Throughout this book, references to numerous research studies provide documentation for its assertions. Readers who wish to learn more about the studies cited are encouraged to consult the original sources listed in the notes for each chapter. Much of this literature requires basic familiarity with the statistics and research methods used in human resource management (HRM). This appendix describes some of the basic concepts needed to understand the original research reports. Although this appendix uses examples relevant to selection and placement (described in Chapter 7), the basic concepts described also apply to research on most other topics.

Regardless of the method or site used to study human resource issues, researchers need to be concerned about the reliability and validity of measurement devices. Reliability refers to the consistency of measurement, and validity relates to the truth or accuracy of measurement. Both are expressed by a correlation coefficient (denoted by the symbol $r$).

**CORRELATION COEFFICIENT**

A correlation coefficient expresses the degree of linear relationship between two sets of scores. A positive correlation exists when high values on one measure (e.g., a job knowledge test) are associated with high scores on another measure (e.g., overall ratings of job performance). A negative correlation exists when high scores on one measure are associated with low scores on another measure. The range of possible correlation coefficients is from $+1$ (a perfect positive correlation coefficient) to $-1$ (a perfect negative correlation coefficient). Several linear relationships represented by plotting actual data are shown in Exhibit B.1.

The correlation between scores on a predictor ($x$) and a criterion ($y$) is typically expressed for a sample as $r_{xy}$. If we did not have a sample but were able to compute our correlation on the population of interest, we would express the correlation as $\rho_{xy}$ ($\rho$ is the Greek letter rho). In almost all cases, researchers do not have access to the entire population. Therefore, they must estimate the population correlation coefficient based on data from a sample of the population. For instance, an organization may desire to know the correlation between a test for computer programming ability (scores on a predictor, $x$) and job performance (scores on a performance appraisal rating form, $y$) for all its computer programmers, but decides it cannot afford to test all computer programmers (i.e., the population of interest). Instead, the organization may select a sample of computer programmers and estimate the correlation coefficient in the population ($\rho_{xy}$) based on the observed correlation in the sample ($r_{xy}$).
For each scatterplot in Exhibit B.1, a solid line represents the pattern of data points. Each line is described by an equation, which takes the general form for a straight line: \( y = a + bx \), where \( a \) is the point at which the line intercepts the \( y \)-axis and \( b \) is the slope of the line. Such equations, called *prediction equations*, allow researchers to estimate values of \( y \) (the criterion) from their knowledge of \( x \) (the predictor). For example, we may conduct a study to determine the correlation between sales performance (i.e., dollar sales volume) and number of years of sales experience for a group of salespeople. Once we have developed a prediction equation, we can then estimate how well a sales applicant might perform on the job. The equation might read,

\[
\text{Dollar sales per month} = 50,290 + 2,000 \times \text{Years of sales experience}
\]

Using this equation, we would predict that a salesperson with 10 years of experience will generate $70,290 a month in sales, while a new salesperson with only one year of experience will be expected to generate only $52,290 a month in sales.

As you might expect, decision makers often employ more than one predictor in their equations. For example, organizations often use multiple predictors when making selection and placement decisions. Similar to that of the single-predictor approach, the purpose of multiple prediction is to estimate a criterion (\( y \)) from a linear combination of predictor variables: \( y = a + b_1x_1 + b_2x_2 + \ldots + b_mx_m \). Such equations are called *multiple-prediction equations* or *multiple-regression equations*. The \( b \)s are the regression weights applied to the predictor measures. The relationship between the predictors and the criterion score is referred to as the *multiple-correlation coefficient* (denoted by the symbol \( R \)).

For example, suppose a manager believes that sales are a function of years of experience, education, and shyness (scored 0 for people who are not very shy to 7 for people who are very shy). To determine whether the manager’s intuition is correct, we might conduct a research study. Information on each of these predictors could be collected, along with sales performance, or criterion, data. A multiple-regression equation such as the following could be generated:

\[
\text{Dollar sales per month} = 25,000 + (1,500 \times \text{Years of experience}) + (500 \times \text{Years of schooling}) - (25 \times \text{Score on shyness})
\]
This equation would indicate that a salesperson who scores low on shyness \((1)\) and has 10 years of experience plus a college degree (16 years of schooling) will be expected to generate $47,975 in sales \([25,000 + (1,500 \times 10) + (1,500 \times 16) - (25 \times 1)]\).

Multiple-regression analysis has been used to determine how to combine information from multiple selection devices. Multiple regression has also been used to assess such things as the effects of rater and ratee characteristics (e.g., sex, experience, prior performance rating, and rate of pay) on performance appraisal and pay decisions and to measure the effects of organizational characteristics (e.g., size, industry, and sales) on human resource planning and policies.

**RELIABILITY**

If a measure such as a selection test is to be useful, it must yield reliable results. The reliability of a measure can be defined and interpreted in several ways. Each of these methods is based on the notion that observed scores \((x)\) comprise true scores \((T)\) plus some error \((E)\) or \(x = T + E\) where \(T\) is the expected score if there were no error in measurement. To the extent that observed scores on a test are correlated with true scores, a test is said to be reliable. That is, if observed and true scores could be obtained for every individual who took a personnel selection test, the squared correlation between observed and true scores in the population is \((\rho_x^2)\) would be called the reliability coefficient for that selection test.

One means of estimating reliability is test-retest reliability \((\rho_{xx})\). This method is based on testing a sample of individuals twice with the same measure and then correlating the results to produce a reliability estimate.

Another means of estimating the reliability of a measure is to correlate scores on alternate forms of the measure. Alternate test forms are any two test forms that have been constructed in an effort to make them parallel; their observed score means that variances (i.e., measures of the spread of scores about the means) and correlations with other measures may be equal or very similar. They are also intended to be similar in content.

A problem with test-retest reliability and alternate forms reliability is the necessity of testing twice. In contrast, internal consistency reliability is estimated based on only one administration of a measure. The most common method, coefficient \(\alpha\) (alpha), yields a split-half reliability estimate. That is, the measure (e.g., a test) is divided into two parts, which are considered alternate forms of each other, and the relationship between these two parts is an estimate of the measure’s reliability.

Researchers are interested in assessing the reliability of measures based on one or more of these methods because reliability is a necessary condition for determining validity. The reliability of a measure sets a limit on how highly the measure can correlate with another measure because it is very unlikely that a measure will correlate more strongly with a different measure than with itself.

**ESTIMATING POPULATION COEFFICIENTS**

If we were able to assess the relationship between a predictor and a criterion in a population of interest with no measurement error, then we would have computed the true correlation coefficient for the population, \(\rho_x\). Because we almost never have the population available and almost always have measurement error, our observed correlation coefficients underestimate the population coefficients.
That is, predictor and criterion unreliability are statistical artifacts that lower predictor–criterion relationships.

Two other statistical artifacts that obscure true relationships are sampling error and range restriction. A sampling error is an inaccuracy resulting from the use of a sample that is smaller than the population when computing the validity coefficient. A range restriction is a correlation or validity coefficient computed between the predictor and criterion scores for a restricted group of individuals. For example, suppose you were interested in determining the correlation between height and weight. If you had data that reflected the entire range of human variability, the correlation would be fairly substantial. However, if you studied only retired adult men, the correlation would be artificially low owing to the restricted range of heights and weights represented in your sample.

Formulas have been developed to remove the effects of predictor unreliability, criterion unreliability, and range restriction, and to determine sampling error variance. Researchers can use these correction formulas to remove the influence of statistical artifacts and, consequently obtain a better idea of the predictor–criterion relationship in the relevant population. A number of studies have examined the effects of variations in sample size, range restriction, and reliability on the size and variability of observed validity coefficients. These studies have improved our understanding of how observed validity coefficients are affected by measurement error and statistical artifacts.

**VALIDITY**

As defined in the American Psychological Association’s *Principles for the Validation and Use of Personnel Selection Procedures*, validity is the degree to which inferences from scores on tests or assessments are supported by evidence. This means that validity refers to the inferences made from the use of a measure, not to the measure itself. Two common strategies used to justify the inferences made from scores on measures are criterion-related validation and content-oriented validation.

**Criterion-Related Validation**

Criterion-related validation empirically assesses how well a predictor measure forecasts a criterion measure. Usually, predictor measures are scores on one or more selection “tests” (e.g., a score from an interview or a score to reflect the amount of experience), and criteria measures represent job performance (e.g., dollar sales per year or supervisory ratings of performance). Two types of criterion-related validation strategies are concurrent validation and predictive validation. These are shown in Exhibit B.2. Concurrent validation evaluates the relationship between a predictor and a criterion for all participants in the study at the same time. For example, the HR department could use this strategy to determine the correlation between years of experience and job performance. The department would collect from each person in the study information about years of experience and performance scores. All persons in the study would have to be working in similar jobs, generally in the same job family or classification. Then a correlation would be computed between the predictor scores and criterion scores.

The steps in determining predictive validity are similar, except that the predictor is measured sometime before the criterion is measured. Thus, predictive validity is determined by measuring an existing group of employees on a predictor and then later gathering their criterion measures.
The classic example of a predictive validation analysis is AT&T’s Management Progress Study.¹ In that study, researchers at AT&T administered an assessment center to 422 male employees. Then, they stored the scores from the assessment center and waited. After eight years, they correlated the assessment center scores with measures of how far the same individuals progressed in AT&T’s management hierarchy. For a group of college graduates, the predictions were highly accurate; a correlation of .71 was obtained between the assessment center predictions and the level of management achieved.

**Content-Oriented Validation**

On many occasions, employers are not able to obtain sufficient empirical data for a criterion-related study. Consequently, other methods of validation are useful. One of the most viable is *content-oriented validation*. It differs from a criterion-related strategy in that it uses subjective judgments and logic as the basis for arguing that a predictor is likely to be effective in forecasting a criterion. To employ a content validation strategy, one must know the duties of the actual job. As discussed in Chapter 5, information about job tasks and duties can be obtained using one or more job analysis procedures. Once the duties are known, logic is used to argue that a predictor is relevant to performance in the job. For example, if a job analysis showed that word processing duties were a substantial portion of a job, then you might logically conclude that a test of word processing skills would be a valid predictor of job performance. For this type of validation, the most important data are job analysis results.

**Validity Generalization**

Since the mid-1900s, hundreds of criterion-related validation studies have been conducted in organizations to determine the predictive effectiveness of HR measures (e.g., ability tests) for selecting and placing individuals.² Often, the validity coefficients for the same or a similar predictor-criterion relationship differed substantially from one setting to another. Although researchers were aware that these differences were affected by range restriction, predictor unreliability, criterion unreliability, and sampling error, only recently were corrections for these statistical artifacts integrated into systematic procedures for estimating to what degree true validity estimates for the same predictor–criterion relationship generalize across settings.

A series of studies has applied validity generalization procedures to validity coefficient data for clerical jobs, computer programming jobs, petroleum industry jobs, and so on. In general, these investigations showed that the effects of
range restriction, predictor unreliability, criterion unreliability, and sample size accounted for much of the observed variance in validity coefficients for the same or similar test-criterion relationship within an occupation (i.e., a job grouping or job family). Thus, the estimated true (corrected) validity coefficients were higher and less variable than the observed (uncorrected) validity coefficient.

The implication of these findings is that inferences (predictions) from scores on selection tests can be transported across situations for similar jobs. That is, if two similar jobs exist in two parts of an organization, a given selection test may have approximately the same validity coefficients for both jobs. If validity generalization can be successfully argued, an organization can save a great deal of time and money developing valid, job-related predictors when the inferences from a predictor for a job have already been established.

The concept of validity generalization, or meta-analysis, has also been applied to other areas of HR research. Such research has led to a better understanding of the effectiveness of interventions such as training programs, goal-setting programs, and performance measurement (appraisal) programs.

Cross-Validation

HR researchers are also interested in how stable their prediction equations are across samples. Cross-validation studies address this concern. For the prediction equations developed by researchers to be of any practical use, they must produce consistent results. Cross-validation is a procedure for determining how much capitalization on chance has affected the prediction equation (or regression weights). In the case of HR selection research, for example, one is interested in how well the regression weights estimated in a sample of job incumbents will predict the criterion value of new job applicants not tested in the sample.

Traditional or empirical cross-validation typically involves holding out some of the data from the initial sample and then applying the equation developed in the initial sample to the holdout sample to evaluate the equation’s stability. In general, this procedure is less precise than one using a formula for estimating the stability of regression equations. The reason for this is that in formula-based estimates, all the available information (the total sample) is used at once in estimating the original weights.

Credits: The authors wish to express their thanks to Michael Burke, Tulane University, for his contributions, comments, and suggestions for the materials contained in this appendix.

ENDNOTES

