CH8

204

Chapter 8 • Measurement

meAsurement

the process of assigning numbers according to some rule or convention to aspects of people, jobs, job success, or aspects of the staffing system

dAtA

the numerical outcomes of measurement

predictive dAtA

information about measures used to make projections about outcomes

criterion dAtA

information about important outcomes of the staffing process

companies in the United States, states they were able to work with PeopleAnswers to reduce turnover by 50 percent using screening and assessment technologies developed using data ana- lytics. The cornerstone of using data effectively is collecting high-quality information on the right predictors and outcomes.9

What Is MeasureMent?

Evaluating people, processes, or outcomes requires collecting data, which requires measure- ment. In staffing, measurement is the process of assigning numbers according to some rule or convention to aspects of people, jobs, job success, or aspects of the staffing system.10 These measures can be used to select individual employees or to assess the staffing process in gen- eral. Interview ratings, knowledge test scores, and performance data are measures used to select employees. Measures relevant to the staffing process in general assess the following: (a) characteristics of the job (which enables the firm to create job requirements and job rewards matrices), (b) aspects of the staffing system (such as the number of days a job posting is run, where it is run, and the nature of the recruiting message), (c) characteristics of job candidates (such as personality or ability), and (d) staffing outcomes (such as performance or turnover). These types of measures enable the firm to improve its staffing system by identifying impor- tant patterns and relationships that can be used to make decisions or to design assessments and interventions.

The numerical outcomes of measurement are data. Strategic staffing centers around col- lecting and analyzing the relationship between two types of data:11

1. Predictive data is information used to make projections about outcomes. For example, what data might you collect to predict turnover or job performance? Similarly, could you measure the conscientiousness of job candidates to see if it predicts some component of job success? This is predictive data. In terms of the general staffing system, predictive data can come from any part of the hiring process and can include information on sourc- ing quality, the basic qualifications of applicants, and their traits, competencies, and values.

2. criterion data is information about important outcomes of the staffing process. Traditionally, this data includes measuring the job success of employees. More gener- ally, criterion data should also include all outcome information relevant to the evaluation of the effectiveness of the staffing system against its goals. This can include measuring a company’s return on investment related to its staffing measures, employee job success, time-to-hire, time-to-productivity, promotion rates, turnover rates, and new hire fit with company values.

In short, criterion data is information about outcomes of the staffing or selection process. Predictive data gives you information about the possible predictors of those outcomes.

Once you have collected data, you have to do something with it. Next, we discuss basic tools and techniques for describing and interpreting data, followed by a discussion of the charac- teristics of useful measures. It is critical to note that it is nearly pointless to analyze and interpret data that is of low quality. The data must be accurate, dependable, and relevant to be worth col- lecting and analyzing.

DescrIbIng anD InterpretIng Data

When a measure, such as an assessment test, is administered to job candidates, the data needs to be interpreted before it can be useful in making hiring decisions. Describing the scores within a distribution is important for interpreting what they mean for the entire group of candidates, as well as understanding the significance of any one particular score.

types of Measurement

The tools you can use to describe and interpret the data depend on the level of measurement. The data can come from nominal, ordinal, interval, or ratio measures.

noMInal In nominal measurement, numbers are assigned to discrete labels or categories. No ordering is implied in the assigned values. Gender, race, and college major are examples of nominal measures. You could assign a “0” to males and a “1” to females to create a nominal measure for gender.

orDInal In ordinal measurement, attributes are ranked by assigning numbers in ascending or descending order. For example, the first person finishing a work sample task might receive a “1,” the second person “2,” and so on. Ordinal measures don’t tell you anything about the dis- tance between the scores, though—just their rank.

Interval In interval measurement, the distance between scores has meaning. The distance from 40 to 50 degrees in Fahrenheit is the same as the distance from 80 to 90 degrees. Thus, the interval is constant. However, the zero point on interval measures is arbitrary, so ratios computed using two different measures of the same attribute will not yield the same result. For example, 100 degrees Fahrenheit is twice that of 50 degrees Fahrenheit but when converted to Celsius the 2:1 ratio doesn’t hold. Examples of interval measurement in selection may include intelligence scores, personality assessment scores, and scoring keys for interview questions.

ratIo ratio measurement includes a true and meaningful zero point. Thus, you can con- struct ratios from the measure. Salary, weight, height, typing speed, and sales per month are examples of ratio-level measures. If one person can lift 200 pounds and another 100 pounds, then the first person can lift twice as much as the second person whether the weight is in grams or pounds. Thus, the ratio holds because there is a true zero point. In a selection context, years of experience is a ratio measure because ratios will hold whether time is measured in years, min- utes, or seconds.

The distinctions among the different types of measures are important because they influ- ence how you can describe and interpret data. For example, it is generally not useful to compute an average of ordinal scores.

scores

The process of assigning numerical values during measurement is scoring. In order to interpret scores properly, we need to understand the scoring system used. Data is often presented in terms of numerical scores, such as raw scores, standard scores, and percentile scores, which we discuss next.

raW scores raw scores are the unadjusted scores on a measure. On a job knowledge test, the raw score might represent the number of items answered correctly. For measures such as personality inventories that have no “right” or “wrong” answers, the raw score may represent the number of positive responses for a particular trait. Raw scores do not provide much useful infor- mation by themselves. Consider your score on a midterm. If you get 30 out of 50 questions cor- rect, it is hard to know whether this is a good or a poor score. You may believe 30 is a poor score, but if you compare the results to the results of other people who took the same test, you may discover that 30 is the highest score. For criterion-referenced measures, or standards-based assessments, the scores have meaning in and of themselves. For example, candidates might be expected to exceed a certain level on a criterion measure, such as typing at least 90 words per minute, before they can advance to the next stage of the hiring process.

On criterion-referenced measures it is easy to see what a particular score indicates about proficiency or competence. In general, scores on norm-referenced measures have meaning only when compared to the scores of others. For example, candidates who reach a certain norm- referenced measure—for example, who score in the top third of their applicant group on a typing test—would advance to the next stage of the hiring process. Converting raw scores into standard scores (or percentiles), as we describe next, provides you with the kind of comparative informa- tion you need to use a norm-referenced measure.

norMal curve Many human characteristics, such as height, weight, math ability, and typing skill, are distributed in the population in a typical pattern known as the normal curve. In other words, the characteristics display a symmetrical bell-shaped appearance like

Chapter 8 • Measurement 205 nominAl meAsurement

a measurement in which numbers are assigned to discrete labels or categories

ordinAl meAsurement

a measurement in which attributes are ranked by assigning numbers in ascending or descending order

intervAl meAsurement

a measurement in which the distance between scores on an attribute has meaning

rAtio meAsurement

a measurement in which the distance between scores has meaning; it includes a true and meaningful zero point

rAw scores

the unadjusted scores on a measure

criterion-referenced meAsures

measures in which the scores have meaning in and of themselves

norm-referenced meAsures

measures in which the scores have meaning only when compared to the scores of others

normAl curve

a curve representing the bell-shaped symmetrical distribution of some factor



206

Chapter 8 • Measurement

normAl distribution

the distribution of scores under the normal curve

percentile score

a raw score that has been converted into an expression of the percentage of people who score at or below that score

centrAl tendency

the midpoint, or center, of the data

meAn

a measure of central tendency reflecting the average score

mediAn

the middle score, or the point below which 50 percent of the scores fall

mode

the most commonly observed score

␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣ ␣␣ ␣␣ ␣␣ ␣␣␣ ␣␣␣ ␣␣␣ ␣␣␣

FIgure 8-1 The Normal Curve Illustrating Standard Scores and Percentiles

the one shown in Figure 8-1. The distribution of scores under the normal curve is called the normal distribution. As you can see, a large number of individual cases cluster in the mid- dle of the curve. The farther from the middle (or average) you go, the fewer the cases. Many distributions of scores follow the same normal curve pattern. Most individuals get scores in the middle range, near average. As score extremes are approached, fewer and fewer cases exist, indicating that progressively fewer individuals get lower scores (represented by the left tail of the curve) or higher scores (represented by the right tail of the curve). Other dis- tributions are possible but we will focus on the normal distribution because it is one of the most commonly used.

percentIle score A percentile score is a raw score that has been converted into an expression of the percentage of people who score at or below that score. For example, in Figure 8-1, a score of 55 on Test X or 120 on Test Y would place a person at about the 84th percentile. This means that 84 percent of the people taking the test scored at or below this individual’s score.

The second horizontal line below the curve in Figure 8-1 labeled “Percentiles” rep- resents the distribution of scores in percentile units. By knowing the percentile score of an individual, you already know how that individual compares with others in the group. An in- dividual at the 98th percentile scored the same as or better than 98 percent of the individuals in the group. This is approximately equivalent to getting a raw score of 60 on Test X or 140 on Test Y.

Percentiles can yield useful information. Assume you want to make highly competitive job offers. You can use data sources such as the Bureau of Labor and Statistics (BLS), which typically report the 10th, 25th, 50th, 75th, and 90th percentiles in the distribution of salaries for a given occupation. If you wish to pay salaries at the top 10 percent of the distribution, then you can use the BLS’s percentiles to figure out how much you should pay.

You can also collect salary information from within your firm to determine what might be a competitive job offer. Assume you collected data from all your employees on their current sal- ary levels. How might you describe the data? We discuss this next.

central tenDency central tendency describes the midpoint, or center, of the data. Typical measures of central tendency include the mean, median, and mode. The mean is a measure of central tendency reflecting the average score. For example, you could compute the average sal- ary and then pay at or above this level to be competitive. The median is the middle score, or the 50th percentile, which is the point below which 50 percent of the scores fall below. The mode is the most commonly observed score.

If scores are normally distributed, as they are in Figure 8-1, then the mean, median, and mode are in the same position. This is not always the case if scores are not normally distributed. In the case of data on annual pay, a distribution could have most employees at the left (lower end) of the range and relatively few employees at the high end. This is not a normal distribution. This is regularly observed in organizations, and it is called positive skew. In this case, the mode would be to the left, near the bulk of the distribution, because it is the most commonly observed

␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣

␣␣␣␣␣␣␣␣␣␣␣

␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣

≈ ␣␣␣ ␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣ ≈ ␣␣␣ ␣␣␣ ␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣␣␣␣ ␣␣ ␣␣ ␣␣ ␣ ␣␣ ␣␣ ␣␣



 score. The mean would be shifted to the right due to the high annual pay for a limited number of employees. The median would be somewhere in between. In this case, the firm might extend different job offers depending on which measure it used to describe the data. In labor disputes, it is not uncommon for managers to use average pay as a reference point (indicating higher pay across employees) whereas unions use the mode or median as a reference point (indicating lower pay across employees).

Alternatively, you might see a bimodal distribution, or a distribution with two modes, for annual pay. The mean and median would fall between the two modes but neither would be repre- sentative of true compensation levels because there are probably two separate employee groups represented in the data. Perhaps the two modes represent some employees who are paid on a sal- ary basis and other employees who are paid on an hourly basis. In this case, you might use one of the two modes to determine your competitive offer for compensation, or you could compute the mean, median, and mode separately for the two groups.

varIabIlIty variability describes the “spread” of the data around the midpoint. If you were told that an applicant scored 76 out of 100 points on a work sample test, what would you think? It’s hard to know what to conclude without more information. What if you were told the mean was 70? This additional information helps because you can tell that the applicant did better than average, but you are still missing important information. How much better or worse did the ap- plicant actually fare? To answer this, you need to know the variability of scores. What would you think if the lowest score was 64 and the highest was 76? What if you were told the lowest score was 40 and the highest 100? Knowing the variability of a distribution changes your interpreta- tion of scores.

There are a number of alternative measures of variability but typical measures include the range, variance, and standard deviation. The range is the difference between the highest and lowest observed scores. The range is highly influenced by any single extreme score (an outlier) so it may not effectively represent the true variability in the data. Other measures of variability such as the variance and standard deviation are less affected by outliers. The variance is a math- ematical measure of the spread based on squared deviations of scores from the mean. You can find the formula for variance in the supplement at the end of this chapter. The standard devia- tion is conceptually similar to the average distance from the mean of a set of scores. It is the positive square root of the variance. A data set with a larger standard deviation has scores with more variance and a larger range. For example, if the average score on a measure was 70, and the standard deviation was 3, the scores would be more tightly clustered around the mean than if the standard deviation was 15. If all the scores were the same, the standard deviation would be 0. If everyone scores the same on a measure, the measure isn’t useful in predicting job performance or deciding who to hire. You can see in Figure 8-1 that the range and standard deviation are smaller for Test X than they are for Test Y.

stanDarD scores standard scores are converted raw scores that indicate where a person’s score lies in comparison to a referent group. A common standard score is a z score, which mea- sures the distance of a score from the mean in standard deviation units. There are three determi- nants of a z score: the raw score and the mean and standard deviation of the entire set of scores.

Look at Figure 8-1. Test X and Test Y have different raw score means. Notice that Test X has a mean of 50 and Test Y has a mean of 100. If an individual got a score of 65 on Test X, that person did very well. However, a score of 65 on Test Y would be a poor score. Raw scores often carry limited information by themselves.

Figure 8-1 shows the percent of cases 1, 2, and 3 standard deviations above the mean and 1, 2, and 3 standard deviations below the mean. As you can see, 34 percent of the cases lie between the mean and +1 standard deviation, and 34 percent of the cases lie between the mean and -1 standard deviation. Thus, approximately 68 percent of the cases lie between -1 and +1 standard deviations. Note that for Test X, the standard deviation is 5, and 68 percent of the test takers scored between 45 and 55. For Test Y, the standard deviation is 20, and 68 percent of the test takers scored between 80 and 120.

A z score is calculated by subtracting the referent group’s mean from the target individ- ual’s raw score, and dividing the difference by the measure’s standard deviation in the referent group. The resulting standard z score indicates how many standard deviations the individual’s

vAriAbility

a measure that describes the “spread” of the data around the midpoint

rAnge

the difference between the highest and lowest observed score

outlier

a score that is much higher or lower than most of the scores in a distribution

vAriAnce

a mathematical measure of the spread based on squared deviations of scores from the mean

stAndArd deviAtion

the positive square root of the variance; it is conceptually similar to the average distance from the mean of a set of scores

stAndArd scores

converted raw scores that indicate where a person’s score lies in comparison to a referent group

z score

a standard score that indicates the distance of a score from the mean in standard deviation units

Chapter 8 • Measurement 207



 208

Chapter 8 • Measurement

score is above or below the mean of the referent group. It can be seen in Figure 8-1 that approxi- mately 84 percent of scores fall below a z score of +1, whereas nearly 100 percent of people fall below a z score of +3. The simple formula for a z score is

z

score

= 1Individual’s raw score - Referent group mean2/ Referent group standard deviation

A z score is negative when the target individual’s raw score is below the referent group’s mean, and positive when the target individual’s raw score is above the referent group’s mean.

To compare candidates, we often need a single overall score that represents each can- didate’s combined performance on all of the assessment methods used to evaluate them. Combining a candidate’s raw scores on two or more measures that use different scoring systems is difficult. Imagine an assessment system that evaluates candidates using an interview scored on a 1-to-10 scale and a job knowledge test that is scored on a 0-to-100 scale. Simply averaging the two scores would give disproportionate weight to one of the tests, depending on the mean and standard deviation. Standardizing both scores by converting them to z scores allows them to be easily combined, as shown in Table 8-1.

In Table 8-1, the interview scores have a range of 15 to 22, a mean of 18.25, and a stan- dard deviation of 3. The job knowledge test has a range of 69 to 87, a mean of 78.25, and a standard deviation of 7.46. If you subtract the mean from each raw score and divide by the standard deviation, you will obtain the standard score. For Felix, the calculations would be 115 - 18.252/3 = -1.1 and 187 - 78.252/7.46 = 1.2 (after rounding).

Although meaningfully combining the raw scores would be difficult, combining the z scores is easy and results in a single number reflecting how each candidate did on both of the as- sessment methods relative to the other candidates. In this case, Sue’s outstanding interview score allowed her to overcome her slightly below-average job knowledge test score to be the candidate with the highest overall score. If a company wants to weight multiple assessment methods differ- ently, each standard score can be multiplied by the desired weighting percentage. For example, the formula for weighting the interview score 60 percent and the job knowledge test score 40

percent would beOverall score = 10.6 \* z 2 + 10.4 \* z 2 interview job knowledge test

For Felix, this would be 10.6 \* -1.12 + 10.4 \* 1.22 = -0.18 rather than the 0.1 he received when the interview and knowledge test were equally weighted.

shifting the normal curve

When making candidate selection decisions, it is often assumed that the distribution of appli- cants’ fit with the job reflects the normal curve as depicted by the current talent pool shown in Figure 8-2. If this is true, then a large burden is placed on the selection system to accurately iden- tify which candidates fall to the far right of the curve (the best hires). In practice, however, many of the most desirable people for a position are not in the applicant pool at all. The most talented and competent people are often successfully employed because they are usually being promoted and rewarded for the work they do. As a result, most of these people are semi-passive job seekers at best. Without an effective sourcing and recruiting process, they will not apply to your firm. For example, during the 2008 economic downturn, it was difficult to get passive job seekers with deep experience and a proven track record in advertising to apply for positions at other companies.12

TaBle 8-1 Converting Raw Scores to Standard Scores

Candidate

Felix Sue Lin Pierre

Interview Score

Job Knowledge Test Score

Raw Standard

87 1.2 77 -0.2 69 -1.2 80 0.2

Overall Score

Raw

15 22 19 17

Standard

-1.1 1.3 0.3 -0.4

1-1.1 + 1.22 = 0.1 11.3 - 0.22 = 1.1 10.3 - 1.22 = -0.9 1-0.4 + 0.22 = -0.2

Standard Units



 Chapter 8 • Measurement 209

␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

FIgure 8-2 Shifting the Normal Curve

A passive sourcing approach can result in a distribution of applicants that is shifted to the left, or lower end, as depicted by distribution A. An alternative way of looking at this is to think about the role that sourcing and recruiting play in terms of shaping the qualifications of the appli- cant pool. If done strategically, sourcing and recruiting can discourage applicants who are a poor fit from applying, and increase the number of high-quality passive and semi-passive candidates who do apply. In this case, the distribution would be shifted to the right, as depicted by distribu- tion B. This recruiting and sourcing approach would yield candidates of higher quality, clearly reducing the burden on the selection system to identify the best candidates. This can significantly increase the likelihood of hiring excellent employees.

Describing and interpreting data is part of the process of using data strategically. Strategic staffing is further enhanced when you can use data to understand relationships between measures and variables. In particular, if you can identify predictors of desired staffing outcomes, then this can lead to new selection tools and interventions, such as recruiter training. The next section explains how you can assess the relationship between predictors and outcomes.

usIng Data strategIcally

correlations

A correlation indicates the strength of a linear relationship between two variables. If people who score higher on a measure tend to perform better on the job, or if people who score higher on a measure perform lower on the job, scores and job performance are said to be correlated. A correlation coefficient, also called “Pearson’s r” or the “bivariate correlation,” is a single number that ranges from -1 to +1; it reflects the direction (positive or negative) and magnitude (the strength) of the relationship between two variables. A value of r = 0 indicates that the val- ues of one measure are unrelated to the values of the other measure. A value of r = +1 means that there is a perfectly linear, positive relationship between the two measures. In other words, as the value of one of the measures increases, the value of the other measure increases by an exactly determined amount. By contrast, a value of r = -1 means that there is a perfectly nega- tive (inverse) relationship between the two measures. In other words, as the value of one of the measures increases, the value of the other variable decreases by an exactly determined amount. The information provided by correlations is useful for making staffing decisions. The health care literature is full of studies that document the positive correlation between patient outcomes and proper staffing in health care organizations, for example.13 You can find the formula for cor- relation in the supplement to this chapter along with the formula for Excel. Correlations can be easily computed using spreadsheets or software such as Microsoft Excel, SAS, or SPSS. In most circumstances, we rarely see correlations remotely approaching +1 or -1. Even the correlation between people’s height and weight is typically less than .80. In staffing contexts, we rarely have such precisely measured and highly correlated data. Measurement error, to be discussed later, also reduces the magnitude of the correlations we observe. In addition, restricting the variability of our applicant and hired pools can also reduce the size of the correlations we observe.

The typical values we might see in staffing contexts are + .30 or - .30. Although much lower than the theoretical maximum and minimum, these values can result in significant improvements in the quality of hires. Unstructured interviews, one of the most commonly used selection techniques, often have a correlation of +.20 or less with job performance. A well-structured personality test can have a correlation of + .30 with job performance. Thus, using such a test can have a significant positive economic impact on an organization by significantly improving the hiring process.

correlAtion

the strength of a linear relationship between two variables

correlAtion coefficient

a single number that ranges from -1 to +1; it reflects the direction (positive or negative) and magnitude (the strength) of the relationship between two variables



 210

Chapter 8 • Measurement

scAtter plot

graphical illustration of the relationship between two variables

predictor vAriAble

a variable used to predict the value of an outcome

r2

XY

Very Highly Correlated

XY

Completely Uncorrelated

r2

XY

Moderately to Strongly Correlated

FIgure 8-3 Diagrams for Correlations

One way of thinking about correlations is depicted in the diagram14 shown in Figure 8-3. Think of the variance of a given variable as depicted by a circle. If the circles of two different vari- ables are perfectly overlapping, then the variance of one variable is perfectly correlated with the other. In the first example, the two circles are nearly overlapping, suggesting the correlation between X and Y is approximately either +.90 or –.90. Why either positive or negative? Because the variance is shared regardless of direction of the sign. In the second example, the two circles do not overlap at all, indicating that the correlation between X and Y is 0. In the third example, the two circles overlap nearly half, suggesting the correlation is about +.70 or -.70. Why a correlation of {.70 for a nearly 50 percent overlap? As it turns out, the amount of variance shared by two variables is equal to the square of the correlation, or r2, and .72 is .49 or about 49 percent overlap. Another good way to understand the correlational relationship between two variables is to graph them. Figure 8-4 illustrates the correlations corresponding to several different patterns of data in scatter plots, or graphical illustrations of the relationship between two variables. Each point on the chart corresponds to how a particular person scored on a test and a measure of how he or she performed on the job.

From the scatter plots in Figure 8-4, you can see that a correlation of +1 occurs when the data points are in a perfect line. A correlation of + 1 means that higher test scores on the measure correspond with an exact improvement in performance scores. The test score is called a predictor variable. A predictor variable is a variable used to predict the value of an outcome. In this case, the predictor variable (test score) perfectly predicts the outcome (performance). Now notice the lack of a relationship between scores and performance in the graph showing a correlation of r = -.05. In this case, the scores are almost completely independent of job performance ratings, and these scores are a poor predictor of performance.

When the relationship is perfect, as it is in the “+1” graph, it is easy to see how the trend line should be drawn. However, when the data do not depict a perfect relationship, it’s harder to figure out how to draw the line. In this case, you will need to draw the line in such a way that it minimizes the distance of all the points from the line (i.e., minimizes errors of prediction). This is called a regression line, which will be discussed in the next section. When there is almost no relationship, the regression line will be nearly flat. When there is a negative relationship, the regression line will slope downward.



 Chapter 8 • Measurement 211

12 6 10 5 84 63 42 21

00 0 1 2 3 4 5 6 7 8 9 10

0 1 2 3 4 5 6 7 8 9 10

Test Scores r = +1.00 r = –.43

Test Scores

8 10

7 6 5

9 8 7 6

45

3

2

1

4 3 2 1

00 0 1 2 3 4 5 6 7 8 9 10

Test Scores

r = .32

0 1 2 3 4 5 6 7 8 9 10 Test Scores

r = .12

10 99 88 77 66 55 44 33 22 11

10

00 0 1 2 3 4 5 6 7 8 9 10 0 1 2 3 4 5 6 7 8 9 10

Test Scores Test Scores Curvilinear Relationship r = –.05

r = .04

FIgure 8-4 Correlations Expressed as Scatter Plots

If you found a correlation of r = -.43 between a measure and job performance, would the measure be useful in predicting which candidates are likely to do better on the job? Absolutely— just hire people who perform lower on the measure. The correlation of r = -.43 is of a rea- sonably high magnitude. Thus, it reflects a fairly strong relationship between the measure and on-the-job performance. The direction of the correlation isn’t important. To make this easier to understand, imagine that the measure was assessing typing errors on a test and the job per- formance dimension was typing performance. Candidates scoring lower on the measure made fewer errors, and thus are likely to be better typists. In other words, negative correlations are just

Performance Performance Performance

Performance Performance Performance



 212

Chapter 8 • Measurement

sAmpling error

the variability of sample correlations due to chance

stAtisticAl significAnce

the degree to which the observed relationship is not likely due to sampling error

as useful as positive correlations—negative correlations involving a desirable staffing outcome just mean that lower-scoring candidates are preferable to those with higher scores.

An additional type of relationship between two variables is a curvilinear relationship in which scores are related to outcomes in a nonlinear fashion. This can occur when higher scores are related to higher performance to a point, after which higher scores relate to lower perfor- mance. Curvilinear relationships are sometimes found between the personality trait of conscien- tiousness and job performance.15 Conscientiousness16 refers to being self-disciplined, striving to achieve, and tending to think carefully before acting. If you have ever worked with someone who was extremely detail oriented, strove for perfection, and had a hard time making decisions, you probably understand how too much conscientiousness can sometimes be a detriment to per- formance. Note, however, that in most situations more conscientious people tend to perform better on the job.

If you were to rely solely on the correlation coefficient to evaluate whether or not being conscientious is a predictor of people’s job performance, you might underestimate the measure’s usefulness. If a curvilinear relationship exists, rather than selecting candidates with the highest conscientiousness scores, it would be better to select candidates who score closer to the middle range on the measure. There are specialized statistical techniques for testing for curvilinearity, and it is important to collect data to determine whether a linear or curvilinear relationship exists for the position you’re filling.

Other uses of the correlation coefficient include:

• Relatingstoresizeswithstaffinglevels • Relating seniority in a firm with how well employees perform on the job • Relating the time to fill a job with new-hire quality • Relating the quality of new hires with a business’s performance and the satisfaction of its

customers

Interpreting correlations

Suppose you find a correlation between a job knowledge test and a measure of job success equal to .15. Should you use the measure? Answering this question requires assessing the likelihood that the observed relationship really exists, and then evaluating whether the rela- tionship is strong enough to make the measure useful given its cost. Whenever two variables are correlated using a subset of the total population, there is always the chance that the sample used does not represent the total population. If you had a group of twenty employees, ten women and ten men, and randomly chose four of them, would you always choose two women and two men? On average, you would. But there would be some instances in which you would end up choosing four women. At other times, you would end up choosing four men. Similarly, when computing a correlation coefficient from a sample of people, the correlation might not accurately represent the correlation that exists in the general population or your applicant pool. sampling error is simply the variability of sample correlations due to chance. The usefulness of a correlation can be evaluated by considering statistical significance and practical signifi- cance. Next, we discuss each.

the statIstIcal sIgnIFIcance oF a correlatIon statistical significance is the degree to which the observed relationship is not likely due to sampling error. A correlation is said to be sta- tistically significant if it has less than a certain probability of being due to sampling error––usually 5 percent (called a p-value for probability value). If the probability of a correlation due to chance is .03 (versus, say, .30), then the correlation is said to be “statistically significant.” In other words, the observed correlation is far enough from zero that it is unlikely to be due to sampling error, yielding a probability below .05 or 5 percent. The larger a sample, the more likely observed relationships capture the “true” relationship, and even small effect sizes will be statistically significant. In a small sample, the observed relationship must be larger for the relationship to reach statistical significance because there is a greater probability that the observed relationship is due to sampling error.

the practIcal sIgnIFIcance oF a correlatIon Unfortunately, statistical significance does not guarantee that a predictor is useful. If a sample size is large enough, then even very small correlations can be statistically significant because large samples tend to result in correlation



 estimates with little sampling error. For example, when the military studies predictors of troop performance, because they have a sample in the tens of thousands, even predictors with a very small correlation with the outcome are statistically significant.

After establishing statistical significance, the focus shifts to practical significance. Practical significance means that the observed relationship is large enough to be of value in a practical sense. A correlation that is statistically significant is not necessarily large enough to be of practical significance. Whether or not a correlation is large enough to be practically sig- nificant is in the eyes of the measurer. For example, if hiring errors for a particular job are not costly, then a correlation of .2 might be acceptable to an organization. However, a correlation of .2 might be too low for critical jobs in which hiring errors are costly. In yet other situations, a correlation as low as .15 can still be practically significant.

Practical significance is irrelevant unless the relationship is also statistically significant because otherwise the observed correlation could be due to chance. To be useful, a relationship needs to have both practical and statistical significance. Other factors will determine whether or not a correlation is useful: An assessment system that is inexpensive, for example, might still be useful even if the correlation is not large. Alternatively, if an assessment method that correlated .15 with job success was expensive, took a long time to administer, and was only moderately liked by job candidates, it might not be worth using even if it was a statistically significant predictor of a person’s job success. It depends on the degree to which the assessment yields a return on the money a firm has invested in its use. For example, even if an organization used an assessment method with a low correlation with job success, the company might still earn a good return on the method. Consequently, it is important to look beyond the magnitude of the correlation.

Identifying statistically and practically significant relationships can help organizations execute their business strategies more effectively. The food wholesaler Sysco is a good exam- ple. Sysco, headquartered in Houston, Texas, periodically assesses the correlation between its customers’ satisfaction and its employees’ satisfaction. The company has found that customer loyalty and operational excellence are affected by a satisfied, productive, and committed work- force. Retaining its employees has also helped Sysco cut its operating costs. After discovering the correlation, Sysco implemented a rigorous set of programs to enhance the retention and satisfaction of its employees.17

regressions

Generally, staffing professionals use more than one measure to assess job applicants because it improves the overall validity of a firm’s selection process. However, with a correlation analy- sis, only two variables can be related to one another so only one predictor variable can be used. Multiple regression is a statistical technique that predicts outcomes using one or more predic- tor variables. Assume the predictors and outcomes are measured on an interval or ratio level. Specialized techniques exist for variables that are measured on a nominal or ordinal level but such approaches are beyond the coverage of this chapter. A human resource professional can do a multiple regression analysis to identify the ideal weights to assign to each assessment score (each predictor) to maximize the validity of a set of assessment methods. The analysis is based on each assessment method’s correlation with job success (the outcome) and the degree to which the assessment methods are intercorrelated. For an example of what we mean by intercorrelated, suppose that cognitive ability is highly correlated with interview performance. In this case, it might not make sense to use both variables to predict job success because it would be redundant to do so. If the redundancy is too great, the regression analysis will retain only one of the pre- dictors to use in the final prediction equation. In other words, the redundant predictor would be assigned a near-zero weight.

One way of visualizing relationships among the variables in multiple regression is de- picted in the diagram18 shown in Figure 8-5. Assume Y is the criterion, or outcome being predicted, and X and Z are the predictor variables. In the first example you can see by the overlap that both X and Z predict Y, and both X and Z are uncorrelated with each other. In the second example, both X and Z predict Y but X and Z are highly correlated. You can eas- ily see in this second case that X and Z contribute little unique prediction beyond the other. This is similar to a case in which you measure the same concept (e.g., intelligence) using

prActicAl significAnce

an observed relationship that is large enough to be of value in a practical sense

multiple regression

a statistical technique that predicts an outcome using one or more predictor variables; it identifies the ideal weights to assign each predictor so

as to maximize the validity of a set of predictors; the analysis is based on each predictor’s correlation with the outcome and the degree to which the predictors are themselves intercorrelated

Chapter 8 • Measurement 213



 214

Chapter 8 • Measurement

␣␣␣␣␣␣␣␣␣ ␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

␣␣␣␣ ␣␣␣␣

ZYX ␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

Z

XY

␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

FIgure 8-5 Diagrams of Multiple Regression

␣

␣

␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣

two different but highly related assessments. The redundancy doesn’t add new information so you can either eliminate one of the assessments or combine them (depending on cost). In the third example, you see a more typical situation. Here X and Z are moderately correlated with Y and with each other. Both X and Z add unique information to the prediction of Y but they are related to each other. As an illustration, a firm might use a cognitive ability test and structured interview scores to predict job performance and cognitive ability and interview scores might be related.

The equations used to do a regression analysis can be computed by hand, but this is a cum- bersome process if there are more than two predictor variables. The equation for two predictor variables (a multiple regression) can be found in the supplement at the end of this chapter. A variety of software packages, including Excel, SPSS, SAS, and Stata, can be used to easily per- form an analysis using three or more variables and we provide the instructions for using Excel to conduct a multiple regression analysis in the supplement.

Job success

+ (b2 \* Test score2) + (b3 \* Test score3) c

Each variable is examined for its statistical relationship to the predicted outcome and a regression (or multiple regression) equation is derived. The regression equation is of the format

= Constant + 1b \* Test score 2 predicted 1 1

The constant, or intercept, is a number added to everyone’s predicted job success score, and the bs are the regression weights that are multiplied by each test score. An applicant’s scores on the test(s) are entered into the resulting model and used to calculate the test taker’s predicted job success. Because the regression analysis operates on raw scores, these scores do not need to be standardized because the weights take into account the differences in means and standard deviations. For example, if Miguel scored 50 on an interview, 27 on a personality measure, and 20 on a job knowledge test, his predicted job success would be 141 based on the following equation:

Job success = 10 + 12 \* Interview2 + 11 \* Personality2 + 1.2 \* Job knowledge2 predicted

The intercept (10) and weights (2, 1, and .2) come from the results of the statistical analy- sis. Miguel’s score of 141 would then be compared with the other candidates to determine if he should be hired. In general, only equations with variables found to be statistically significant should be used to make staffing decisions.



 Regression analysis is also used to predict future headcount requirements. Consider the regression equation that uses projected sales per month and the number of expected customers to determine a firm’s headcount requirements:

Full@time employees = 60 + 1.00015 \* Sales2 + 1.3 \* Expected customers2 If the firm’s projected sales are $1,000,000, and the company projects that it will acquire

Full@time employees = 60 + 1.00015 \* 1,000,0002 + 1.3 \* 2502 = 60 + 150 + 75 = 285

To prevent giving different variables credit for predicting the same part of the criterion, multiple regression analysis examines the effect of each variable (e.g., each test score) on the cri- terion (e.g., job success) after controlling for other variables. Using multiple regression requires high quality measures.

What are the characterIstIcs oF useFul Measures?

Two properties of a good measure are its reliability and validity. We discuss each next as well as the importance of a measure’s standard error of measurement.

reliability

reliability refers to how dependably, or consistently, a measure assesses a particular character- istic. If you obtained wildly different weights each time you stepped on a scale, would you find the scale useful? Probably not. The same principle applies to measures relevant to staffing, such as job knowledge, personality, intelligence, and leadership skills.

A measure that yields similar scores for a given person when it is administered multiple times is reliable. Reliability sets boundaries around the usefulness of a measure. A measure can- not be useful if it is not reliable, but even if it is reliable, it still might not be useful—for example, if it doesn’t measure what you’re seeking to determine but something else instead. Reliability is a critical component of any staffing measure, including candidate assessment. If a person com- pletes a personality test twice, will he or she get a similar score or a much different score? If the scores radically change, then perhaps the test isn’t reliable. Why would a job candidate score dif- ferently when completing a personality test again, you might wonder? Think of why you might score differently on a midterm given on Monday and one given on Friday, and you should have some insights. Some possible reasons are the following:19

• the respondent’s temporary psychological or physical state. For example, differing lev- els of anxiety, fatigue, or motivation can affect test results. If you are stressed the first time you are tested but are relaxed the second time, you might respond differently.

• environmental factors. Differences in the environment, such as room temperature, light- ing, noise, or even the test administrator, can influence an individual’s performance. If it is quiet on one occasion, and you hear distracting construction equipment on the other, you might obtain different scores.

• the version, or form, of the measure. Many measures have more than one version, or form. For example, the ACT and SAT college entrance examinations have multiple forms. The items differ on each form, but each form is supposed to measure the same thing. Because the forms are not exactly the same, a respondent might do better on one form than on another. If one version happened to be harder, or it was equally challenging but tapped into material you knew less well, then you would perform more poorly. In the case of the ACT and SAT, scores can be adjusted to reflect the difficulty of each form.

• different evaluators. Certain measures are scored subjectively—that is, they are de- termined by an evaluator’s judgments of the respondent’s performance or responses. Differences in the training, experience, and frame of reference among the evaluators can result in different scores for a respondent. This is why two interviewers evaluating the same job candidate might come to completely different conclusions about the quality of that job candidate.

250 new customers, then it will need 285 full-time employees:

Chapter 8 • Measurement 215

reliAbility

how dependably, or consistently, a measure assesses a particular characteristic



 216 Chapter 8 • Measurement

rAndom error

error that is not due to any consistent cause

systemAtic errors

error that occurs because of consistent and predictable factors

deficiency error

occurs when you fail to measure important aspects of the attribute you would like to measure

contAminAtion error

occurs when other factors unrelated to whatever is being assessed affect the observed scores

These factors are sources of measurement error in the assessment process. Measurement error can be systematic or random. In some cases, the measurement error can be random, as in the flip of a coin. You don’t always get 50 heads and 50 tails in 100 flips. This is an example of a random error. Similar things can happen in a staffing context. Running into traffic on the way to work or experiencing bad weather can cause employee productivity to randomly fluctuate in unpredictable ways. systematic errors are errors that occur because of consistent and predict- able factors. For example, an employee’s productivity may go down every Tuesday because he or she works late at a second job Monday night.

The sources of systematic errors can include factors such as the time of day or day of the week. Administering a difficult work sample test in the morning might, for example, lead to dif- ferent results than if the test was administered in the afternoon or late at night. Systematic errors can also result when there are systematic differences across evaluators. For example, some inter- viewers might regularly tend to rate most interviewees near the middle of a 1-to-10 scale, whereas other interviewers might tend to regularly rate most interviewees on the high end of the scale. In this example, the differences among the evaluators are a source of systematic error. Another source may be due to the measurement items themselves. Items that are reverse-worded, confus- ing, or overly complex can lead to systematic errors. The following question is a good example:

Using a 1-to-5 scale where 1 is very true and 5 is very untrue, answer this question: “In previous leadership roles you rarely failed to set goals on a timely and consistent basis while providing good feedback to your team.”

If you reflect for a moment on this item, you can see how difficult it is to understand and how it could lead to systematic error due to wording. People may systematically vary in the ac- curacy of their responses, depending on their verbal ability, motivation to complete the survey quickly, or simple attention to detail. This is not random error because it is attached to specific and identifiable personal characteristics.

If there were no systematic or random errors of measurement, then the respondent would get the same score each time. If you step on a scale then you will get a reading. If the scale is perfectly reliable and you step on it again 10 seconds later, you would see the same reading. This is your true score. If in real life, your scale, like ours, slightly fluctuates, then you might see a slightly different reading 10 seconds later. This is random measurement error because it is due to a variety of factors unrelated to your actual weight. However, if you weigh yourself every day in the morning when you wake up and after lunch, you might find you systematically weigh more due to having eaten lunch. In a staffing context, test and performance scores also contain a true score plus some variation due to random and systematic errors. An applicant is unlikely to obtain exactly the same score on a knowledge test every time. Part of this could be due to random fac- tors, or the error could be systematic.

The distinction between random and systematic errors is important: Systematic errors can be controlled or eliminated. For example, you can weigh yourself at the same time each day while wearing the same clothes or you can use only highly trained interviewers. Random errors, however, cannot be controlled but still affect the quality of measurement. Some measures are subject to more random errors than others, just as some scales provide more reliable measures of weight than do others.

A deficiency error is yet another type of error. It occurs when you fail to measure impor- tant aspects of the attribute you would like to measure. The underlying attribute being measured is called a construct. If you wanted to measure an applicant’s ability to use calculus for an en- gineering job, then this ability is your construct of interest. However, if the test you were using focused only on algebra, a deficiency error would result. A contamination error occurs when other factors unrelated to one’s advanced math skills (in this case) affect the observed scores. For example, if the calculus test had many complex word problems, then language could influence the results. If the test was administered under varying conditions (a loud versus a quiet setting, early morning versus late at night, using calm and helpful administrators versus loud and anxious administrators), then these factors could affect scores.

The diagram shown in Figure 8-6 illustrates deficiency, contamination, and relevance. Assume you used supervisory ratings to measure job performance. In what ways might supervi- sory ratings be deficient? It is possible that supervisors focus on productivity and meeting dead- lines more than teamwork, quality, and safety. Or supervisors may not know about the quality of



 Chapter 8 • Measurement 217

Construct

␣␣

FIgure 8-6

␣␣␣␣␣␣␣␣␣␣

␣␣␣␣␣␣␣␣␣␣

␣␣␣␣␣␣␣␣␣␣␣␣␣␣

Measure

Deficiency, Contamination, and Relevance

customer service provided and only attend to quantity of sales. In each case, the supervisor may overlook important aspects of job performance and using their ratings alone may result in defi- ciency. Supervisory ratings can be contaminated, too. What might affect supervisory ratings other than actual job performance? The research literature is filled with information about sources of rater contamination, including stereotypes, halo effects, and similar-to-me bias, among many other sources.20 This is contamination. The overlapping area between the construct and the measure indi- cates relevance, or the degree to which the measure captures the intended concept to be measured.

It is impossible to eliminate all sources of error, but measures can be made more reliable by standardizing the measurement process as much as possible. For example, you can pretest the items on a test to ensure they are clear and they statistically correlate with each other consis- tently. Interviewers can be trained to ask the same questions, avoid bias, and use the same behav- iorally based scoring key to make their ratings. Test administrators can give tests at consistent times, under similar conditions, and so forth.

Conceptually, reliability is the correlation of an item, scale, or measurement instrument with a hypothetical set of true scores. However, in practice, true scores are not available for the computation of a correlation. Instead, reliability must be estimated by correlating different types of observations. This will be elaborated upon later.

The reliability of a measure is indicated by the reliability coefficient, which is expressed as a number ranging between 0 and 1, with 0 indicating no reliability (no correlation between the measure and the true score) and 1 indicating perfect reliability (perfect correlation between the measure and the true score). Like a correlation, we express reliability as a decimal––for ex- ample, .70 or .91. The closer the reliability coefficient is to 1.0, the more repeatable or reliable the scores are. Near-perfect reliability is extremely rare. The reason that reliability coefficients are only positive (as opposed to correlations, which can range from -1 to +1) is that observed scores and true scores should relate to each other in a consistently positive manner. Table 8-2 presents some general guidelines for interpreting the reliability of a measure.

TaBle 8-2 General Guidelines for Interpreting Reliability Coefficients21

Reliability Coefficient Value

.90 and up .80–.89 .70–.79 .50–.69 .00–.49

Interpretation

Superior Good Adequate for most needs Limited applicability Not useful at all



 218

Chapter 8 • Measurement

test-retest reliAbility

reflects the repeatability of scores over time and the stability of the underlying construct being measured

AlternAte or pArAllel form reliAbility

indicates how consistent scores are likely to be if a person completes two or more forms of the same measure

internAl consistency reliAbility

indicates the extent to which items on a given measure assess the same construct

inter-rAter reliAbility

indicates how consistent scores are likely to be if the responses are scored by two or more raters using the same item, scale, or instrument

The reliability coefficient is not the only thing to consider in selecting or rejecting an as- sessment method. To evaluate a measure’s reliability, you should consider the type of measure, the type of reliability estimate reported, and the context in which the measure will be used. There are several types of reliability estimates. Before deciding to use a measure, such as a personality evaluation or cognitive ability test, it is important to learn about the reliability of the measure. Organizations sometimes purchase tests or assessment tools, and information about reliability is often provided by the creator and publisher of these tests and tools. You should be familiar with the different kinds of reliability estimates reported. Next, we discuss several different types of reliability. Reliability can be estimated in one of four ways.

test-retest relIabIlIty test-retest reliability reflects the repeatability of scores over time and the stability of the underlying construct being measured (e.g., a person’s math skills, personality, intelligence, honesty, and other relevant characteristics). Test-retest reliability is estimated by the correlation between two (or more) administrations of the same measure across different times or locations on the same sample. This assesses stability over time. Some constructs are more stable than others. For example, mechanical ability is more stable over time than is mood or anxiety. Therefore, we would expect a higher test-retest reliability coef- ficient on a mechanical aptitude test than on a measure of anxiety. For constructs like mood, which vary over time, an acceptable test-retest reliability coefficient may be lower than is suggested in Table 8-2.

alternate or parallel ForM relIabIlIty Developers often make multiple forms or ver- sions of a measure that are intended to assess the same thing and be of the same difficulty level. alternate or parallel form reliability indicates how consistent scores are likely to be if a per- son completes two or more forms of the same measure. This reliability is estimated by the cor- relation between two (or more) administrations of different forms that are supposed to measure the same construct to the same population. A high parallel form reliability coefficient indicates that the different forms are very similar, which means that it makes virtually no difference which version of the measure is used. On the other hand, a low parallel form reliability coefficient suggests that the different forms are probably not comparable and may be measuring different things. In this case, the multiple forms cannot be used interchangeably and scores on each form cannot be directly compared.

Internal consIstency relIabIlIty internal consistency reliability indicates the extent to which items on a given measure assess the same construct. A high internal consistency reli- ability coefficient indicates that the items on a measure function in a similar manner. Internal consistency is based on the correlation among the items comprising a measure. For example, you might have 10 items measuring math skill. If all the items measure math skill reliably, and they are internally consistent, then scores on one item should correlate highly with scores on another. Items in the measure can be split into even and odd items or first half and second half items. Scores on these halves can then be correlated with each other. This is called split-half reliability, which is one indicator of internal consistency. The most commonly used indicator of internal consistency is Cronbach’s alpha. It is an estimate of the average of all possible split-half reliabilities.

If finance and history questions were included on an exam for a staffing class, the test would have lower internal consistency reliability than if the test contained only staffing-related questions because the different types of items would yield varying patterns of scores. Measures that assess multiple characteristics are usually divided into distinct sections, and a separate inter- nal consistency reliability coefficient is reported for each section in addition to one for the whole measure.

Inter-rater relIabIlIty inter-rater reliability indicates how consistent scores are likely to be if the responses are scored by two or more raters using the same item, scale, or instrument. On some measures, like during Olympic gymnastic events, different raters evaluate responses or behaviors and subjectively determine a score. Often in business contexts different people evalu- ate the same job applicant (e.g., the HR recruiter and the hiring manager). Differences in raters’ judgments create variations in a person’s scores for the same measure or event.



 Inter-rater reliability is based on the correlation of scores between or among two or more raters who rate people or objects using the same item, scale, or instrument. A high inter-rater re- liability coefficient indicates that the judgment process is consistent and that the resulting scores are reliable. Although inter-rater reliability coefficients are typically lower than other types of reliability estimates, rater training can increase inter-rater reliabilities. This type of reliability is particularly important for understanding the usefulness of interview evaluations.

These four reliability estimation methods are not necessarily mutually exclusive, nor do they need to yield the same results. A measure of job knowledge that has many different dimen- sions within the test might show low internal consistency reliability. However, people’s job knowledge characteristics might be relatively stable, in which case the test scores will be similar across administrations (high test-retest reliability). In this case, you should compute separate in- ternal consistency reliabilities for each dimension. As another illustration, two distinct measures of leadership capability might yield high parallel forms reliability. Additionally, the items within each of the measures might be internally consistent. However, the test-retest reliability could be low if leadership training occurred between the administration of the two tests. In this case, you would expect to see low test-retest reliability because the training ought to change or improve people’s leadership capabilities.

Clearly, the acceptable level of reliability will differ depending on the type of measure and the reliability estimate used. A measure of mood may exhibit low reliability across admin- istrations because, as we explained, moods fluctuate over time. However, the measure might still be useful for predicting how applicants will react during interviews. In this case, the items measuring mood must yield consistent scores among themselves, even if they vary over time as an overall score.

standard error of Measurement

As we have explained, the measurement process always contains some type of error. The problem is that we wish to use imperfect scores to make decisions despite the presence of error. It is help- ful to know how much error exists when we use a given score. The standard error of measure- ment (seM) is the margin of error that you should expect in an individual score because of the imperfect reliability of the measure. (The formula for SEM is given in this chapter’s supplement.) SEM represents the spread of scores you might have observed had you tested the same person repeatedly. The lower the standard error, the more accurate the measurements. If the SEM is zero, then the observed score is the true score. However, we know that error exists so we can compute a range of possibilities around the observed score. This is a confidence interval. Although not technically precise, you can think of it in this manner. If you score 85 out of a possible 100 on a measure that has an SEM of 2, there is a 68 percent chance that the “true” score lies between 83 and 87, and about a 95 percent chance that the true score lies between 81 and 89.

In a normal distribution, 68 percent of cases fall between +1 and -1 standard deviations from the mean, and approximately 95 percent of cases in a population fall between +2 and -2 standard deviations from the mean. The SEM tells us the standard deviation of errors. If we cen- ter our mean around 85 (the observed score), then we can use the SEM to determine the chance that the true score will fall within a given range. With an SEM of 2, one standard deviation of error below the mean is 83185 - 22 and one standard deviation above the mean is 87185 + 22. This gives us a 68 percent confidence interval. A similar computation can be made for the 95 percent confidence interval.

The SEM is a useful measure of the accuracy of individual scores. If you have received a manual with a test or assessment tool, then when you’re evaluating the reliability coefficients of a measure, it is important to review the explanations provided for the following:

• the types of reliability used. The manual should explain why a certain type of reliability coefficient was reported and discuss any known sources of random measurement error. • How the reliability studies were conducted. The manual should describe the conditions

under which the data was obtained, including the length of time that passed between the administrations of a measure in a test-retest reliability study. In general, reliabilities tend to drop as the time between administrations increases.

• the characteristics of the sample group. The manual should indicate the important characteristics of the sample group used to gather the reliability information, such as the

stAndArd error of meAsurement (sem)

the margin of error that you should expect in an individual score because of the imperfect reliability of the measure

Chapter 8 • Measurement 219



 220

Chapter 8 • Measurement

vAlidity

how well a measure assesses a given construct and the degree to which you can make specific conclusions or predictions based on observed scores

education levels, ages, occupations, and other relevant characteristics of the people in the group. This will allow you to compare the characteristics of the people you want to mea- sure with the sample group. If they are sufficiently similar, then the reported reliability estimates will probably hold true for your population as well.

The important thing to remember is that high reliability measures will have lower SEMs, which means that observed scores are more likely to reflect true scores. Additionally, reliabili- ties can drift over time. With longer periods between administrations, test-retest correlations are likely to go down, and the SEMs will then go up. Moreover, as we have explained, even if a measure is reliable, it doesn’t mean it’s useful. Reliable measures may or may not measure what you intend to measure, and they may or may not predict desired staffing outcomes. This is where the issue of validity comes into play, which we discuss next.

validity

validity is the most important issue in selecting a measure. It refers to how well a measure as- sesses a given construct and the degree to which you can make specific conclusions or predic- tions based on observed scores. Validity is the cornerstone of strategic staffing. If you wish to use data to make decisions, then the data must relate in meaningful ways to desired outcomes. If you can predict high-quality talent using various kinds of tests, then they will give you a com- petitive edge over firms that do not use valid tests for selection.

It is important to understand the differences between reliability and validity. Validity will tell you how useful a measure is for a particular situation; reliability will tell you how consistent scores from that measure will be. You cannot draw valid conclusions unless the measure is reli- able. But even when a measure is reliable, it might not be valid. For example, you might be able to measure a person’s shoe size reliably, but it probably won’t be useful as a predictor of the person’s job performance. Any measure used in staffing needs to be both reliable and valid for the situation.

Figure 8-7 shows a popular bull’s-eye illustration of the relationship between reliability and validity. The center of the target is whatever construct you are trying to measure, usually some aspect of job success. Each “shot” at the bull’s-eye is a measurement for a single person. A bull’s-eye means that your measure is perfectly assessing the person on that construct. The further you are from the center, the more your measurement is off for that person.

The dots close together in Figure 8-7 reflect higher reliability than the dots more spread out. Dots centered on the bull’s-eye reflect higher reliability and validity than dots clustered away from the bull’s-eye. You can easily see that if the measure is not reliable (the dots are widely scattered), it is not possible for them to be valid (on target).

Figure 8-7 shows three possible situations. In the first one, shots are consistent, but miss the center of the target—we are consistently measuring the wrong value for all observations. This measure is thus reliable (consistent), but not valid (not accurate). An everyday example might be a scale that consistently registers a weight that is 20 pounds too heavy. A staffing example might be a math test that gives consistent results but that is too easy. In the second bull’s-eye, hits are spread across the target, and we are consistently missing the center, reflect- ing a measure that is neither reliable nor valid. This is like a scale that gives random readings and is on average 20 pounds off. A math test that does a poor job of measuring math and is

␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣ ␣␣␣␣␣␣␣␣␣

FIgure 8-7 Illustration of Reliability and Validity



 plagued by error (it is both contaminated and deficient) might yield the second pattern. The third bull’s-eye shows hits that are consistently in the center of the target, reflecting a measure that is both reliable and valid. This is like a scale that gives the same weight each time, and the weight is accurate. In staffing, this pattern might be exhibited by a high-quality math test that is consistent in results and neither deficient nor contaminated. This is our goal in measurement and assessment.

A measure’s validity is established in reference to a specific purpose. Thus, the measure might be valid for some purposes but not be valid for others. For example, a measure you use to make valid predictions about someone’s technical proficiency on the job may not be valid for predicting his or her leadership skills, job commitment, or teamwork effectiveness.

Similarly, a measure’s validity is established in reference to specific groups called refer- ence groups. Thus, the same measure might not be valid for different groups. For example, a problem-solving skills measure designed to predict the performance of sales representatives might not be valid or useful for predicting the performance of clerical employees.

As we have explained, the manuals that accompany assessment tools, or tests, should de- scribe the reference groups used to develop the measures. The manuals should also describe the groups for whom the measure is valid and how the scores for the individuals belonging to each of the groups were interpreted. You, then, must determine if the measure is appropriate for the particular type of people you want to assess. This group of people is called your target popula- tion, or target group.

Although your target group and the reference group might not have to match perfectly, they must be sufficiently similar so that the measure will yield meaningful scores for your group. For example, you will want to consider factors such as the occupations, reading levels, and cultural and language differences of the people in your target group. Use only assessment pro- cedures and instruments demonstrated to be valid for your target group(s) and for your specific purpose. This is important because the Uniform Guidelines on Employee Selection Procedures require assessment tools to have adequate supporting evidence for the conclusions reached with them in the event adverse impact occurs. Although all employee selection procedures—for ex- ample, interviews—do not have to be validated, scored assessments that have an adverse impact should be validated if technically feasible.

The user of an assessment tool is ultimately responsible for making sure that validity evi- dence exists for the conclusions reached using the measures. This applies to all measures and procedures used (including interviews), whether the measures have been bought off-the-shelf, developed externally, or developed in-house. This means that if you develop your own measures or procedures, you should conduct your own validation studies. If validation is not possible, the scored assessment should be eliminated. If informal or nonscored assessments have an adverse impact, the employer should either eliminate the tool or use a more formal one that can be validated.

Although the Uniform Guidelines focus on adverse impact and legal liability, validation is even more important from a strategic perspective. Strategically, it makes sense to use only those measures that reliably and validly assess what is important to job success and that predict desired outcomes. Anything else is potentially an expensive waste of time. Invalid measures can lead to missed opportunities for selecting high-quality talent, or worse yet, the selection of people who will perform poorly. The cost of selection-related errors is high. However, they can be dramatically reduced by using valid measures for selection. There are many types of validity, all of which address the usefulness and appropriateness of using a given measure. We discuss them next.

Face valIDIty One aspect of validity is whether the measure seems to measure what it is supposed to measure. This is face validity. It is a subjective assessment of how well items or measures seem to be related to the requirements of the job. Face validity is often important to job applicants who tend to react negatively to assessment methods if they perceive them to be unre- lated to the job (or not face valid). Even if a measure seems face valid, if it does not predict job performance, then it should not be used. Hypothetically, a measure of extroversion might look like an acceptable way to measure job candidates applying for a sales position. Nonetheless, it might still fail to predict whether or not an extroverted person performs well as a sales represen- tative. Perhaps outgoing salespeople talk too much and sell too little, for example.

fAce vAlidity

a subjective assessment of how well items seem to be related to the requirements of the job

Chapter 8 • Measurement 221



 222 Chapter 8 • Measurement

vAlidAtion

the cumulative and ongoing process of establishing the job relatedness of a measure

content-relAted vAlidAtion

the process of demonstrating that the content of a measure assesses important job-related behaviors

construct-relAted vAlidAtion

the process of demonstrating that a measure assesses the construct, or characteristic, it claims to measure

criterion-relAted vAlidAtion

the process of demonstrating that there is a statistical relationship between scores from a measure (the predictor) and the criterion (the outcome)

valIDatIon In order to be certain an employment measure is useful and valid, you must col- lect evidence relating the measure to a job. The process of establishing the job relatedness of a measure is called validation. Validation is the cumulative and ongoing process of establishing the job relatedness of a measure.

The Uniform Guidelines on Employee Selection Procedures discuss the following three methods of conducting validation studies and describe conditions under which each type of vali- dation method is appropriate:

• content-related validation is the process of demonstrating that the content of a measure assesses important job-related behaviors. For example, a mathematical skills test would have high content validity for an engineering position, but a typing skills test might have low content validity if the job required only minimal typing. However, the same typing test might have strong content validity for a clerical position.22 Content validity also applies to the items making up a measure. A math test might have low content validity if it includes items focusing on, for example, psychology or biology, or other facets unrelated to the position being hired for.

• construct-related validation is the process of demonstrating that a measure assesses the construct, or characteristic, it claims to measure. This method often applies to measures that attempt to assess the abstract traits of candidates, such as their personalities, honesty, or aptitudes. A construct-related validation would need to be done if, for example, a bank wanted to test its tellers for a trait such as “numerical aptitude.” In this case, the aptitude is not an observable behavior, but a concept created to explain possible future behaviors. To demonstrate that the measure possesses construct validity, the bank would need to show (1) that the measure did indeed assess the desired trait (numerical aptitude) and (2) that this trait corresponded to success on the job.23 Construct validity is established by the pattern of correlations among items within a measure and the pattern of correlations of the scores from that measure with other relevant outcomes. Content validity can also be used to help establish construct validity.

• criterion-related validation is the process of demonstrating that there is a statistical re- lationship between scores from a measure (the predictor) and the criterion (the outcome), usually some aspect of job success such as job performance, training performance, or job tenure. This form of validation uses either correlational or regression-based procedures. In other words, in the case of a positive relationship, individuals who score high on the measure should tend to perform better on the job success criterion than those who score low. If the criterion is obtained at the same time the predictor measure is collected, it is called concurrent validity; if the criterion is obtained after the initial measure (the pre- dictor) is collected, then it is called predictive validity. Consider the position of a mill- wright, who installs, repairs, replaces, and dismantles machinery and heavy equipment. A measure might be designed to assess how employees’ mechanical skills are related to their performance when it comes to servicing machines (criterion). A strong relation- ship would validate using the measure.24 Predictive validity would be estimated if you measured employees’ mechanical skills before they were hired and then correlated those skills with their subsequent performance. Concurrent validity would be estimated if at a single point in time you measured the mechanical skills of a company’s current employ- ees as well as correlated their scores with their performance. The criterion-related validity of a measure is measured by the validity coefficient, which we discuss in more detail in the next section of this chapter.

All types of validity are important. You can establish content validity for a math skills test using job analysis techniques to determine the level, type, and difficulty of math required for a position and the importance of math to job performance. You can then construct a large number of items that would potentially measure math skills and establish that job experts agree that the items are related to the math skills. You can also use job expert ratings to establish that each math test item is essential by computing a content validity ratio (the formula is available in the chapter supplement). You can also ask job incumbents and supervisors to determine if any important math skills are missing from the measure. This would help to establish the content and face validity of the measure. Then you could correlate the math skills test with the performance of engineers to see if they predict job performance. This establishes criterion-related validity. All these forms of



 Chapter 8 • Measurement 223

validity, combined with reliability information and information about how the items within the math skill test relate to each other can then establish the construct validity of the measure.

the valIDIty coeFFIcIent The validity coefficient is a number between 0 and + 1 that indi- cates the magnitude of the relationship between a predictor (such as test scores) and the criterion (such as a measure of actual job success). The validity coefficient is the absolute value of the correlation between the predictor and criterion. The larger the validity coefficient, the more con- fidence you can have in predictions made from the scores. Because jobs and people are complex, a single measure can never fully predict what a person’s job performance will be because suc- cess on the job depends on so many factors. Therefore, validity coefficients rarely exceed .40 in staffing contexts.25

As a general rule, the higher the validity coefficient, the more beneficial it is to use the measure. Validity coefficients of .21 to .35 are typical for a single measure. The validities of se- lection systems that use multiple measures will probably be higher because you are using differ- ent tools to measure and predict different aspects of performance. By contrast, a single measure is more likely to measure or predict fewer aspects of total performance. Table 8-3 shows some general guidelines for interpreting a single measure’s validity. It is difficult to obtain validity coefficients above .50 even if multiple measures are used.

evaluatIng valIDIty Evaluating a measure’s validity is a complex task. In addition to the magnitude of the validity coefficient, you should also consider at a minimum the following factors:

• Thelevelofadverseimpactassociatedwithyourassessmenttool • Thenumberofapplicantscomparedtothenumberofopenings • Thenumberofcurrentlysuccessfulemployees • Thecostofahiringerror

• Thecostoftheselectiontool • Theprobabilityofhiringaqualifiedapplicantwithoutusingascoredassessmenttool.

Here are three scenarios illustrating why you should consider these factors, individually and in combination with one another, when evaluating validity coefficients26:

Scenario One: You have few applicants for each open position. Most of the applicants will be hired because the positions do not require a great deal of skill. In this situation, you might be willing to accept a selection tool that has a validity in the range of “potential to be useful” or “useful in certain circumstances” if the assessment method is cheap, you need to fill the positions quickly, you do not have many applicants to choose from, and the level of skill required is not that high.

Scenario Two: You are recruiting for jobs that require a high level of accuracy, and mis- takes could be dangerous or costly. In this case, a slightly lower validity coefficient would probably not be acceptable to you because hiring an unqualified worker would be too much of a risk. Instead, you would need to use a selection tool that reported validities con- sidered to be “very beneficial.”

Scenario Three: The company you are working for is considering a very expensive assess- ment system that results in fairly high levels of adverse impact. There are other assessment tools on the market associated with lower adverse impact, but they are less valid and just as costly. Additionally, making a hiring mistake would put your company at too much risk.

vAlidity coefficient

a number between 0 and +1 that indicates the magnitude of the relationship between a predictor (such as test scores) and the criterion (such as a measure of actual job success)

TaBle 8-3 General Guidelines for Interpreting Validity Coefficients27

Validity Coefficient Value

Above .35 .21–.35 .11–.20 Below .11

Interpretation

Very beneficial Potential to be useful Useful in certain circumstances Unlikely to be useful



 224

Chapter 8 • Measurement

vAlidity generAlizAtion

the degree to which evidence of validity obtained in one situation can be generalized to another situation without further study

Consequently, your company decides to implement the assessment given the difficulty in hiring for the particular positions, the “very beneficial” validity of the assessment, and your failed attempts to find alternative instruments with less adverse impact. However, your com- pany will continue to try to find ways to reduce the adverse impact of the system.

Clearly, most situations require you to consider multiple factors. For example, the recruit- ing and selection context must be considered along with validity. Even if a staffing system is valid and predicts job success well, unintended consequences may result from the use of the system. For example, the following might be adversely affected:

• Applicants. A valid assessment system can result in adverse impact by differentially se- lecting people from various protected groups, have low face validity, and result in law- suits. As we discussed in Chapter 3, fair employment legislation prohibits the use of tests to discriminate against job applicants because of their race, color, religion, sex, or national origin. In some cases, job candidates may perceive valid measures as irrelevant to the job in question.

• the organization’s time and cost. A valid assessment system can have an unacceptably long time to fill or high cost per hire; result in the identification of high-quality candidates who demand high salaries, resulting in increasing payroll costs; and be cumbersome, dif- ficult, or complex to use.

• future recruits. A system can be valid but if it is too long or onerous then applicants, particularly high-quality applicants, are more likely to drop out of consideration; word that a firm is using time-consuming selection practices could reduce the number of applica- tions; a valid system could result in differential selection rates and reduce the number of applicants from a particular gender, ethnicity, or background; and valid systems can still be viewed as unfair, resulting in fewer future applicants.

• current employees. The assessment system may favor external applicants or not give all qualified employees an equal chance of applying for an internal position; employees might therefore question its fairness.

The point here is not to ignore validity. Rather, it is to highlight the need to address these factors so that highly valid measures can be used for selection while minimizing the downsides of using them.

Validity is typically evaluated using single samples for specific jobs. There are limitations, both practical and statistical, to conducting validity studies in cases in which there are relatively few people in a given position. Computing validities with small samples can lead to large sam- pling errors and reduce the likelihood that your findings will be statistically significant. One method of dealing with this problem is to use validity generalization.

valIDIty generalIzatIon validity generalization refers to the degree to which evidence of a measure’s validity obtained in one situation can be generalized to another situation without further study.28 A statistical technique called meta-analysis is used to combine the results of validation studies done for a single measure on many different target groups and for a variety of jobs. The goal of the meta-analysis is to estimate a measure’s “true validity” and to identify whether we can generalize the results to all situations or determine if the same measure works differently in different situations.

Validity generalization studies can often give staffing professionals insight about the strength of the relationship between a measure and a person’s job performance. However, there is no guar- antee that all employers would find the same level of validity of a study when it comes to their own workforces. Every organization has different situational factors that can drastically impact the valid- ity of a measure. Although the legal acceptability of validity generalization has yet to be thoroughly considered in the courts, online assessment companies, such as preVisor [(www.previsor.com)](https://jigsaw.brytewave.com/books/9780133597066/pages/108660660/www.previsor.com), are increasingly using validity generalization as part of their validation of their collection of products.

using existing assessment Methods

Conducting your own validation study is expensive. Moreover, as we have explained, many smaller firms do not have enough employees in a relevant job category to make it feasible to con- duct a study. One alternative is to conduct cooperative studies across firms within an association



 Chapter 8 • Measurement 225

to collect more validation data more quickly. For example, insurance companies can share data to obtain large amounts of validation data on specific positions. Another alternative is that it can be advantageous to use professionally developed assessment tools and procedures for which documentation on validity already exists. However, you must ensure that the validity evidence obtained from an “outside” study can be suitably “transported” to your particular situation. In fact, the Uniform Guidelines require as much. To determine if a particular measure is valid for your intended use, consult the manual and available independent reviews such as those in Buros Institute’s Mental Measurements Yearbook29 and Test Critiques.30

When evaluating validity information purchased from a vendor, you should consider the following:

• Available validation evidence supporting the use of the measure for specific purposes. The manual should include a thorough description of the procedures used in the validation studies and the results of those studies. Also consider the definition of job success used in the validation study.

• the possible valid uses of the measure. The purposes for which the measure can legiti- mately be used should be described, as well as the performance criteria that can validly be predicted.

• the similarity of the sample group(s) on which the measure was developed with the group(s) with which you would like to use the measure. For example, was the measure developed on a sample of high school graduates, managers, or clerical workers? What was the racial, ethnic, age, and gender mix of the sample?

• Job similarity. A job analysis should be performed to verify that your job and the original job are substantially similar in terms of ability requirements and work behavior.

• Adverse impact evidence. Consider the adverse impact reports from outside studies for each protected group that is part of your labor market. If this information is not available for an otherwise qualified measure, conduct your own study of adverse impact, if feasible.

In addition, if an organization would like to use a vendor’s assessment or other tool glob- ally, it is important to thoroughly evaluate this capability. Many vendors that claim to be global are actually not capable of delivering a product globally.31

This chapter’s Develop Your Skills feature provides some advice on measuring the char- acteristics of job applicants.

Develop Your SkillS

Assessment Tips32

To effectively assess job candidates, employers must be aware of the inherent limitations of any assessment procedure as well as how to properly use their chosen assessment methods. Here are 10 tips on conducting an effective assessment program:

1. The measures should be used in a purposeful manner— have a clear understanding of what you want to measure and why you want to measure it.

2. Use a variety of tools—because no single measurement tool is 100 percent reliable or valid, use a variety of tools to measure job-relevant characteristics.

3. Use measures that are unbiased and fair to all groups— this will allow you to identify a qualified and diverse set of finalists.

4. Use measures that are reliable and valid. 5. Use measures that are appropriate for the target popula-

tion—a measure developed for use with one group might

not be valid for other groups. 6. Ensure that your administration staff is properly trained—

the training should include how to administer the mea- sure as well as how to handle special situations with

sensitivity—for example, how and when to provide rea-

sonable accommodations for people with disabilities. 7. Ensure suitable and uniform assessment conditions—noise, poor lighting, inaccurate timing, and damaged equipment

can adversely affect respondents. 8. Keep your assessment instruments secure—developers

and administrators should restrict access to the instru- ment’s questions, and the measures should be periodically revised.

9. Maintain the confidentiality of the results—the results should be shared only with those who have a legitimate need to know. Personal information should not be re- leased to other organizations or individuals without the informed consent of the respondent.

10. Interpret the scores properly—the inferences made from the results should be reasonable, well founded, and not based on superficial interpretation; careful attention should be paid to contamination and deficiency errors; the manual for the tools should also provide instructions on how to properly interpret the results.



 226

Chapter 8 • Measurement

selection errors

not hiring someone who would have been successful at the job or hiring someone who is not successful at the job

wHole-person ApproAcH

the practice of using a variety of measures and procedures to more fully assess people

stAndArdizAtion

the consistent administration and use of a measure

obJectivity

the amount of judgment or bias involved in scoring an assessment measure

selection errors

Professionally developed measures and procedures that are used as part of a planned assessment program can help you select and hire more qualified and productive employees even if the mea- sures are not perfect. It is essential to understand that all assessment tools are subject to errors, both in measuring a characteristic, such as verbal ability, and in predicting job success criteria, such as job performance. This is true for all measures and procedures.

• Do not expect any measure or procedure to measure a personal trait or ability with perfect accuracy for every single person.

• Donotexpectanymeasureorproceduretobecompletelyaccurateintermsofpredictinga candidate’s job success.

Certainly, selecting employees who are highly able is important. However, there are many factors that affect a person’s performance. You also need a motivated employee who clearly understands the job to be performed, for example. The employee also needs the time and resources necessary to succeed in the job. Several of these factors can be predicted using good measurement tools. This is why selection procedures typically involve three to five dis- tinct selection measures (at a minimum) that are combined in some fashion to make a final hiring decision.

Despite these efforts, there always will be cases in which a score or procedure will predict someone to be a good worker, who, in fact, is not. There will also be cases in which an individual receiving a low score will be rejected when he or she would actually be a capable and good worker. In the staffing profession, these errors are called selection errors. False positives and false negatives are two types of selection errors. False positives occur when you erroneously classify a weak applicant as being a good hire. False negatives occur when you erroneously classify a strong applicant as being a weak hire. As you try to reduce one type of error you may increase the other so there are trade-offs in how you make your hiring decision. These issues will be covered more in the following chapters.

Selection errors cannot be completely avoided, but they can be reduced, for example, by using a variety of measures. Using a variety of measures and procedures to more fully as- sess people is referred to as the whole-person approach to assessment. This approach will help reduce the number of selection errors and boost the effectiveness of your overall decision making.33

standardization and objectivity

standardization is the consistent administration and use of a measure. Standardization reflects the consistency and uniformity of the conditions as well as the procedures for administering an assessment method. Computerization helps to ensure that all respondents receive the same instructions and the same amount of time to complete the assessment. Because maintaining stan- dardized conditions is the responsibility of the people administering the assessment, training all administrators in proper procedures and control of conditions is critical. This is true for inter- viewing as well as any other assessment approach. In addition to being legally important, stan- dardization is also valuable because recruiters should consistently evaluate candidates on their competencies, styles, and traits.

Norms reflect the distribution of scores of a large number of people whose scores on an assessment method are to be compared. The standardization sample is the group of respondents whose scores are used to establish norms. These norms become the comparison scores for deter- mining the relative performance of future respondents.

Objectivity refers to the amount of judgment or bias involved in scoring an assessment measure. The scoring process for objective measures is free of personal judgment or bias. The number of words typed in a minute is an objective measure, as is the amount of weight a firefighter candidate can lift. Subjective measures, on the other hand, contain items (such as essay or interview questions) for which the score can be influenced by the attitudes, biases, and personal characteristics of the person doing the scoring. Whenever hiring decisions are subjective, it is also a good idea to involve multiple people in the hiring process, preferably of diverse gender and race, to generate a more defensible decision.34 Because they produce the most accurate measurements, it is best to use standardized, objective measures whenever possible.



 creatIng anD valIDatIng an assessMent systeM

Creating an effective assessment and selection system for any position in any organization begins with a job analysis. As you learned in Chapter 4, after understanding the requirements of job suc- cess, you identify the important knowledge, skills, abilities, and other characteristics (KSAOs) and competencies required of a successful employee. You then identify reliable and valid methods of measuring these KSAOs and competencies, and create a system for measuring and collecting the re- sulting data. The integrity and usefulness of the data generated by each measure needs to be consid- ered when deciding which measures to use. The data collected from each measure is then examined to ensure that it has an appropriate mean and standard deviation. Remember, a measure on which everyone scores the same or nearly the same is not as useful as a measure that produces a wide range of scores. Candidates’ scores on each assessment method are then correlated or entered into a regres- sion equation to evaluate any redundancies among the measures and to assess how well the group of measures predicts job success. Adverse impact and the cost of the measures are also considered in evaluating each measure. After the final set of measures is identified, selection rules are developed to determine which scores are passing. The usefulness and effectiveness of the system is then peri- odically reevaluated to ensure that it is still predicting job success without adverse impact.

benchmarking

It is sometimes useful to compare an organization’s staffing data with those of other organizations to understand better whether the organization is doing well or poorly on a particular dimension. For example, is a voluntary turnover rate of 30 percent good or bad? In some positions, such as the po- sitions held by retail employees, this would be a good turnover level compared to the industry aver- age. In other positions, a 30 percent turnover rate would be unusually high. Benchmarking other firms can give a company comparative information about dimensions including the following:

• Application rates • Averagestartingsalaries • Averagetimetofill • Averagecostperhire

There are numerous sources of relatively high-quality benchmark information, but it can be expensive. Some sources of benchmarking data include

• CorporateLeadershipCouncil • WatsonWyattandotherstaffingconsultingfirms • HackettGroup • TheSaratogaInstitute(nowpartofPricewaterhouseCoopers) • Staffing.org • Many industry associations, such as the Society for Human Resource Management, track

benchmark information and make it available to their members.

evaluating assessment Methods

The determinants of the effectiveness of any internally or externally developed assessment method include

1. validity—whether the assessment method predicts the relevant components of job success 2. return on investment—whether the assessment method generates a financial return that

exceeds the cost associated with using it 3. Applicant reactions—including the perceived job relatedness and fairness of the assess-

ment method 4. usability—the willingness and ability of people in the organization to use the method con-

sistently and correctly 5. Adverse impact—whether the method can be used without discriminating against mem-

bers of a protected class 6. the selection ratio—whether the method has a low selection ratio

The importance of a firm’s selection ratio and base rate to the effectiveness of an assess- ment method deserve further elaboration. Taylor and Russell35 were among the first to dem- onstrate that validity alone will not determine the usefulness of an assessment. The tables they generated, taking into account selection ratio and base rate, demonstrated that assessments with

Chapter 8 • Measurement 227

