that provide the most accurate summary of the data in a condensed form. If possible, it is best to keep the size of the intervals the same.

DISTRIBUTION SHAPES

The shape of a distribution can tell you a lot about the nature of the scores. When data are presented in the form of a list, such as in Figure 6.1, it is not easy to think

of the overall distribution of scores as having a particular shape. To better understand shape, the data can be presented on a two-dimensional graph. The scores are placed, in ascending order from lowest to highest, on the horizontal part of the graph (x-axis), and values for the frequency with which each score was obtained are placed on the vertical part of the graph (y-axis). A *frequency polygon* is created when a line is drawn to connect the frequencies of each score.

With a large number of scores, the shape becomes smoother and is often referred to as a "curve." Different types of curves provide generic information about the nature of the distribution. The most commonly used shape is the *normal curve* as shown in Figure 6.2. The normal curve represents a symmetrical, bell-shaped distribution. The normal curve is important for two reasons. First, this distribution is found in nature when most human and other kinds of traits or characteristics are measured, such as height, weight, size, temperature, wind velocity, intelligence, intensity, athletic prowess, and so on. Second, properties of the normal curve are used extensively in large-scale, standardized testing.

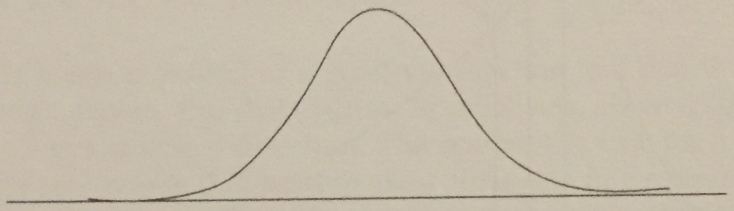


Figure 6.2 The Normal Distribution

In a normal distribution, most of the scores cluster around the middle, with few at the highest and lowest ends. Because the normal curve is symmetrical, one side is a mirror image of the other side. That is, if the normal curve is divided down the middle, the halves are the same; one is a mirror image of the other. Being bell-shaped is also important. A symmetrical curve can be shaped in various ways, only one of which is bell-shaped as apparent in Figure 6.3.

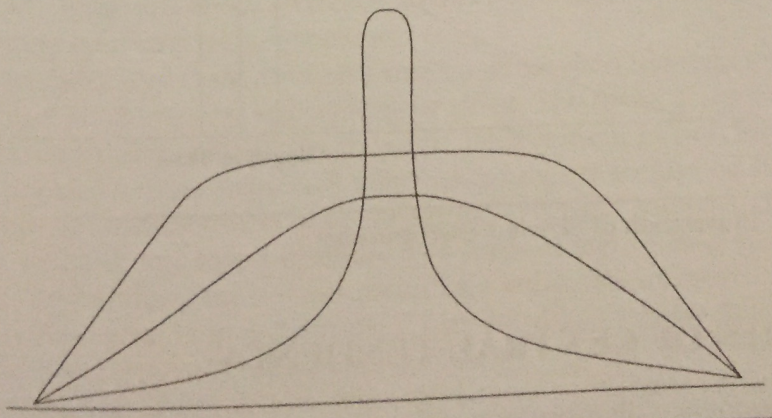


Figure 6.3 Symmetrical Distributions That Are Not Bell-Shaped

If a distribution is not symmetrical, then it may be characterized as *positively skewed*, *negatively skewed*, illustrated in Figure 6.4, or *flat*. In a positively skewed, or "skewed to the right" distribution, most of the scores pile up at the lower end, and there are just a few very high scores. This forms a tail that points in a positive direction. Conversely, when there are mostly high scores with just a few low scores, the distribution is negatively skewed, or "skewed to the left" (tail points in a negative direction). In a flat, or rectangular, distribution, most of the scores have about the same frequency. Negatively skewed distributions are common in classroom assessments. In these tests, often most students do well, while just a few students do poorly. (Note: Positive and negative skew does not mean good and bad!)

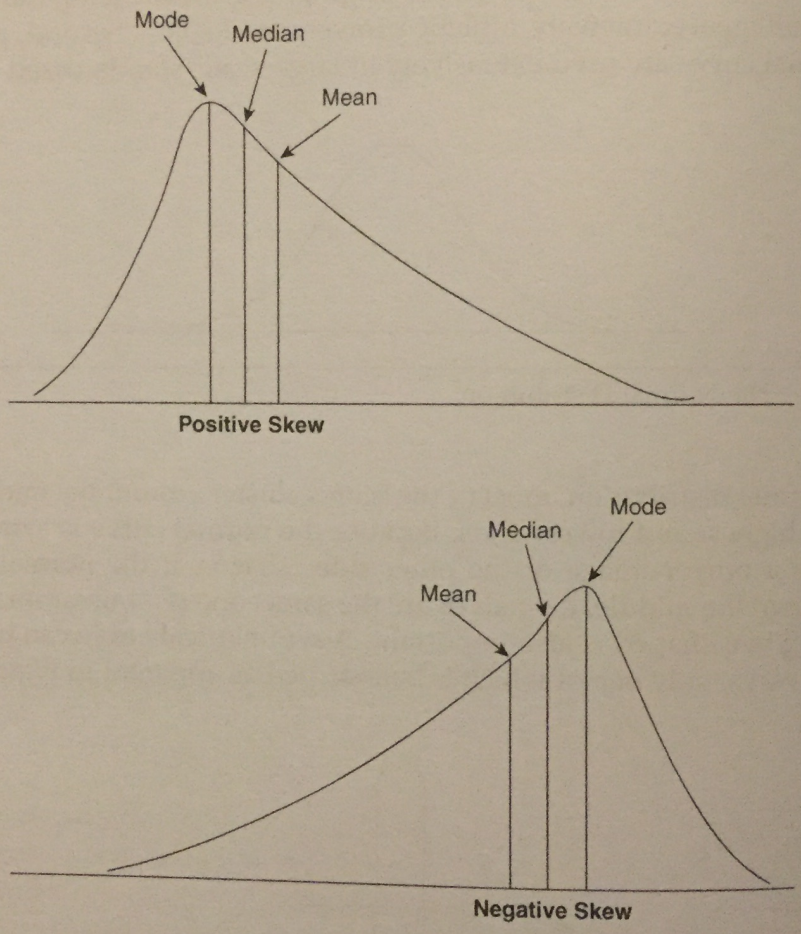


Figure 6.4 Illustration of Skewed Distributions

MEASURES OF CENTRAL TENDENCY

Although frequency distributions and curves can show how scores are distributed, there is usually a need for even more succinct indexes that can capture the

essence of the distribution. These indexes are called *measures of central tendency* because they are used to indicate, with a single number, the most typical or "average" score.

Mean

The *mean* is the arithmetic average. It is calculated by adding all the scores in the distribution and then dividing that sum by the number of scores. The mean is represented by $\overline{X}$ or M. For the set of scores in Figure 6.1, the mean is 81.

One drawback of the mean as a measure of the typical score is that it may be distorted with extremely high or low scores. Consider the following distributions and means. The distributions as a whole are the same with the exceptions of two scores. Yet this small difference results in vastly different means.

1. 1, 2, 2, 3, 4, 4, 5, 5, 5, 6, 6, 8, 8, 8, 10 mean = 5.13
2. 1, 2, 2, 3, 4, 4, 5, 5, 5, 6, 6, 8, 8, 80, 100 mean = 15.93

Because the mean is pulled in a positive direction in a distribution with a few atypical high scores, the distribution is positively skewed. It is as if the mean is skewed in a positive direction. The opposite is true for a distribution with a few very low scores that tend to skew the mean in a negative direction. This is what happens to students when a zero is averaged in as one of the scores in calculating interim or semester grades. The zero may distort the mean so much that it no longer indicates the typical performance of the student.

Another use of the mean is to examine the overall performance of a class or school. Typically, a class or school average is computed and compared with other classes or schools and/or performance in earlier years. Important issues are involved when using the data in these ways. First, just as an extreme score can significantly affect the average for individual students, it also can affect the average for a class or school. This means that a few very low scores can skew the average for the entire class. Consequently, when analyzing scores for the class as a whole, it is prudent to examine the frequency distribution along with the mean. This will show extreme scores and clusters of scores that will help make interpretations more accurate.

Second, teachers find class averages most useful when the data are reported for subgroups of students and knowledge or skills. The mean score for students in a class who have demonstrated very high achievement in previous quizzes, homework, in-class worksheets, and other assignments should be higher than the mean score of students who have struggled with the content. This type of report helps validate the interpretation that high scores indeed suggest high achievement. By breaking an overall mean into subscales or parts, the teacher is able to use these results in a diagnostic way to identify areas that may need further instruction or even remediation.

Third, comparing mean scores with the scores of students from previous years, or with students from other schools, requires care because although the scores are based on the same test, there are invariably differences in other factors

that affect the average score, such as ability levels of students, motivation, and administration procedures. For example, in one school, all students may take the test, whereas in another school, students whose primary language is not English are exempted. For a variety of reasons, students as a group change from year to year, so even longitudinal data from the same school can be misleading if these differences are not taken into account in interpreting the averages.

Median and Mode

The *median* is the midpoint or middle point of the distribution. In other words, the median is the value of the score that has 50 percent of the scores below it and 50 percent of the scores above it. Thus, the median is the score that is at the 50th percentile. The median is found by rank ordering every score in the distribution, including each score that is the same, and locating the score that has half of the scores above it and half below it. For the distribution in Figure 6.1, the median is 80 (in distributions that have an even number of scores, the median is the sum of the two middle scores divided by 2). The median is not distorted by extreme low or high scores and is a better indication of typical score in skewed distributions than the mean.

The *mode* is simply the score that occurs most frequently. In the distribution in Figure 6.1, more students scored an 80 (three) than any other score, so the mode is 80. In some distributions, there can be more than a single mode. For example, if two scores occur the same and both are the highest frequency scores, the distribution is *bimodal*.

In a normal distribution, the mean, median, and mode are the same. In a positively skewed distribution, the mean is greater than the median and mode, whereas in a negatively skewed distribution, the mean is less than the median and mode. The relationship between the three measures of central tendency is illustrated in skewed distributions in Figure 6.4.

MEASURES OF DISPERSION

Although a measure of central tendency is a good indicator of the most typical score in a particular group, it is also useful to know something about how much the scores cluster around the mean or median. Statistics that show how much the scores spread out from the mean are called measures of *dispersion* or measures of *variability*. If the scores are highly dispersed, different, scattered, spread, or dissimilar, then the distribution is characterized as having high *variability* or *variance*. That is, the scores vary considerably. If the scores are bunched together close to the mean, then there is little dispersion and low variability or small variance.

The need for a measure of dispersion to describe a distribution is illustrated in Figure 6.5. These two distributions have the same mean, median, and mode but portray different groups of scores. A complete description is possible only if a measure of dispersion is included.

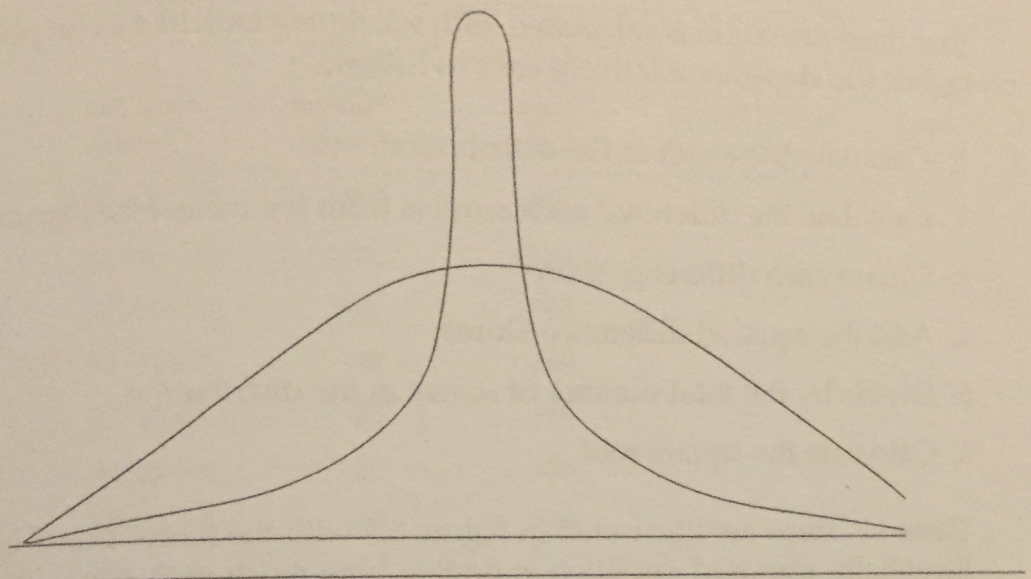


Figure 6.5 Distributions With Different Score Dispersion

Although it is sometimes helpful to use general terms such as *small, large, great, little,* and *high* to describe the amount of dispersion in the distribution, three measures are typically used to more specifically indicate variance: range, interquartile range, and standard deviation.

Range and Interquartile Range

The *range* is simply the numerical difference between the highest and lowest scores in the distribution. It is calculated by subtracting the lowest score from the highest score. The range is a crude measure of dispersion because it is based on only two scores from the distribution, and it does not indicate anything about relative cluster of scores. In a highly skewed distribution, the range is particularly misleading, suggesting a higher degree of dispersion than actually exists.

The *interquartile range* indicates the middle 50 percent of the scores in the distribution. By limiting the measure to the middle 50 percent of the scores, the majority of the scores are included in the calculation of dispersion, but extreme high or low scores that would influence the range are excluded. The interquartile range is determined by subtracting the score at the 25th percentile from the score at the 75th percentile. In the distribution in Figure 6.1, the interquartile range is 13 (87–74).

Standard Deviation

A more complicated but informative and precise measure of dispersion is *standard deviation,* a number that indicates the "average" distance of the scores from the mean. A distribution that has scores that are bunched together close to the mean will have a small standard deviation, whereas distributions with scores spread way out from the mean will have a large standard deviation.

Standard deviation is calculated with what may look like a complicated formula, but the steps are relatively easy to follow:

1. Calculate the mean of the distribution.
2. Calculate the difference each score is from the mean (see Figure 6.6).
3. Square each difference score.
4. Add the squared difference scores.
5. Divide by the total number of scores in the distribution.
6. Calculate the square root.

These six steps are illustrated in Figure 6.7 with the scores from Figure 6.1.

Essentially, standard deviation is finding how much each score differs from the mean and then finding the average difference score, or, in other words, finding the average distance of the scores from the mean. Simply calculate the squared deviation scores, find the average deviation score, and then take the square root to return to the original unit of measurement. The most common convention in reporting and using standard deviation is to indicate that one standard deviation is equal to some number (e.g., 1 *SD* = 5).

In a normal distribution, there are certain properties to standard deviation that are universal and that help in understanding the scores. The meaning of how scores are related when we say "one standard deviation" is always the same in a normal distribution, regardless of the unit of standard deviation. For instance, a score in a normal distribution that is at +1 standard deviation will be at the 84th percentile. This is true for any normal distribution. If one distribution has a mean of 40 and a standard deviation of 5, a score of 45 is at the same percentile as a score of 6.5 in a distribution that has a mean of 6 and a standard deviation of 0.5.

Because the normal curve is symmetrical, it can describe the approximate percentage of scores that are contained within given units of standard deviation. This is illustrated in Figure 6.8, where 1 *SD* = 5. On both sides of the mean (15), there is a line that designates –1 and +1 *SD*. The negative and positive

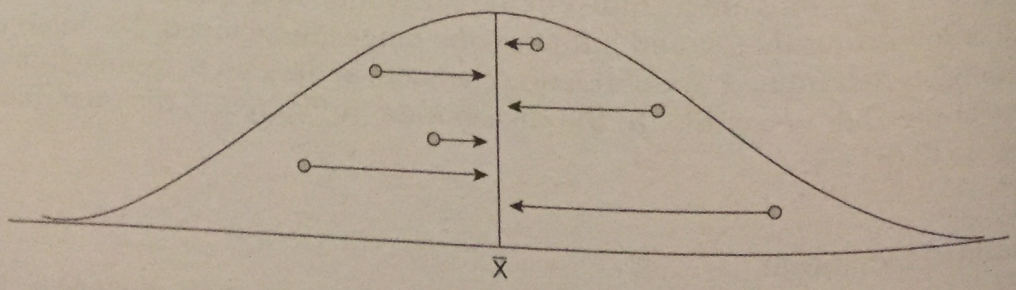


Figure 6.6 Illustration of Distance of Each Score From the Mean

Source: From James H. McMillan & Sally Schumacher. *Research in Education: A Conceptual Introduction*, 6/e. Published by Allyn and Bacon, Boston, MA. Copyright © 2006 by Pearson Education. Reprinted by permission of the publisher.

(1) Calculate the Mean	(2) Deviation Scores	(3) Deviation Scores Squared	(4) Scores Added	(5) Added Scores Divided by N	(6) Square Root
98	98–81 = 17	289	+289		
94	94–81 = 13	169	+169		
94	94–81 = 13	169	+169		
88	88–81 = 7	49	+49		
86	86–81 = 5	25	+25		
86	86–81 = 5	25	+25		
85	85–81 = 4	16	+16		
82	82–81 = 1	1	+1		
80	80–81 = –1	1	+1		
80	80–81 = –1	1	+1		
80	80–81 = –1	1	+1		
79	79–81 = –2	4	+4		
77	77–81 = –4	16	+16		
77	77–81 = –4	16	+16		
75	75–81 = –6	36	+36		
74	74–81 = –7	49	+49		
72	72–81 = –9	81	+81		
72	72–81 = –9	81	+81		
71	71–81 = –10	100	+100		
70	70–81 = –11	121	+121		
1620/20 = 81			1250	1250/20 = 62.5	$\sqrt{62.5} = 7.9$

Figure 6.7 Steps in Calculating Standard Deviation

directions from the mean are equivalent in score units. That is, both –1 and +1 *SD* are 5 score units. Between –1 and +1 *SD* is about 68% of the total number of scores in the distribution. This is determined by knowing that if the mean is the 50th percentile, which it is for a normal distribution, and +1 *SD* is at the 84th percentile, then subtracting 50 from 84 shows that 34 percent of the scores in the distribution must be between the mean and +1 *SD*. Because the distribution is symmetrical, the same is true for the percentage of scores between the mean and –1 *SD* (34 percent). Thus, adding 34 percent and 34 percent yields a total of 68 percent of the scores of the distribution within one standard deviation of the mean.

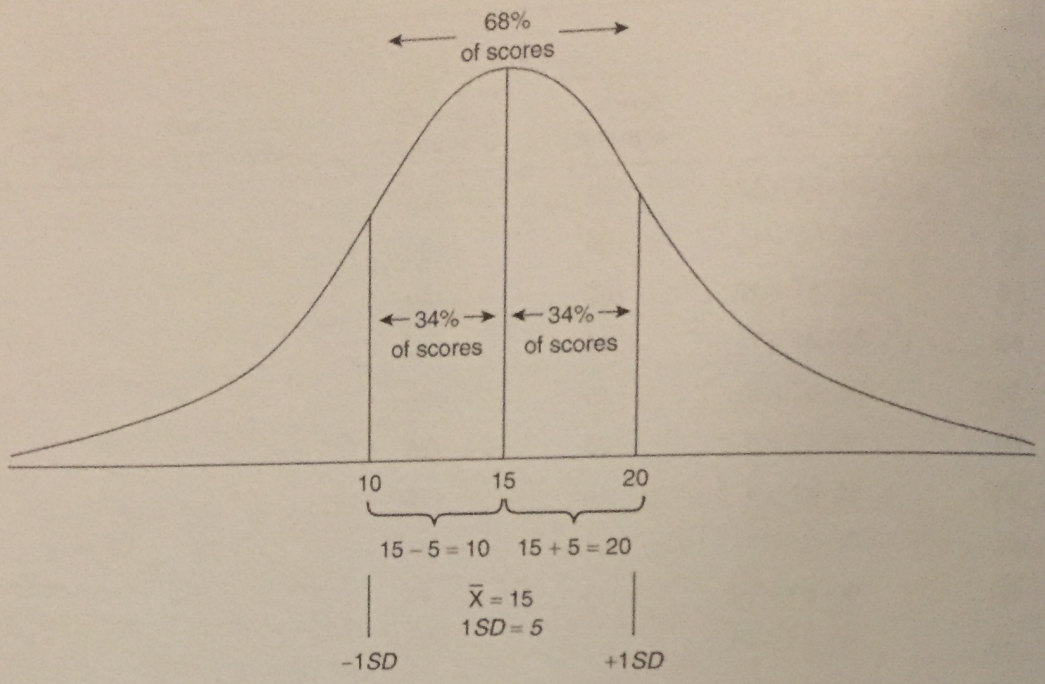


Figure 6.8

Source: Adapted from McMillan & Schumacher (2006).

Figure 6.9 shows a more complete description of the normal curve and units of standard deviation. As long as the distribution is normal, +2 *SD* will be at the 98th percentile, and –2 *SD* will be at the 2nd percentile. In other words, if a student's score is at two standard deviations above the mean, the student did better than 98 percent of the other scores in the distribution. In a normal distribution, 96 percent of the scores are between +2 and –2 *SD*.

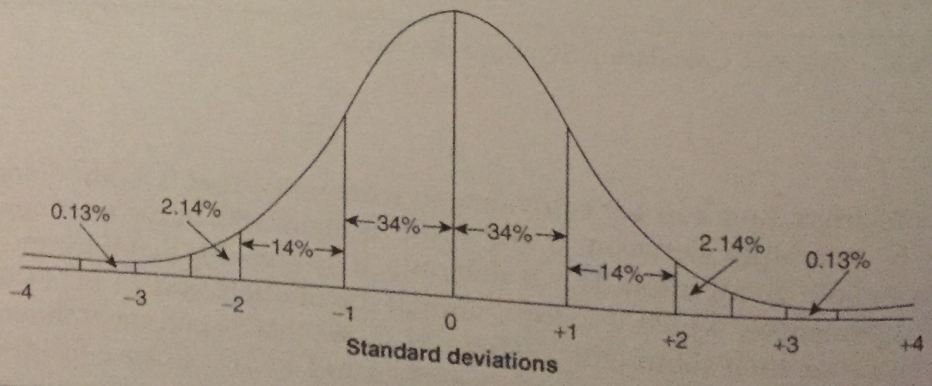


Figure 6.9 Standard Deviations in the Normal Curve

MEASURES OF RELATIONSHIP

Measures of relationship are used to indicate the degree to which two scores from different measures are related. That is, are scores from one assessment associated with, or predictable from, scores from another assessment? Suppose Ms. Lopez has a small class of students and is interested in the relationship between her classroom semester geometry test and the scores from a large-scale geometry test of the same content. Both tests are taken by her students. The scores for the students are summarized in Figure 6.10. When these scores are rank-ordered and compared it is evident that there is some degree of relationship because students who obtained a high score on the classroom assessment also obtained a high score on the large-scale test. In other words, in a rough sense, scores from one test can be used to predict approximate scores on the other test. If there was no relationship, there would be no pattern or prediction. Students with high scores on the semester test might have low large-scale test scores, and those obtaining low classroom test scores might have high standardized scores.

The kind of description of relationship shown by these two sets of scores is conceptual. Generally, a more specific and precise measure of relationship is used, known as the correlation coefficient.

Student	*Semester Exam Score*	*Standardized Test Score*	*Semester Exam Rank*	*Standardized Test Rank*
1. Molly	97	93	1	3
2. Ted	95	95	2	1
3. Dan	94	94	3	2
4. Cheryl	90	89	4	6
5. Ryann	88	91	5	4
6. Jon	86	90	6	5
7. Hannah	85	84	7	8
8. Tyron	82	85	8	7
9. Jim	79	82	9	10
10. Jan	78	79	10	12
11. Bill	77	83	11	9
12. Maria	75	77	12	13
13. Frank	72	71	13	15
14. Carl	71	80	14	11
15. Tom	68	73	15	14

Figure 6.10 Classroom Exam Scores and Standardized Test Scores

Correlation Coefficients

A correlation measures the degree to which the scores of two or more variables or factors are related. The *correlation coefficient* (r) is a number between –1 and +1 that is calculated to indicate the strength and direction of the relationship. A correlation coefficient is calculated by a formula and is reported as r = .76, r = –.35, r = .04, and so on (notice that there is a minus sign before a negative correlation but no plus sign in front of a positive correlation). Although there are a number of types of correlation coefficients, the one encountered most in assessment is called the Pearson product-moment correlation. A positive correlation means that as the value of one variable increases, so does the value of the other variable. This is also called a *direct* relationship. If the correlation coefficient is between 0 and +1, it is positive. A negative or inverse correlation is indicated by a negative coefficient and indicates that as the value of one variable increases, the value of the other variable decreases. A negative or inverse correlation is represented by a number between 0 and –1.

A positive correlation is not necessarily any better than a negative one. For example, a desirable positive correlation exists between time spent studying and achievement, whereas an undesirable positive correlation would be student anxiety and referrals to the counselor's office. Conversely, there are many helpful negative relationships, such as those between student attention and teacher rebukes, or time teachers lecture and student attitudes. An undesirable negative correlation would be student disruptions and student achievement.

The coefficient also indicates the strength or magnitude of the relationship, independent of direction. Strength refers to the degree of the relationship, that is, how powerful or helpful it is in predicting one variable from another. A high positive value (e.g., r = .93, r = .85, or r = .88) represents a high or strong positive relationship (+1 is a perfect relationship). The same is true for a high negative value (r = –.93, r = –.85, or r = –.88). Low values, those close to zero, indicate a weak or small relationship, whereas values midway between 0 and +1 or –1 indicate moderate relationships (e.g., r = –.45, r = .62). Thus, the strength of the relationship becomes stronger as the correlation coefficient approaches either +1 or –1 from 0. This is illustrated in Figure 6.11.

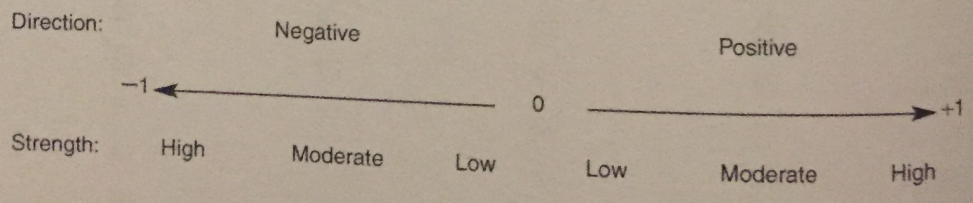


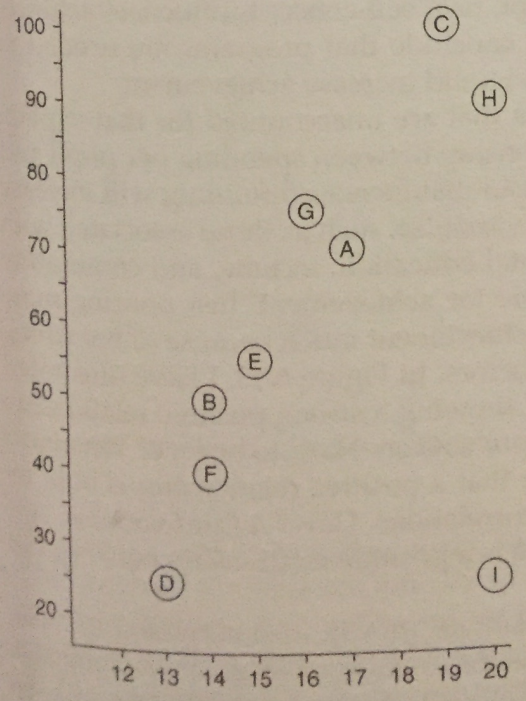
Figure 6.11 Correlation Strength and Direction

Scatterplots

Although the correlation coefficient is used extensively to report relationships, the scatterplot or scatter diagram is needed to interpret the coefficient

correctly. The scatterplot is a graphic representation of the relationship. It is formed by making a visual array of the intersection of each student's score on the two variables or measures. The two-dimensional graph is made by rank ordering the values of one variable on the horizontal axis, from low to high, and rank ordering the values of the second variable on the vertical axis. For example, to correlate student achievement scores with self-concept scores, the range of results for the achievement scores could be placed on the horizontal axis and range of self-concept results on the vertical axis. This relationship is illustrated in Figure 6.12 with a few students. The student scores are summarized next to the graph in random order. The intersections of each set of scores are indicated in the graph with the letters. Together, these intersection points form a pattern that provides a general indication of the relationship. Figure 6.12 shows a positive relationship. As achievement scores increase, self-concept scores also increase.

The scatterplot is helpful in identifying two aspects of correlation that affect the interpretation of the coefficient. The first is to determine if there are any atypical scores as related to the overall pattern. In the illustration in Figure 6.12, for instance, the intersection of Helen's scores is not at all consistent with the pattern. This atypical score, or *outlier*, lowers the coefficient to give the impression of less relationship than actually exists. It is similar to the effect that an extreme score has on the mean. In this case, there is a *spurious* correlation, rather than a *skewed* distribution.



Subject	Self-Concept	Achievement
Ryann (A)	17	70
Jesse (B)	14	50
Amanda (C)	19	100
Meghan (D)	13	25
Katie (E)	15	55
Cristina (F)	14	40
Emma (G)	16	75
Jan (H)	20	90
Helen (I)	20	25

Figure 6.12 Scatterplot of Achievement Related to Self-Concept

The general pattern also indicates if the relationship is linear or curvilinear. The Pearson product-moment correlation coefficient is calculated as if the relationship is a linear one. Thus, if the scatterplot identifies a curvilinear pattern, then the coefficient will be lower than the actual relationship.

Interpreting Correlations

Because correlations are used extensively in assessment, it is important to interpret the meaning of correlation coefficients accurately. There are three primary limitations to consider: correlation and causation, restricted range, and the size of coefficients.

Correlation and Causation

It is tempting to think that a correlation describes a cause-and-effect relationship, but that is rarely the case. An accurate interpretation of a correlation *always* begins with the understanding that the relationship is descriptive only of a predictive relationship—that to some degree, the value of one variable or measure can be predicted from knowledge of value for another one. You should not conclude that one variable *caused* the change in the other or was the *reason* that the values of the other measure or variable changed.

Correlation does not imply causation for two reasons. First, a relationship between A and B may be high, but there is no way to know if A caused B or B caused A. For example, consider the relationship between achievement and self-concept illustrated in Figure 6.12. Although it is clearly positive, we don't know if achievement affects self-concept, or if self-concept influences achievement. That is, it would be incorrect to conclude that programs are needed to enhance self-concept, thinking that this would increase achievement.

Second, there may also be variables that are unaccounted for that explain the relationship. Think about the relationship between spending per pupil and achievement. If it is positive, does it mean that increased funding will increase achievement? Perhaps, but many other variables, such as those associated with family background and SES, like parental education, income, and community attitudes, are probably more responsible for achievement. Just pouring more money in the schools would not raise achievement much because of the strong effect of these family and community factors. In Figure 6.13, I have illustrated the principle of additional variables by showing a strong positive relationship between body weight and reading comprehension. Hard to believe? When you follow the steps in Figure 6.13, you see that a positive relationship is built by stringing together a series of near-zero correlations. How? A third variable, age, is related to weight, and obviously there is a positive relationship between age and reading comprehension.

Despite these two limitations, correlations are still misinterpreted to mean something causal—perhaps because it seems so reasonable, given the language that is used. For example, when there is a positive correlation between time on task and achievement, it seems obvious that increasing time on task will increase

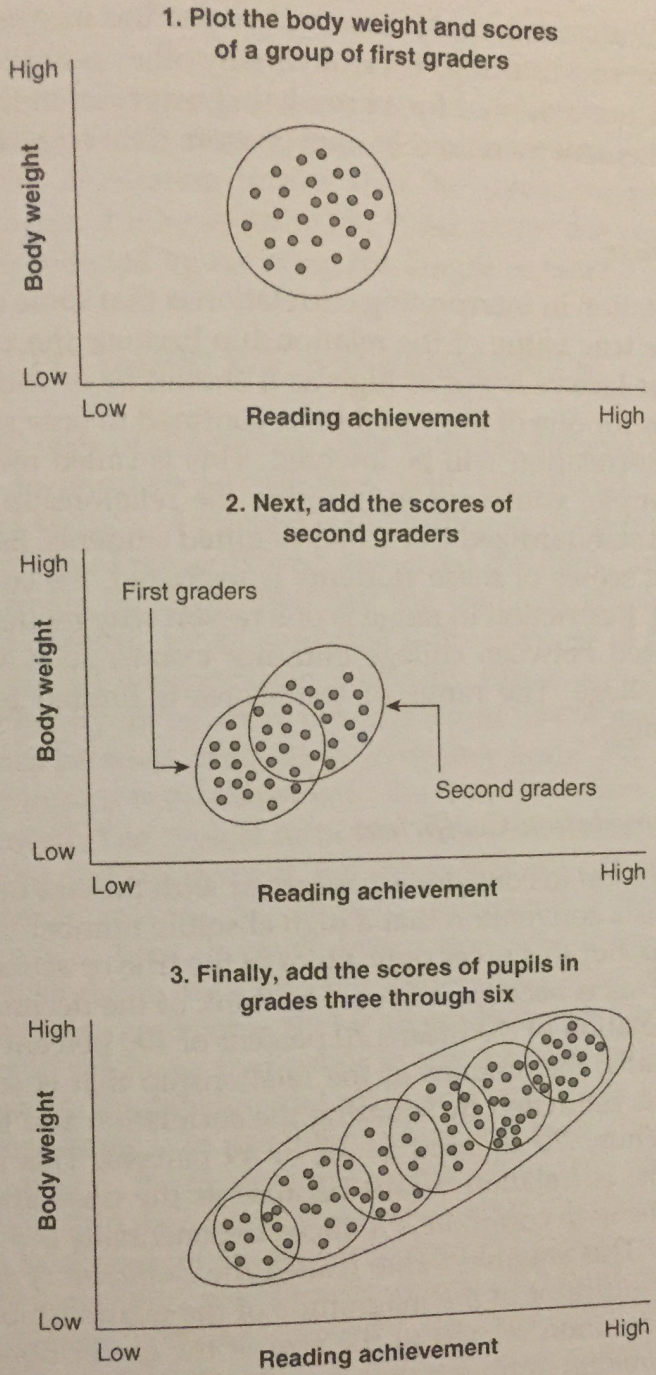


Figure 6.13 Correlation Between Weight and Reading Comprension

achievement, and in fact, this might be true. It simply sounds logical, like the positive correlation between amount of rain and growth of crops. But do not be lured into thinking that the reason or cause of increases in achievement is an increase in time on task. The reality is that we don't know if increased achievement caused students to be on task more, or if other unaccounted-for variables are actually responsible for the increases. Maybe students were given different

incentives along with more study time, or maybe the increased time on task consisted of one-on-one tutoring. There are many other possible causal explanations, which if left unaccounted for, necessitate great caution in concluding that in this case, achievement increased *because* of more time on task.

Restricted Range

A second limitation in interpreting correlation is that some correlations may underestimate the true value of the relationship because the variability of one of the measures or factors is not as high as it should be or could be. That is, if the range of scores for one of the variables is confined to only a part of the total distribution, the correlation will be lowered. This is called *restriction in range.* Suppose, for example, you want to examine the relationship between grade point average and standardized test scores of gifted students. Because the range of grades and test scores of these students is restricted, the correlation would probably be small. Restriction in range is one reason why modest relationships, at best, are reported between college entrance exams, such as the SAT, and achievement in college. The range of test scores is limited to students who scored relatively high.

Size of the Correlation Coefficient

The third limitation to consider is concerned with the size of the correlation coefficient. There is a convention that a high absolute number is described as a strong relationship, but this gives only a hint of the true or actual magnitude of the relationship. This is because it is easy to think of the decimal as a percentage, so that a correlation of .70 means 70 percent of 100 percent of the possible relationship. Actually, the amount of the relationship that is common among the two variables is estimated by squaring the correlation and thinking of that as a percentage. Thus, .70 squared is .49, or 49 percent. This has a dramatic impact on what the correlation means because as the correlation lessens, the amount of relationship in common is reduced exponentially (r = .50, 25 percent; r = .30, 9 percent). This squared value (called the *coefficient of determination*) is the more accurate indicator of the magnitude of the relationship.

Another consideration related to the size of the correlations is that many will be labeled *significant* although they are small and account for a very small amount of common variance (e.g., r = .20, 4%). Researchers use the term *significant* in this context to mean that the correlation is *statistically* different from no relationship at all. It is probable that in studies that have a large number of participants, that a small correlation will be reported as statistically significant. This does not, however, mean that the correlation is strong, high, important, or meaningful.

The importance of the size of correlations is illustrated further in Figure 6.14, which shows scatterplots of four sets of data. The correlation of r = .84 shows a strong positive relationship. This makes sense for the two measures, grade point average (GPA) and SAT score. But if you look at the grade point

averages that predict an SAT score of about 1000, the range of predicted scores is from 2.2 to 3.0. This is not nearly as precise as the "very strong" correlation of $r = .84$ might seem to imply. Examine the remaining three scatterplots. The one of no relationship, $r = .06$, doesn't look too different from moderate positive relationship, $r = .58$, illustrating further that the actual predictive power of a correlation is not what it may seem to be. What about the correlation of –.17. Can SAT score be predicted by knowing the length of hair?

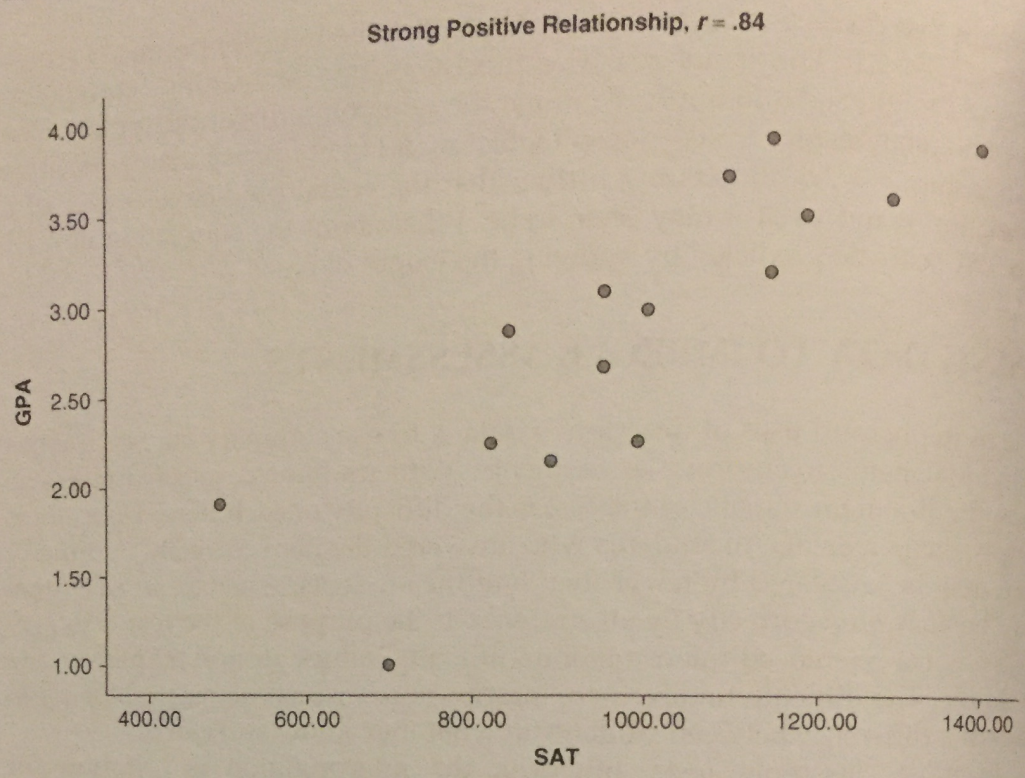
USING DATA TO IMPROVE ASSESSMENTS

One of the helpful uses of descriptive data is to use summary statistics about individual items to improve assessments. With traditional selected-response tests, the first item statistic of interest is the difficulty of each item. *Difficulty* is simply the percentage of students who answered the item correctly. A difficult item may be answered by fewer than half the students, whereas an easy item may be answered correctly by all students. If the purpose of the test is to provide a norm-referenced interpretation, difficulty values around 50 percent are desirable. The difficulty index, then, informs about which items are helping to discern a difference between students in what they know and can do.

In most classroom tests, however, the interpretation is criterion- or standards-referenced. The typical difficulty index for items will usually vary between 60 percent and 100 percent. For items that are answered correctly by only a few students, review the items to determine if the scores are low because the item was poor, because instruction was inappropriate for the items, or because students in fact do not understand what is being assessed. For items that are answered correctly by all students, review the items to see if there are flaws in the items that result in correct answers even by students who do not have the knowledge and skills that are being assessed. Another way to check these items is to use them before and then after instruction in a pretest-posttest design. If students are unable to answer the item correctly prior to instruction, and then know the answer after instruction, then the item is working well to document student learning. For items answered correctly by some students, check to see if the students who answered correctly did well on other, related work and whether students who did not know the answer did poorly on related work. If both of these conditions are present, then the item is a good one.

A second procedure to use with items that are used for norm-referenced interpretations is to calculate the discriminating power of the item. The *discriminating power* of an item refers to the ability of the item to discriminate between students whose total score is high and students whose total score is low. A discrimination index for each item can be calculated by completing five steps:

1. Rank order all students on the basis of total score.
2. Divide the scores into a high and a low group (usually the top quarter or third and bottom quarter or third).



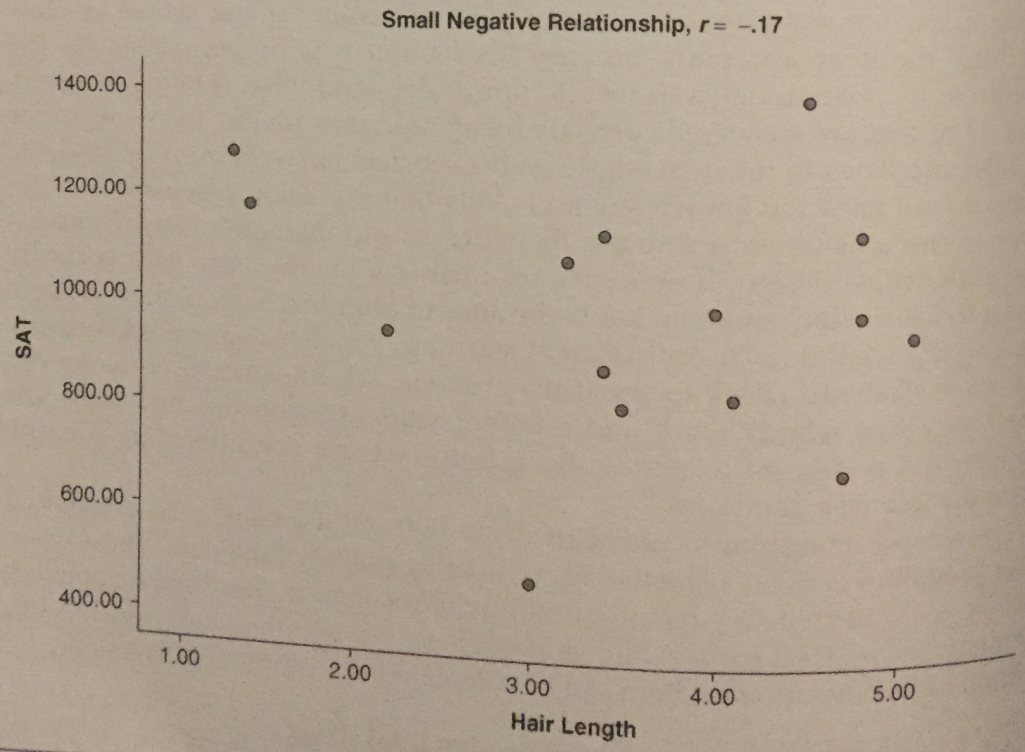
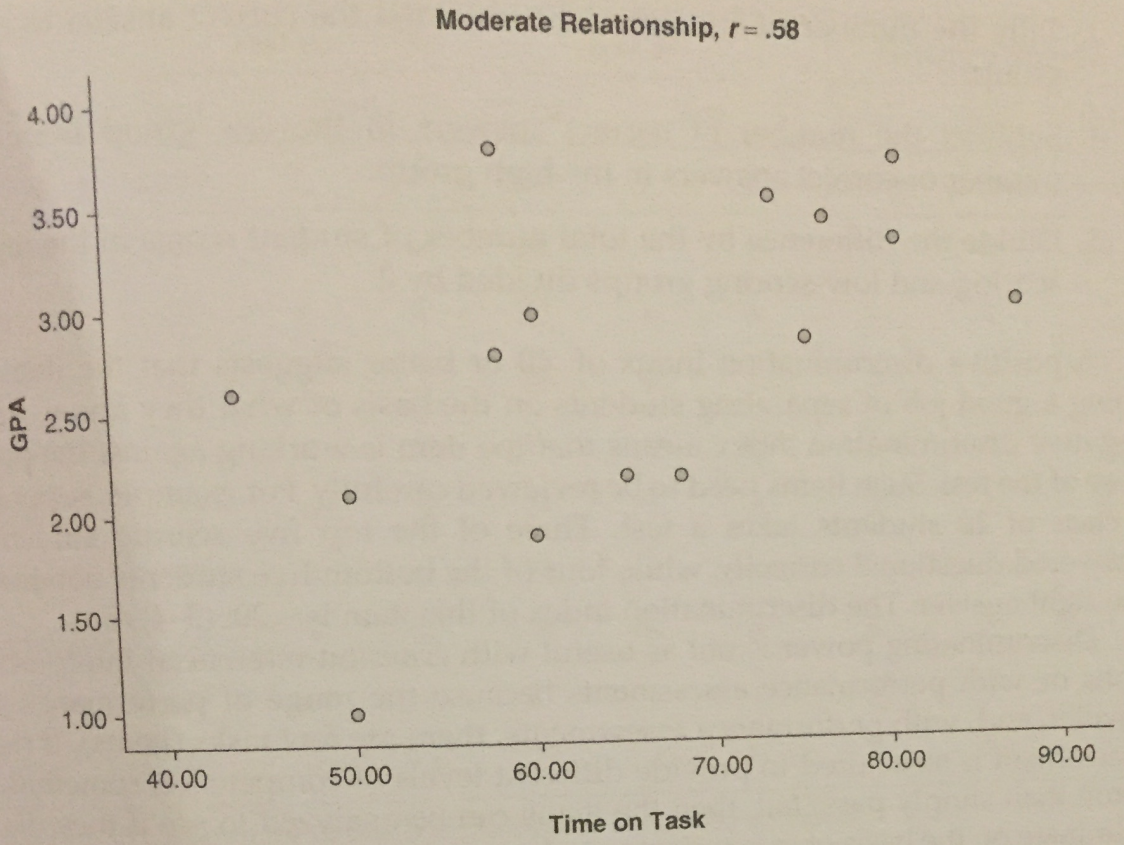


Figure 6.14 Scatterplots of Different Correlations



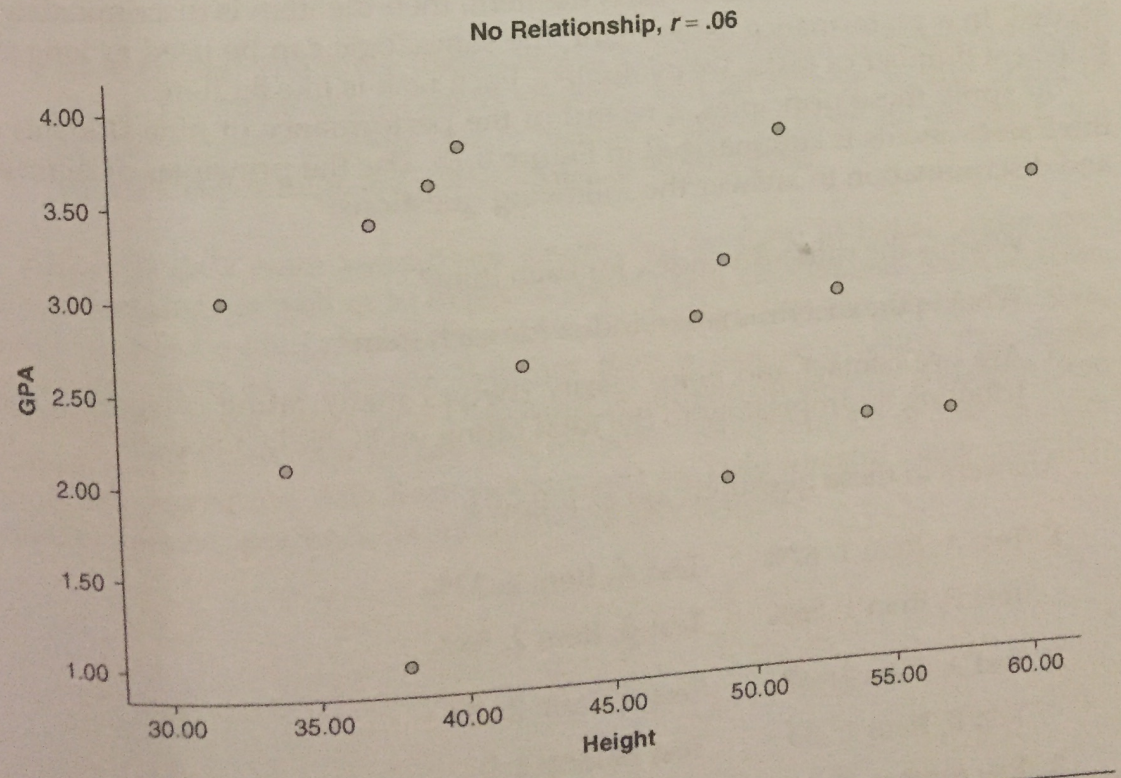


Figure 6.14 *(Continued)*

3. Tally the number of students who obtained the correct answer in each group.
4. Subtract the number of correct answers in the low group from the number of correct answers in the high group.
5. Divide the difference by the total number of student scores in the high-scoring and low-scoring groups divided by 2.

A positive discrimination index of .40 or better suggests that the item is doing a good job of separating students on the basis of what they know. Any negative discrimination index means that the item is working against the purpose of the test. Such items need to be reviewed carefully. For example, suppose a class of 20 students takes a test. Three of the top five scoring students answered question 3 correctly, while four of the bottom five students obtained the right answer. The discrimination index of this item is –.20: (3–4)/5.

Discriminating power is not as useful with criterion-referenced interpretations or with performance assessments because the range of performance is smaller, and, with performance assessments, there are few tasks (items). If the assessment is structured to provide different levels of competence, something more than simply pass/fail, then the items can be analyzed to see if they discriminate on the basis of categorizing students in one of the levels. That is, if all students categorized as "exemplary" answer an item correctly, and students designated as "fail to understand" miss the item, then the item is discriminating as needed. In a performance assessment, the same logic can be used as long as a sufficient number of tasks are evaluated. Each task is like an item.

To apply these principles, a record of the performance of nine students on three assessments is summarized in Figure 6.15. Use the principles of difficulty and discrimination to answer the following questions:

1. What is the difficulty index for each item?
2. What is the discrimination index for each item?
3. Are eye contact and voice clarity (two of many rating categories) contributing appropriately to the total rating on "Giving a Speech"?

Answers to these questions are as follows:

1. Test A, Item 1: 67% Test A, Item 2: 33%
   Test B, Item 1: 56% Test B, Item 2: 56%
2. Test A, Item 1: .67 Test A, Item 2: .67
   Test B, Item 1: .33 Test B, Item 2: 0
3. Eye contact: The three highest-rated students had an eye contact average score of 4.7. The three lowest-rated students had an average eye contact score of 1.7. Yes, eye contact is contributing as appropriate.

	Test A			Test B			Giving a Speech		
	Total	*Item*		*Total*	*Item*		*Total*	*Eye*	*Voice*
Student	*Score*	*1*	*2*	*Score*	*1*	*2*	*Rating*	*Contact*	*Clarity*
George	70	w	w	24	r	r	5	5	2
Ron	60	r	w	30	r	w	5	4	3
Sally	75	r	r	28	w	r	3	2	3
Tamika	83	r	w	19	w	w	4	4	1
Mica	72	w	w	24	r	r	2	1	4
Helen	88	r	r	22	r	r	2	2	3
Rosemary	85	r	w	18	w	w	4	3	5
Ashley	78	w	w	27	w	r	4	2	5
Kenya	92	r	r	29	r	w	5	5	5

Figure 6.15 Summary of Student Scores on Three Assessments

Voice clarity: The three highest-rated students had a voice clarity rating of 3.3. The three lowest-rated students had an average clarity rating of 3.3. Because the average clarity ratings of the students are the same, the voice clarity ratings are not working as appropriate.

Although data from assessment results can be used to better understand student learning, as well as to improve subsequent assessments, such analysis should always be tempered with professional judgment. Descriptive data are useful, but more as a general indicator than a precise measure that drives specific conclusions or practice. Data must be judged on validity, reliability, and fairness, along with other factors that influence interpretation or use. In other words, the descriptive data from assessments are only some of many considerations in making implications and drawing conclusions.