CHAPTER 8

How Do We Gather Evidence of Validity Based on a Test’s Relationships With External Criteria?

After completing your study of this chapter, you should be able to do the following:

- Identify evidence of validity of a test based on its relationships to external criteria (criterion-related validity), and describe two methods for obtaining this evidence.
- Read and interpret validity studies.
- Discuss how restriction of range occurs and its consequences.
- Describe the differences between evidence of validity based on test content and evidence based on relationships with other variables.
- Describe the difference between reliability and validity.
- Define and give examples of objective and subjective criteria, and explain why criteria must be reliable and valid.
- Calculate a validity coefficient and the coefficient of determination, and conduct a test of significance for the validity coefficient.
- Explain the concept of regression, calculate and interpret a linear regression formula, and interpret a multiple regression formula.
“The graduate school I’m applying to says they won’t accept anyone who scores less than 1,000 on the GRE. How did they decide that 1,000 is the magic number?”

“Before we married, my fiancé and I went to a premarital counselor. She gave us a test that predicted how happy our marriage would be.”

“My company uses a test for hiring salespeople to work as telemarketers. The test is designed for people selling life insurance and automobiles. Is this a good test for hiring telemarketers?”

Have you ever wondered how psychological tests really work? How can we be comfortable using an individual’s answers to test questions to make decisions about hiring him for a job or admitting her to college? Can mental disorders really be diagnosed using scores on standard questionnaires?

Psychologists who use tests for decision making are constantly asking these questions and others like them. When psychologists use test scores for making decisions that affect individual lives, they as well as the public want substantial evidence that the correct decisions are being made.

This chapter describes the processes that psychologists use to ensure that tests perform properly when they are used for making predictions and decisions. We begin by discussing the concept of validity evidence based on a test’s relationships to other variables, specifically external criteria. As we discussed in Chapter 7, this evidence has traditionally been called criterion-related validity. We also discuss the importance of selecting a valid criterion measure, how to evaluate validity coefficients, and the statistical processes that provide evidence that a test can be used for making predictions.

What Is Evidence of Validity Based on Test-Criteria Relationships?

In Chapter 7, we introduced you to the concept of evidence of validity based on a test’s relationship with other variables. We said that one method for obtaining evidence is to investigate how well the test scores correlate with observed behaviors or events. When test scores correlate with specific behaviors, attitudes, or events, we can confirm that there is evidence of validity. In other words, the test scores may be used to predict those specific behaviors, attitudes, or events. But as you recall from Chapter 7, we cannot use such evidence to make an overall statement that the test is valid. We also said that this evidence has traditionally been referred to as criterion-related validity (a term that we use occasionally in this chapter as it is still widely used by testing practitioners).

For example, when you apply for a job, you might be asked to take a test that is designed to predict how well you will perform on the job. If the job is clerical and the test really predicts how well you will perform on the job, your test score should be related to your skill in performing clerical duties such as word processing and filing. To provide evidence that the test predicts clerical performance, psychologists correlate test scores for a large number of people with another measure of their performance on clerical tasks, such as supervisor ratings. The measure of performance that we correlate with test scores is called the criterion. And if higher test scores are associated with higher performance ratings, then we can say that the test has demonstrated evidence of validity based on the relationship between these two variables. In more traditional usage people say that the test has demonstrated criterion-related validity.
Educators use admissions tests to forecast how successful an applicant will be in college or graduate school. The SAT and the Graduate Record Examination (GRE) are examples of admissions tests used by colleges. The criterion of success in college is often the student’s first-year grade point average (GPA).

In a clinical setting, psychologists often use tests to diagnose mental disorders. In this case, the criterion is the diagnoses made by several psychologists or psychiatrists independent of the test. Researchers then correlate the diagnoses with the test scores to establish evidence of validity.

**Methods for Providing Evidence of Validity Based on Test-Criterion Relationships**

As you also recall from Chapter 7, there are two methods for demonstrating evidence of validity based on test-criterion relationships: predictive and concurrent. This section defines and gives examples of each method.

**The Predictive Method**

When it is important to show a relationship between test scores and a future behavior, researchers use the predictive method to establish evidence of validity. In this case, a large group of people take the test (the predictor), and their scores are held for a predetermined time interval. When the time interval has elapsed, researchers collect a measure of some behavior, for example, a rating or other measure of performance, on the same people (the criterion). Then researchers correlate the test scores with the criterion scores. If the test scores and the criterion scores have a strong relationship, we say the test has demonstrated **predictive evidence of validity**.

Researchers at Brigham Young University used the predictive method to demonstrate evidence of validity of the PREPation for Marriage Questionnaire (PREP-M). For Your Information Box 8.1 describes the study they conducted.

Psychologists might use the predictive method in an organizational setting to establish evidence of validity for an employment test. To do so, they administer an employment test (predictor) to candidates for a job. Researchers file test scores in a secure place, and the company does not use the scores for making hiring decisions. The company makes hiring decisions based on other criteria such as interviews or different tests. After a predetermined time interval, usually three to six months, supervisors evaluate the new hires on how well they perform the job (the criterion). To determine whether the test scores predict the candidates who were successful and unsuccessful, researchers correlate the test scores with the ratings of job performance. The resulting correlation coefficient is called the **validity coefficient**, a statistic used to infer the strength of the evidence of validity that the test scores might demonstrate in predicting job performance.

To get the best measure of validity, everyone who took the test would need to be hired so that all test scores could be correlated with a measure of job performance. It is desirable to get the widest range of test scores possible (including the very low ones) to understand fully how all the test scores relate to job performance.

Gathering evidence of predictive validity can present problems for some organizations because it is important that everyone who took the test is also measured on the criterion. Some organizations might
Evidence of Validity Based on Test-Criteria Relationships of a Premarital Assessment Instrument

In 1991, researchers at Brigham Young University (Holman, Larson, & Harmer, 1994) conducted a study to determine the evidence of validity based on test-criteria relationships of the PREParation for Marriage Questionnaire (PREP-M; Holman, Busby, & Larson, 1989). Counselors use the PREP-M with engaged couples who are participating in premarital courses or counseling. The PREP-M has 206 questions that provide information on couples’ shared values, readiness for marriage, background, and home environment. The researchers contacted 103 married couples who had taken the PREP-M a year earlier as engaged couples and asked them about their marital satisfaction and stability.

The researchers predicted that those couples who had high scores on the PREP-M would express high satisfaction with their marriages. The researchers used two criteria to test their hypothesis. First, they drew questions from the Marital Comparison Level Index (Sabatelli, 1984) and the Marital Instability Scale (Booth, Johnson, & Edwards, 1983) to construct a criterion that measured each couple’s level of marital satisfaction and marital stability. The questionnaire showed an internal consistency of .83. The researchers also classified each couple as either “married satisfied,” “married dissatisfied,” or “canceled/delayed” and as either “married stable,” “married unstable,” or “canceled/delayed.” These classifications provided a second criterion.

The researchers correlated the couples’ scores on the PREP-M with their scores on the criterion questionnaire. The husbands’ scores on the PREP-M correlated at .44 (p < .01) with questions on marital satisfaction and at .34 (p < .01) with questions on marital stability. The wives’ scores on the PREP-M were correlated with the same questions at .25 (p < .01) and .20 (p < .05), respectively. These correlations show that PREP-M is a moderate to strong predictor of marital satisfaction and stability—good evidence of the validity of the PREP-M. (Later in this chapter, we discuss the size of correlation coefficients needed to establish evidence of validity.)

In addition, the researchers compared the mean scores of those husbands and wives classified as married satisfied, married dissatisfied, or canceled/delayed and those classified as married stable, married unstable, or canceled/delayed. As predicted, those who were married satisfied or married stable scored higher on the PREP-M than did those in the other two respective categories. In practical terms, these analyses show that counselors can use scores on the PREP-M to make predictions about how satisfying and stable a marriage will be.

For Your Information Box 8.1

FOR YOUR INFORMATION BOX 8.1
Evidence of Validity Based on Test-Criteria Relationships of a Premarital Assessment Instrument

In 1991, researchers at Brigham Young University (Holman, Larson, & Harmer, 1994) conducted a study to determine the evidence of validity based on test-criteria relationships of the PREParation for Marriage Questionnaire (PREP-M; Holman, Busby, & Larson, 1989). Counselors use the PREP-M with engaged couples who are participating in premarital courses or counseling. The PREP-M has 206 questions that provide information on couples’ shared values, readiness for marriage, background, and home environment. The researchers contacted 103 married couples who had taken the PREP-M a year earlier as engaged couples and asked them about their marital satisfaction and stability.

The researchers predicted that those couples who had high scores on the PREP-M would express high satisfaction with their marriages. The researchers used two criteria to test their hypothesis. First, they drew questions from the Marital Comparison Level Index (Sabatelli, 1984) and the Marital Instability Scale (Booth, Johnson, & Edwards, 1983) to construct a criterion that measured each couple’s level of marital satisfaction and marital stability. The questionnaire showed an internal consistency of .83. The researchers also classified each couple as either “married satisfied,” “married dissatisfied,” or “canceled/delayed” and as either “married stable,” “married unstable,” or “canceled/delayed.” These classifications provided a second criterion.

The researchers correlated the couples’ scores on the PREP-M with their scores on the criterion questionnaire. The husbands’ scores on the PREP-M correlated at .44 (p < .01) with questions on marital satisfaction and at .34 (p < .01) with questions on marital stability. The wives’ scores on the PREP-M were correlated with the same questions at .25 (p < .01) and .20 (p < .05), respectively. These correlations show that PREP-M is a moderate to strong predictor of marital satisfaction and stability—good evidence of the validity of the PREP-M. (Later in this chapter, we discuss the size of correlation coefficients needed to establish evidence of validity.)

In addition, the researchers compared the mean scores of those husbands and wives classified as married satisfied, married dissatisfied, or canceled/delayed and those classified as married stable, married unstable, or canceled/delayed. As predicted, those who were married satisfied or married stable scored higher on the PREP-M than did those in the other two respective categories. In practical terms, these analyses show that counselors can use scores on the PREP-M to make predictions about how satisfying and stable a marriage will be.

not be able to hire everyone who applies regardless of qualifications, and there are usually more applicants than available positions, so not all applicants can be hired. Also, organizations frequently will be using some other selection tool such as an interview to make hiring decisions, and typically, only people who do well on the interview will be hired. Therefore, predictive studies in organizations may only have access to the scores of a portion of the candidates for the job. Because those actually hired are likely to
be the higher performers, a restriction of range in the distribution of test scores is created. In other words, if the test is a valid predictor of job performance and the other selections tools that are used to make a hiring decision are also valid predictors, then people with lower scores on the test will be less likely to be hired. This causes the range of test scores to be reduced or restricted to those who scored relatively higher. Because a validity study conducted on these data will not have access to the full range of test scores, the validity coefficient calculated only from this restricted group is likely to be lower than if all candidates had been hired and included in the study.

Think of it like this: The worst case of restricted range would be if everyone obtained exactly the same score on the test (similar to what would happen if you hired only those people who made a perfect score on the test). If this situation occurred, the correlation between the test scores and any other criteria would be zero. As you can see, if the test scores do not vary from person to person, high performers and lower performers would all have the same test score. We cannot distinguish high performers from low performers when everybody gets the same score, and therefore these test scores cannot be predictive of job performance. Using the full range of test scores enables you to obtain a more accurate validity coefficient, which usually will be higher than the coefficient you obtained using the restricted range of scores. Fortunately, the correlation coefficient can be statistically adjusted for restriction of range, which, when used properly, can provide a corrected estimate of the validity coefficient of the employment test in the unrestricted population.

These problems exist in educational and clinical settings as well because individuals might not be admitted to an institution or might leave during the predictive study. For Your Information Box 8.2 describes a validation study that might have failed to find evidence of validity because of a flawed design.

The Concurrent Method

The method of demonstrating concurrent evidence of validity based on test-criteria relationships is an alternative to the predictive method that we discussed earlier in this chapter. In the concurrent method, test administration and criterion measurement happen at approximately the same time. This method does not involve prediction. Instead, it provides information about the present and the status quo (Cascio, 1991). A study by Watson and colleagues (1996), described in For Your Information Box 8.3, is a good example of a concurrent validity study.

The concurrent method involves administering two measures, the test and a second measure of the attribute, to the same group of individuals at as close to the same point in time as possible. For example, the test might be a paper-and-pencil measure of American literature, and the second measure might be a grade in an American literature course. Usually, the first measure is the test being validated, and the criterion is another type of measure of performance such as a rating, grade, or diagnosis. It is very important that the criterion test itself be reliable and valid (we discuss this further later in this chapter). The researchers then correlate the scores on the two measures. If the scores correlate, the test scores demonstrate evidence of validity.

In organizational settings, researchers often use concurrent studies as alternatives to predictive studies because of the difficulties of using a predictive design that we discussed earlier. In this setting, the process is to administer the test to employees currently in the position for which the test is being considered as a selection tool and then to collect criterion data on the same people (such as performance appraisal data). In some cases, the criterion data are specifically designed to be used in the concurrent study, while in other cases recent, existing data are used. Then the test scores are correlated with the criterion data and the validity coefficient is calculated.
FOR YOUR INFORMATION BOX 8.2

Did Restriction of Range Decrease the Validity Coefficient?

Does a student’s academic self-concept—how the student views himself or herself in the role of a student—affect the student’s academic performance? Michael and Smith (1976) developed the Dimensions of Self-Concept (DOSC), a self-concept measure that emphasizes school-related activities and that has five subscales that measure level of aspiration, anxiety, academic interest and satisfaction, leadership and initiative, and identification versus alienation.

Researchers at the University of Southern California (Gribbons, Tobey, & Michael, 1995) examined the evidence of validity based on test-criteria relationships of the DOSC by correlating DOSC test scores with grade point average (GPA). They selected 176 new undergraduates from two programs for students considered at risk for academic difficulties. The students came from a variety of ethnic backgrounds, and 57% were men.

At the beginning of the semester, the researchers administered the DOSC to the students following the guidelines described in the DOSC manual (Michael, Smith, & Michael, 1989). At the end of the semester, they obtained each student’s first-semester GPA from university records. When they analyzed the data for evidence of reliability and validity, the DOSC showed high internal consistency, but scores on the DOSC did not predict GPA.

Did something go wrong? One conclusion is that self-concept as measured by the DOSC is unrelated to GPA. However, if the study or the measures were somehow flawed, the predictive evidence of validity of the DOSC might have gone undetected. The researchers suggested that perhaps academic self-concept lacks stability during students’ first semester. Although the internal reliability of the DOSC was established, the researchers did not measure the test–retest method reliability of the test. Therefore, this possibility cannot be ruled out. The researchers also suggested that GPA might be an unreliable criterion.

Could restriction of range have caused the validity of the DOSC to go undetected? This is a distinct possibility for two reasons. First, for this study the researchers chose only those students who were at risk for experiencing academic difficulties. Because the unrestricted population of students also contains those who are expected to succeed, the researchers might have restricted the range of both the test and the criterion. Second, the students in the study enrolled in programs to help them become successful academically. Therefore, participating in the programs might have enhanced the students’ academic self-concept.

This study demonstrates two pitfalls that researchers designing predictive studies must avoid. Researchers must be careful to include in their studies participants who represent the entire possible range of performance on both the test and the criterion. In addition, they must design predictive studies so that participants are unlikely to change over the course of the study in ways that affect the abilities or traits that are being measured.

Barrett, Phillips, and Alexander (1981) compared the two methods for determining evidence of validity based on predictor-criteria relationships in an organizational setting using cognitive ability tests. They found that the two methods provide similar results.

How Validity Differs From Reliability

As you recall from Chapter 6, reliability refers to the consistency of test results and derives from two factors: reliability within the test (internal consistency) and reliability across time (test–retest reliability). The reliability coefficient provides a quantitative estimate of a test’s consistency of measurement. A yardstick, for example, is a reliable or consistent measuring instrument because each
time it measures an item it gives the same answer. In addition, a yardstick is an appropriate measure for measuring distance, and therefore we can say the yardstick has validity as a distance measure. On the other hand, a yardstick is not a valid measure of intelligence. There is no relation between height and intelligence, and therefore we would not use a yardstick to predict IQ. This example demonstrates the principle that reliability and validity are two separate issues. A psychological test might be reliable but not valid, as in the case of the yardstick trying to predict intelligence. Reliability is a characteristic of the test itself, and validity depends on the inferences that are going to be made from the test scores.

While reliability and validity are different concepts, there is a relationship between them when it comes to gathering evidence of validity through test-criterion relationships. Mathematically, the square root of the reliability of a test will set the upper limit of the validity coefficient of a test. So if a test has a relatively low reliability of .64, the maximum correlation that the test could have with any criteria is .8. The point is that while a test can be reliable but not valid for a particular use, a test cannot be valid for any use if the scores are very unreliable.

**FOR YOUR INFORMATION BOX 8.3**

**Diagnosing Alcohol Abuse at a Veterans Affairs Medical Center**

Watson and colleagues (1996) administered two self-report alcoholism measures to 118 volunteers recruited from chemical dependency or psychiatric wards at a Veterans Affairs medical center. At roughly the same time, the researchers asked the volunteers to complete the criterion measure, a computerized version of the Diagnostic Interview Schedule (C-DIS; Blouin, 1987). The C-DIS asks questions that reflect the revised third edition of the *Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R; American Psychiatric Association, 1987)*, a reference book for psychologists and psychiatrists that lists symptoms of various mental disorders. The researchers chose the C-DIS as a criterion because it has shown high test–retest reliability and good evidence of validity based on content.

Correlations of the self-report tests with the criterion, the C-DIS, were .75 for the first self-report alcoholism measure and .60 for the second self-report alcoholism measure. These data suggest that the self-report measures of alcoholism have high evidence of validity using a concurrent study and are appropriate tools for diagnosis of alcohol dependency. (Later in this chapter, we describe how to interpret validity coefficients to provide evidence of validity based on test-criteria relationships.)

**INTERIM SUMMARY 8.1**

**DEMONSTRATING EVIDENCE OF VALIDITY FROM TEST-CRITERIA RELATIONSHIPS**

The two basic methods for showing relation between a test and independent events or behaviors (the criterion) are as follows:

- **Predictive Method**—Established by correlating test scores taken at one time with scores on a criterion measure obtained at a later date, usually months later. This method establishes that the test provides information about events in the future.
- **Concurrent Method**—Established by correlating test scores and criterion scores obtained at approximately the same time. This method establishes that the test can provide information about independent events or behaviors in the present.
Selecting a Criterion

A criterion is an evaluative standard that researchers use to measure outcomes such as performance, attitude, or motivation. Evidence of validity derived from test-criteria relationships provides evidence that the test relates to some behavior or event that is independent of the psychological test. As you recall from For Your Information Box 8.1, the researchers at Brigham Young University constructed two criteria—a questionnaire and classifications on marital satisfaction and marital stability—to demonstrate evidence of validity of the PREP-M.

In a business setting, employers use pre-employment tests to predict how well an applicant is likely to perform a job. In this case, supervisors’ ratings of job performance can serve as a criterion that represents performance on the job. Other criteria that represent job performance include accidents on the job, attendance or absenteeism, disciplinary problems, training performance, and ratings by peers—other employees at the work site. None of these measures can represent job performance perfectly, but each provides information on important characteristics of job performance.

Objective and Subjective Criteria

Criteria for job performance fall into two categories: objective and subjective. An objective criterion is one that is observable and measurable, such as the number of accidents on the job, the number of days absent, or the number of disciplinary problems in a month. A subjective criterion is based on a person’s judgment. Supervisor and peer ratings are examples of subjective criteria.

<table>
<thead>
<tr>
<th>Table 8.1 Common Criteria</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Educational settings</strong></td>
</tr>
<tr>
<td>Grade point average</td>
</tr>
<tr>
<td>Withdrawal or dismissal</td>
</tr>
<tr>
<td>Teacher’s recommendations</td>
</tr>
<tr>
<td><strong>Clinical settings</strong></td>
</tr>
<tr>
<td>Diagnosis</td>
</tr>
<tr>
<td>Behavioral observation</td>
</tr>
<tr>
<td>Self-report</td>
</tr>
<tr>
<td><strong>Organizational settings</strong></td>
</tr>
<tr>
<td>Units produced</td>
</tr>
<tr>
<td>Number of errors</td>
</tr>
<tr>
<td>Ratings of performance</td>
</tr>
</tbody>
</table>
Each has advantages and disadvantages. Well-defined objective criteria contain less error because they are usually tallies of observable events or outcomes. Their scope, however, is often quite narrow. For instance, dollar volume of sales is an objective criterion that might be used to measure a person's sales ability. This number is easily calculated, and there is little chance of disagreement on its numerical value. It does not, however, take into account a person's motivation or the availability of customers. On the other hand, a supervisor's ratings of a person's sales ability may provide more information on motivation, but in turn ratings are based on judgment and might be biased or based on information not related to sales ability, such as expectations about race or gender. Table 8.1 lists a number of criteria used in educational, clinical, and organizational settings.

### Does the Criterion Measure What It Is Supposed to Measure?

The concept of validity evidence based on content (addressed in Chapter 7) also applies to criteria. Criteria must be representative of the events they are supposed to measure. Criterion scores have evidence of validity to the extent that they match or represent the events in question. Therefore, a criterion of sales ability must be representative of the entire testing universe of sales ability. Because there is more to selling than just having the highest dollar volume of sales, several objective criteria might be used to represent the entire testing universe of sales ability. For instance, we might add the number of sales calls made each month to measure motivation and add the size of the target population to measure customer availability.

Subjective measures such as ratings can often demonstrate better evidence of their validity based on content because the rater can provide judgments for a number of dimensions specifically associated with job performance. Rating forms are psychological measures, and we expect them to be reliable and valid, as we do for any measure. We estimate their reliability using the test–retest or internal consistency method, and we generate evidence of their validity by matching their content to the knowledge, skills, abilities, or other characteristics (such as behaviors, attitudes, personality characteristics, or other mental states) that are presumed to be present in the test takers. (Chapters 10 and 15 contain more information on various types of rating scales and their uses in organizations.)

By reporting the reliability of their criteria, researchers provide us with information on how accurate the outcome measure is. As you may have noticed, the researchers at Brigham Young University (Holman et al., 1989) who conducted the study on the predictive validity of the PREP-M reported high reliability for their questionnaire, which was their subjective criterion. Likewise, the researchers at the Veterans Affairs medical center (Watson et al., 1996) chose the computerized version of the Diagnostic Interview Schedule (C-DIS) as a criterion because it reflected the Diagnostic and Statistical Manual of Mental Disorders (DSM-III-R; American Psychiatric Association, 1987) diagnosis of alcohol dependency—an indicator that the C-DIS showed evidence of validity based on content.

Sometimes criteria do not represent all of the dimensions in the behavior, attitude, or event being measured. When this happens, the criterion has decreased evidence of validity based on its content because it has underrepresented some important characteristics. If the criterion measures more dimensions than those measured by the test, we say that **criterion contamination** is present. For instance, if one were looking at the test-criterion relationship of a test of sales aptitude, a convenient criterion might be the dollar volume of sales made over some period of time. However, if the dollar volume of sales of a new salesperson reflected both his or her own sales as well as sales that resulted from the filling of back orders sold by the former salesperson, the criterion would be considered contaminated.
Calculating and Evaluating Validity Coefficients

As you recall from Chapter 5, the correlation coefficient is a quantitative estimate of the linear relationship between two variables. In validity studies, we refer to the correlation coefficient between the test and the criterion as the validity coefficient and represent it in formulas and equations as $r_{xy}$. The $x$ in the subscript refers to the test, and the $y$ refers to the criterion. The validity coefficient represents the amount or strength of the evidence of validity based on the relationship of the test and the criterion.

Validity coefficients must be evaluated to determine whether they represent a level of validity that makes the test useful and meaningful. This section describes two methods for evaluating validity coefficients and how researchers use test-criterion relationship information to make predictions about future behavior or performance.

Tests of Significance

A validity coefficient is interpreted in much the same way as a reliability coefficient, except that our expectations for a very strong relationship are not as great. We cannot expect a validity coefficient to have as strong a relationship with another variable (test-criterion evidence of validity) as it does with itself (reliability). Therefore, we must evaluate the validity coefficient by using a test of significance and by examining the coefficient of determination.
The first question to ask about a validity coefficient is, “How likely is it that the correlation between the test and the criterion resulted from chance or sampling error?” In other words, if the test scores (for example, SAT scores) and the criterion (for example, college GPA) are completely unrelated, then their true correlation is zero. If we conducted a study to determine the relationship between these two variables and found that the correlation was .4, one question that we need to ask is, “What is the probability that our study would have yielded the obtained correlation by chance alone, even if the variables were truly unrelated?” If the probability that the correlation occurred by chance is low—less than 5 chances out of 100 (\( p < .05 \))—we can be reasonably sure that the test and its criterion (in this example, SAT scores and college GPA) are truly related. This process is called a test of significance. In statistical terms, for this example we would say that the validity coefficient is significant at the .05 level. In organizational settings, it can be challenging for validity studies to have statistically significant results at \( p < .05 \) because of small sample sizes and criterion contamination.

Because larger sample sizes reduce sampling error, this test of significance requires that we take into account the size of the group (\( N \)) from which we obtained our data. Appendix E can be used to determine whether a correlation is significant at varying levels of significance. To use the table in Appendix E, calculate the degrees of freedom (\( df \)) for your correlation using the formula \( df = N - 2 \), and then determine the probability that the correlation occurred by chance by looking across the row associated with those degrees of freedom. The correlation coefficient you are evaluating should be larger than the critical value shown in the table. You can determine the level of significance by looking at the column headings. At the level where your correlation coefficient is smaller than the value shown, the correlation can no longer be considered significantly different from zero. For Your Information Box 8.5 provides an example of this process.

When researchers or test developers report a validity coefficient, they should also report its level of significance. You might have noted that the validity coefficients of the PREP-M (reported earlier in this chapter) are followed by the statements \( p < .01 \) and \( p < .05 \). This information tells the test user that the likelihood a relationship was found by chance or as a result of sampling error was less than 5 chances out of 100 (\( p < .05 \)) or less than 1 chance out of 100 (\( p < .01 \)).

If the correlation between the test and the predictor is not as high as the critical value shown in the table, we can say that the chance of error associated with the test is above generally accepted levels. In such a case, we would conclude that the validity coefficient does not provide sufficient evidence of validity.

**The Coefficient of Determination**

Another way to evaluate the validity coefficient is to determine the amount of variance that the test and the criterion share. We can determine the amount of shared variance by squaring the validity coefficient to obtain \( r^2 \)—called the coefficient of determination. For example, if the correlation \( (r) \) between a test and a criterion is .30, the coefficient of determination \( (r^2) \) is .09. This means that the test and the criterion have 9% of their variance in common. Larger validity coefficients represent stronger relationships with greater overlap between the test and the criterion. Therefore, if \( r = .50 \), then \( r^2 = .25 \)—or 25% shared variance.

We can calculate the coefficient of determination for the correlation of husbands’ scores on the PREP-M and the questionnaire on marital satisfaction and stability. By squaring the original coefficient, .44, we
obtain the coefficient of determination, \( r^2 = .1936 \). This outcome means that the predictor, the PREP-M, and the criterion, the questionnaire, shared (or had in common) approximately 19\% of their variance.

Unadjusted validity coefficients rarely exceed .50. Therefore, you can see that even when a validity coefficient is statistically significant, the test can account for only a small portion of the variability in the criterion. The coefficient of determination is important to calculate and remember when using the correlation between the test and the criterion to make predictions about future behavior or performance.
How Confident Can We Be About Estimates of Validity?

Conducting one validity study that demonstrates a strong relationship between the test and the criterion is the first step in a process of validation, but it is not the final step. Studies that provide evidence of a test’s validity should continue for as long as the test is being used. No matter how well designed the validation study is, elements of chance, error, and situation-specific factors that can over- or underinflated the estimate of validity are always present. Ongoing investigations of validity include cross-validation (where the results that are obtained using one sample are used to predict the results on a second, similar sample) and meta-analyses (where the results from many studies are statistically combined to provide a more error-free estimate of validity). Psychologists also inquire about whether validity estimates are stable from one situation or population to another—a question of validity generalization. (Chapters 12 and 15 address these issues in more detail.)

INTERIM SUMMARY 8.2
EVALUATING VALIDITY COEFFICIENTS

- Tests of significance establish the likelihood that a correlation observed between a test and a criterion was obtained by chance.
- The coefficient of determination describes how much variance a test and a criterion share.
- Ongoing investigations of validity include cross-validation and meta-analyses.
- Psychologists also inquire about whether validity estimates are stable from one situation or population to another—a question of validity generalization.

Using Validity Information to Make Predictions

When a relationship can be established between a test and a criterion, we can use test scores from other individuals to predict how well those individuals will perform on the criterion measure. For example, some universities use students’ scores on the SAT to predict the students’ success in college. Organizations use job candidates’ scores on pre-employment tests that have demonstrated evidence of validity to predict those candidates’ scores on the criteria of job performance.

Linear Regression

We use the statistical process called linear regression when we use one set of test scores (X) to predict one set of criterion scores (Y’). To do this, we construct the following linear regression equation:

\[ Y’ = a + bX \]

where
\( Y’ \) = the predicted score on the criterion
\( a \) = the intercept
\( b \) = the slope
\( X \) = the score the individual made on the predictor test
This equation actually provides a predicted score on the criterion \((Y')\) for each test score \((X)\). When the \(Y'\) values are plotted, they form the linear regression line associated with the correlation between the test and the criterion.

We can calculate the slope \((b)\) of the regression line—the expected change in one unit of \(Y\) for every change in \(X\)—using the following formula:

\[
b = r \frac{S_Y}{S_X}
\]

where

- \(r\) = the correlation coefficient
- \(S_x\) = the standard deviation of the distribution of \(X\)
- \(S_y\) = the standard deviation of the distribution of \(Y\)

The intercept is the place where the regression line crosses the \(y\)-axis. The intercept \((a)\) is calculated using the following formula:

\[
a = \bar{Y} - b \bar{X}
\]

where

- \(\bar{Y}\) = the mean of the distribution of \(Y\)
- \(b\) = the slope
- \(\bar{X}\) = the mean of the distribution of \(X\)

For Your Information Box 8.6 shows the calculation of a linear regression equation and how it is used to predict scores on a criterion.

The process of using correlated data to make predictions is also important in clinical settings. For Your Information Box 8.7 describes how clinicians use psychological test scores to identify adolescents at risk for committing suicide.

---

**FOR YOUR INFORMATION BOX 8.6**

**Making Predictions With a Linear Regression Equation**

Research suggests that academic self-efficacy (ASE) and class grades are related. We have made up the following data to show how we could use the scores on an ASE test to predict a student’s grade. (Note: Our fake data set is small to facilitate this illustration.)

For instance, we can ask the question, “If a student scores 65 on the ASE test, what course grade would we expect the student to receive?” We have assigned numbers to each grade to facilitate this analysis, therefore, 1 = D, 2 = C, 3 = B, and 4 = A.
Chapter 8: How Do We Gather Evidence of Validity Based on a Test’s Relationships With External Criteria? 225

FOR YOUR INFORMATION BOX 8.7

Evidence of Validity of the Suicide Probability Scale Using the Predictive Method

Although the general incidence of suicide has decreased during the past two decades, the rate for people between 15 and 24 years old has tripled. Suicide is generally considered to be the second or third most common cause of death among adolescents, even though it is underreported (O’Connor, 2008).

If young people who are at risk for committing suicide or making suicide attempts can be identified, greater vigilance is likely to prevent such actions. Researchers at Father Flanagan’s Boys’ Home, in Boys Town, Nebraska, conducted a validity study using the predictive method for the Suicide Probability Scale (SPS) that provided encouraging results for predicting suicidal behaviors in adolescents (Larzelere, Smith, Batenhorst, & Kelly, 1996). The SPS contains 36 questions that assess suicide risk, including thoughts about suicide, depression, and isolation. The researchers administered the SPS to 840 boys and girls when they were admitted to the Boys Town residential treatment program from 1988 through 1993. The criteria for this study were the numbers of suicide attempts, suicide verbalizations, and self-destructive behaviors recorded in the program’s Daily Incident Report completed by supervisors of the group homes. (The interrater reliabilities for reports of verbalizations and reports of self-destructive behaviors were very high at .97 and .89, respectively. The researchers were unable to calculate a reliability estimate for suicide attempts because only one attempt was recorded in the reports they selected for the reliability analysis.)

(Continued)
Multiple Regression

Complex criteria, such as job performance and success in graduate school, are often difficult to predict with a single test. In these situations, researchers use more than one test to make an accurate prediction. An expansion of the linear regression equation helps in this situation.

We use the statistical process of multiple regression for predicting a criterion \( Y' \) using more than one set of test scores \( (X_1, X_2, \ldots, X_n) \). The multiple regression equation that incorporates information from more than one predictor or test is as follows:

\[
Y' = a + b_1 X_1 + b_2 X_2 + b_3 X_3 \ldots + b_n X_n
\]

where

\( Y' \) = the predicted score on the criterion

\( a \) = the intercept

\( b \) = the slope of the regression line and amount of variance the predictor contributes to the equation, also known as beta \((\beta)\)

\( X \) = the predictor

The numbers following \( b \) and \( X \) indicate the test from which the information was drawn.

In multiple regression, there is one criterion \( (Y) \), but there are several predictors \( (X) \). To describe this relationship, we use a multiple correlation \((R)\). We can evaluate \( R \) by calculating the coefficient of multiple determination \((R^2)\), which indicates the proportion of variance in the criterion accounted for by all of the predictors. The \( R \) can also be subjected to a test of significance to determine whether it is significantly different from zero. We can also calculate the contribution that each predictor \( (X) \) by itself
contributes to the total explained variance ($R^2$). We must bear in mind, however, that when a predictor is first in the multiple regression equation, it is likely to explain more variance than when the predictor is last. This is the case when the predictors share variance (are correlated) with each other as well as with the criterion.

Because each combination of $a + bX$ in the preceding equation indicates the presence of a different regression line, the multiple regression equation is difficult to plot. It works in theory, however, much the same as does the linear regression equation. The value of each $b$ (partial regression coefficient) indicates how many units $Y$ increases for every increase in $X$, and therefore each $b$ indicates the contribution that each predictor makes in determining a predicted score on the criterion ($Y'$).

Chibnall and Detrick (2003) published a study that examined the usefulness of three personality inventories—the Minnesota Multiphasic Personality Inventory-2 (MMPI-2), the Inwald Personality Inventory (IPI; an established police officer screening test), and the Revised NEO Personality Inventory (NEO PI-R)—for predicting the performance of police officers. They administered the inventories to 79 police recruits and compared the test scores with two criteria: academic performance and physical performance. Tables 8.2 and 8.3 show the outcomes of the study for the academic performance criterion.

On the Web Box 8.1 guides you in researching the Inwald Personality Test. Can you find information on evidence of validity based on test-criteria relationships for this test?

When the researchers entered recruit class, marital status, and race first, they accounted for 20% of the prediction of academic performance. In the second step, the researchers entered the test scores from the IPI. The table shows the contribution of the inventory dimensions that contributed significantly to the prediction. Together, the three dimensions of the IPI contributed 16% of the prediction ($R^2\Delta$). In the third step, the researchers entered two dimensions of the MMPI-2, and together they accounted for 8% of the prediction. Finally, the researchers entered three dimensions of the NEO PI-R, and together they accounted for another 11% of the prediction. Altogether, the demographic characteristics and the three inventories accounted for 55% of the prediction of academic performance ($R^2$).

### Table 8.2 Regression Model for Predicting Academic Performance ($R^2 = .55$)

<table>
<thead>
<tr>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Demographic Variables</strong></td>
<td><strong>IPI</strong></td>
<td><strong>MMPI-2</strong></td>
<td><strong>NEO PI-R</strong></td>
</tr>
<tr>
<td><strong>Predictor</strong></td>
<td>$\beta$</td>
<td>$\beta$</td>
<td>$\beta$</td>
</tr>
<tr>
<td>Recruit class</td>
<td>.32</td>
<td>Trouble law</td>
<td>−.26</td>
</tr>
<tr>
<td>Marital status</td>
<td>.27</td>
<td>Antisocial</td>
<td>−.35</td>
</tr>
<tr>
<td>Race</td>
<td>.24</td>
<td>Obsessiveness</td>
<td>.29</td>
</tr>
<tr>
<td>$R^2\Delta$</td>
<td>.20</td>
<td>$R^2\Delta$</td>
<td>.16</td>
</tr>
</tbody>
</table>


**NOTE:** Step refers to the introduction of a predictor into the regression equation for predicting academic performance. The predictors are the individual demographic characteristics or the subscales that reached significance. $R^2\Delta$ is the percentage of prediction of academic performance contributed by each predictor.
Demographic characteristics did not contribute to the prediction of physical performance. Only the NEO PI-R predicted 20% ($R^2 = .20$) of the variance for physical performance. The dimensions of the NEO PI-R that accounted for variance in the criterion were Fantasy ($\beta = .47$), Deliberation ($\beta = .26$), and Feelings ($\beta = .25$).

Two studies (Kolb, Race, & Seibert, 2000; Meyer, Woodard, & Suddick, 1994) illustrate the usefulness of multiple regression in providing evidence of validity from test-criterion relationships. Kolb and colleagues evaluated the evidence of validity of a questionnaire designed to assess the satisfaction of psychiatric inpatients. They distributed the questionnaire to inpatients in 37 psychiatric units. Their data analysis showed that all coefficient alphas for the six scales in the questionnaire were above .74. The results of a multiple regression analysis, in which the six scales were used as predictors and the overall quality-of-care ratings were used as the criterion, are displayed in Table 8.4.

### Table 8.3
Range of Variance Explained ($R^2$) by Each Inventory at Each Step in the Regression Models That Predict Academic Performance

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Step 1</th>
<th>Step 2</th>
<th>Step 3</th>
<th>Step 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>.20</td>
<td>—</td>
<td>.14 to .16</td>
<td>.11 to .16</td>
</tr>
<tr>
<td>NEO PI-R</td>
<td>—</td>
<td>.20</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>IPI</td>
<td>—</td>
<td>.16</td>
<td>.00 to .05</td>
<td>.03 to .03</td>
</tr>
<tr>
<td>MMPI-2</td>
<td>—</td>
<td>.14</td>
<td>.08 to .10</td>
<td>.00 to .08</td>
</tr>
</tbody>
</table>


**NOTE:** Step refers to the introduction of a predictor into the regression equation for predicting academic performance. MMPI-2 = Minnesota Multiphasic Personality Inventory-2, IPI = Inwald Personality Inventory, NEO PI-R = Revised NEO Personality Inventory.

### Table 8.4
Power of Inpatient Psychiatric Scales to Predict Quality-of-Care Ratings

<table>
<thead>
<tr>
<th>Scale</th>
<th>Pearson Product–Moment Correlation Coefficient</th>
<th>Partial $R^2$</th>
<th>$R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Staff</td>
<td>.667</td>
<td>.422</td>
<td>.422</td>
</tr>
<tr>
<td>Nonclinical Services</td>
<td>.642</td>
<td>.199</td>
<td>.541</td>
</tr>
<tr>
<td>Program Components/Activities</td>
<td>.570</td>
<td>.026</td>
<td>.576</td>
</tr>
<tr>
<td>Psychiatric Care</td>
<td>.534</td>
<td>.006</td>
<td>.573</td>
</tr>
<tr>
<td>Medical Outcome</td>
<td>.445</td>
<td>.002</td>
<td>.575</td>
</tr>
<tr>
<td>Patient Education</td>
<td>.425</td>
<td>.002</td>
<td>.577</td>
</tr>
</tbody>
</table>


**NOTE:** Stepwise multiple regression is based on a sample of 1,351 patients. Partial $R^2$ = additional variance explained at each step. The proportion of variance explained at each step is affected by the order in which these scales are entered into the model. Modifying the order of inclusion could result in proportional changes in this variance. $R^2$ = total variance explained at each step.
Meyer and colleagues (1994) conducted a validity study using the predictive method of two of the Descriptive Tests of Mathematics Skills (DTMS) of the College Entrance Examination Board. The researchers’ university requires elementary education majors to pass the Arithmetic Skills and Elementary Algebra Skills tests of the DTMS. The university also requires elementary education majors to take an upper-division mathematics concepts and structures course. In this study, the Arithmetic Skills and Elementary Algebra Skills tests were the predictors, and the grade received in the upper-division math course became the criterion in a study of 60 elementary education majors.

In their data analysis, Meyer and colleagues (1994) found evidence of internal consistency for the two tests: Arithmetic Skills (.72) and Elementary Algebra Skills (.77). Course grades (the criterion) correlated with Arithmetic Skills at .48 ($p < .001$) and correlated with Elementary Algebra Skills at .49 ($p < .001$).

As you can see, Table 8.4 shows the correlation of each scale with the criterion in the Pearson's Product–Moment Correlation Coefficient column. The Partial $R^2$ column shows the increase in percentage of variance that each scale provides. The partial $R^2$ values will vary when the order in which the scales appear in the formula changes. The $R^2$ column (the coefficient of determination for $R$) shows the percentage of variance accounted for at each step in the formula. Using just one scale on the questionnaire accounts for 42.2% of the variance in the overall score, but using all six scales accounts for 57.7% of the variance. We would expect such an outcome because each scale provides information about a different aspect of inpatient satisfaction. Although Kolb and colleagues’ (2000) report of the results of their data analysis is clear, they do not provide information on how the criterion was developed, measured, or calculated. As you know from our discussion of criterion validity, we cannot rely on the results of studies that do not provide information on the criterion measure.

ON THE WEB BOX 8.1

Conducting Research

This book contains detailed information on the revised Minnesota Multiphasic Personality Inventory-2 (see Chapter 11) and the Revised NEO Personality Inventory (see Chapter 14). Using professional databases, such as ProQuest, Mental Measurements Yearbook, Tests in Print, and Education Resources Information Center (ERIC), research the Inwald Personality Inventory. Chapter 1 provides information on using the Mental Measurements Yearbook and Tests in Print. (Hint: Detrick, Chibnall, & Rosso, 2001, is also relevant to this exercise.)

Find the following:

- Test purpose
- Target audience
- Reliability and validity estimates

Taking into account the information you found on the Inwald Personality Inventory, would you recommend that organizations use the test for selection purposes? Back up your recommendation with the information you found.
Then the researchers used both tests in a multiple regression equation to predict course grades. They found a multiple $R$ of .54—higher than either test predicted independently—that accounted for approximately 29% of the variance in the two tests and course grades ($R^2 = .286$). Most of the variance, however, was predicted by one test, Elementary Algebra, which accounted for 24% of the variance ($R^2 = .241$). The multiple $R$ for Elementary Algebra alone and the $R$ for both tests together were not significantly different.

Meyer and colleagues (1994) concluded that each test showed evidence of validity and could be used as a predictor of math performance for elementary education majors. In addition, they suggested simplifying the assessment program by using one test instead of two tests because the tests supplied redundant information about the criterion (course grades).

**Ethical Issues Associated With Test Validation**

Decisions based on test predictions have far-reaching consequences. Each day in the United States and other industrialized nations, individuals are hired or rejected by organizations that base their decisions on employment tests. Therefore, test users must rely on validity studies to ensure that the tests they use make accurate predictions.

Educators also use test results to admit or refuse admittance to programs based on predictions made by educational ability tests, and clinicians use tests to screen clients for residential or outpatient treatment and to admit them to specific treatment programs based on diagnoses made by tests. As you recall from Chapter 3, Michael Elmore’s teacher used Michael’s score on an intelligence test to decide that Michael was borderline retarded even though he was not.

The increasing diversity of the population in the United States presents questions about the suitability of tests for students, clinical clients, employees, and job seekers from various minorities. Also, as the use of psychological assessment spreads to countries whose primary languages are not English, questions arise concerning translations and norms associated with translations of standardized tests developed in the United States and Great Britain. When test takers are members of minorities, especially those who do not speak standard English as their primary language, test users must be aware of test bias and how it affects test validity. Cofresi and Gorman (2004) assert that the test users’ responsibilities include testing assessment tools to ensure that they are valid for the minority population who will be the test takers. Appropriate assessments, they emphasize, should be free of questions that require a specific cultural background (for example, knowledge of sports, holidays, or foods; etiquette related to specific cultures, races, or religions). The use of test norms that were developed without inclusion of the minority being tested is likely to be inappropriate for interpreting test scores of the minority group.

With each decision, the test user is ethically and morally responsible for ascertaining that the test instrument shows acceptable evidence of reliability and validity. In some cases, such as employment decisions in which there is discrimination against protected classes, test users might be held legally liable for improper test use. Test users rely on researchers and test publishers to provide full information about tests. Test publishers have a particular responsibility to prevent test misuse by making test manuals and validity information available and accessible before test users purchase their tests. Publishers should also refuse to provide test materials to persons who do not have testing credentials.
or who are likely to misuse the tests. Finally, psychologists in general have a responsibility to increase public awareness about the importance of test reliability and validity so that the public can understand the role that tests play in decisions that affect individuals’ lives.

Evidence of validity based on test-criteria relations—the extent to which a test is related to independent behavior or events—is one of the major methods for obtaining evidence of test validity. The usual method for demonstrating this evidence is to correlate scores on the test with a measure of the behavior we wish to predict. This measure of independent behavior or performance is called the criterion.

Evidence of validity based on test-criteria relations depends on evidence that the scores on the test correlate significantly with an independent criterion—a standard used to measure some characteristic of an individual, such as a person’s performance, attitude, or motivation. Criteria may be objective or subjective, but they must be reliable and valid. There are two methods for demonstrating evidence of validity based on test-criteria relations: predictive and concurrent.

We use correlation to describe the relationship between a psychological test and a criterion. In this case, the correlation coefficient is referred to as the validity coefficient. Psychologists interpret validity coefficients using tests of significance and the coefficient of determination.

Either a linear regression equation or a multiple regression equation can be used to predict criterion scores from test scores. Predictions of success or failure on the criterion enable test users to use test scores for making decisions about hiring.

Finally, decisions based on test predictions have far-reaching consequences. Researchers, test developers, test publishers, and test users are ethically and morally responsible for ascertaining that any psychological test used for making predictions and decisions shows acceptable evidence of reliability and validity. Their responsibility also extends to guarding against test misuse and to increasing public awareness about the important role that tests play in test takers’ lives.
Learning Activities

The following are some learning activities you can engage in to support the learning objectives for this chapter.

<table>
<thead>
<tr>
<th>Learning Objectives</th>
<th>Study Tips and Learning Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>After completing your study of this chapter, you should be able to do the following:</td>
<td>The following study tips will help you meet these learning objectives:</td>
</tr>
</tbody>
</table>
| Identify validity evidence of a test based on its relationships to external criteria (criterion-related validity), and describe two methods for obtaining this evidence. | • Make a chart of the similarities and differences between predictive and concurrent methods.  
• Be sure to include when the criterion is measured and its purpose. |
| Read and interpret validity studies. | • Do Additional Learning Activities 1, 2, and 3, which follow this table. |
| Discuss how restriction of range occurs and its consequences. | • Review For Your Information Box 8.2.  
• Explain the reasons why restriction of range may have occurred in this study. |
| Describe the differences between evidence of validity based on test-criteria relationships and evidence of validity based on test content. | • Make a chart of the similarities and differences.  
• Be sure to include the method of validation and method of statistical analysis. |
| Describe the difference between reliability and validity. | • List the differences between reliability and validity. |
| Define and give examples of objective and subjective criteria, and explain why criteria must be reliable and valid. | • Study Table 8.1.  
• See whether you can add other examples of objective and subjective criteria to the list.  
• Check your additions with a classmate or your instructor.  
• Do Additional Learning Activity 2, which follows this table. |
| Calculate a validity coefficient and the coefficient of determination, and then conduct a test of significance for the validity coefficient. | • Work through the example in For Your Information Box 8.5 to see how the test of significance works.  
• Consider the following: Would the correlation have been significant at \( p < .001 \)?  
• Do Additional Learning Activities 3, 5, and 6, which follow this table. |
| Explain the concept of regression, calculate and interpret a linear regression formula, and interpret a multiple regression formula. | • Use your calculator to work through the example of making a prediction with linear regression in For Your Information Box 8.6.  
• Write a short essay that explains the difference between linear regression and multiple regression.  
• Do Additional Learning Activities 3 and 4, which follow this table. |
ADDITIONAL LEARNING ACTIVITIES

1. Interpreting Validity Studies. Read the summarized published criterion-related validation studies below. For each summary, identify the following elements:

   (1) Predictor(s)
   (2) Criterion
   (3) Evidence of validity examined
   (4) Validity coefficient and its strength
   (5) Type of reliability
   (6) Reliability of the test scores (where given)

   a. College Students’ Recent Life Experiences. Researchers administered to 216 undergraduate students (in the same time period) the Inventory of College Students’ Experiences and a measure of daily hassles. The total coefficient alpha was .92 for the inventory and .96 for the measure of daily hassles. The inventory correlated with the measure of daily hassles at .76 (p < .001). (Adapted from Osman, Barrios, Longnecker, & Osman, 1994)

   b. The Pilot Personality. Test scores for the Eysenck Personality Inventory and Cattell’s 16 Personality Factor Questionnaire were obtained for male army applicants for flyer training. Forms A and B were used for each test, and the correlations between forms for the same test ranged from .39 to .85. Some of the men entered flying school several years after taking the tests. The correlations of the subscales on the two tests with training outcome (pass or fail) averaged approximately .20. (Adapted from Bartram, 1995)

   c. Computer Aptitude and Computer Anxiety. Researchers gave 162 students enrolled in computer courses a test that measured computer anxiety and another test that measured computer aptitude. Both tests were given at the beginning of the course. Student performance in the course was measured by the grades the students earned in the course. Computer aptitude correlated with course grade at .41 (p < .01) for one course and at .13 (ns; note that ns stands for “not significant”) for the other course. Correlations of computer anxiety and course grade were .01 and .16 (ns). (Adapted from Szajna, 1994)

2. Objective and Subjective Criteria. A number of criteria are listed in the table below. Decide what type of criterion each is, and mark either “Objective” or “Subjective” in the Type column. Discuss the advantages and disadvantages you think might be associated with using each of the criteria.

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratings of training success</td>
<td></td>
</tr>
<tr>
<td>Letters of reference</td>
<td></td>
</tr>
<tr>
<td>Completion of a work sample (pass/fail)</td>
<td></td>
</tr>
<tr>
<td>Ratings based on a work sample</td>
<td></td>
</tr>
<tr>
<td>Annual salary</td>
<td></td>
</tr>
<tr>
<td>Number of alcoholic drinks</td>
<td></td>
</tr>
<tr>
<td>Self-ratings of drug use</td>
<td></td>
</tr>
<tr>
<td>Course grade</td>
<td></td>
</tr>
<tr>
<td>Number of weeks in therapy</td>
<td></td>
</tr>
<tr>
<td>Therapist’s estimate of weekly progress</td>
<td></td>
</tr>
</tbody>
</table>
3. **Interpreting Statistics.** The following table contains symbols that stand for statistics used in validation studies. Identify each, and explain when to use it and what it means.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td></td>
</tr>
<tr>
<td>$r^2$</td>
<td></td>
</tr>
<tr>
<td>$R^2$</td>
<td></td>
</tr>
<tr>
<td>$\gamma'$</td>
<td></td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
</tr>
<tr>
<td>$X$</td>
<td></td>
</tr>
<tr>
<td>$\rho$</td>
<td></td>
</tr>
<tr>
<td>$df$</td>
<td></td>
</tr>
</tbody>
</table>

4. **A Case for Incremental Validity.** Return to the description of the study by Chibnall and Detrick (2003) described in the section on multiple regression. Reread the information about this study, including Tables 8.3 and 8.4, and then answer the following questions:

1. What are the predictors and criteria for this study?
2. What did the predictors have in common?
3. What are the strengths and weaknesses of this study?
4. Would you feel comfortable using the predictors in this study to select men and women for admission to a police academy? Why or why not?

5. **Testing the Significance of Validity Coefficients.** Complete the table below, and then express an opinion on whether each coefficient provides sufficient evidence of criterion-related validity. Test the coefficients using a two-tailed test with $p < .05$. If the degrees of freedom are not listed in the table, use the next lowest degrees of freedom listed.

<table>
<thead>
<tr>
<th>Validity Coefficient</th>
<th>Size of Sample</th>
<th>$df$</th>
<th>Critical $r$ From Appendix E</th>
<th>Significant at $p &lt; .05$?</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>.23</td>
<td>40</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.43</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.33</td>
<td>50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.22</td>
<td>1,000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.50</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6. **Calculating a Validity Coefficient.** The following simulated data were collected when an organization tested job candidates before they were hired and then evaluated them using supervisor ratings after they had worked for six months. Calculate the validity coefficient for these data. What kind of evidence of validity does it represent? Now suppose that the company had not hired those who scored 60 or below on the employment test. What would the validity coefficient be for the six employees only? Can you explain why the validity coefficient changed?
## Practice Questions

The following are some practice questions to assess your understanding of the material presented in this chapter.

### Multiple Choice

Choose the one best answer to each question.

1. When a test is used to predict future performance, there must be evidence of validity
   a. based on test taker's perceptions.
   b. based on test relationship with a criteria.
   c. using the predictive method.
   d. using the concurrent method.

2. Sarah conducted a study in which she correlated students’ scores on the SAT taken in high school with students’ grade point averages at the end of the first year of college. Her study was designed to find evidence of validity
   a. based on test content.
   b. using the concurrent method.
   c. based on test takers’ perceptions.
   d. using the predictive method.

3. In the study at Brigham Young University, the researchers correlated scores on the PREPAration for Marriage Questionnaire with measures of marital satisfaction and marital stability. In this study, the measures of marital satisfaction and marital stability were
   a. predictors.
   b. tests.
   c. criteria.
   d. coefficients.

4. One problem with studies of validity using the concurrent method is that there may be
   a. no evidence of validity based on test content.
   b. no criterion measure.
   c. restriction of range.
   d. low reliability.

5. Both the predictive method and the concurrent method are ways to establish evidence of
   a. validity based on test-criteria relationships.
   b. validity based on test content.
   c. validity based on the perceptions of the test takers.
   d. both validity and reliability.

6. A major difference between the predictive method and the concurrent method is the ______ the criterion is measured.
   a. place where
   b. people to whom
   c. time when
   d. format in which

7. ______ is a characteristic of the test itself; ______ depends on how the test is used.
   a. Reliability; validity
   b. Validity; reliability
   c. Face validity; content validity
   d. Content validity; face validity

<table>
<thead>
<tr>
<th>Candidate</th>
<th>Employment Test</th>
<th>Supervisor Rating (1–5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abel</td>
<td>80</td>
<td>2</td>
</tr>
<tr>
<td>Bartmann</td>
<td>98</td>
<td>3</td>
</tr>
<tr>
<td>Cardoza</td>
<td>95</td>
<td>5</td>
</tr>
<tr>
<td>Dixon</td>
<td>55</td>
<td>3</td>
</tr>
<tr>
<td>Everett</td>
<td>70</td>
<td>2</td>
</tr>
<tr>
<td>Friedman</td>
<td>75</td>
<td>4</td>
</tr>
<tr>
<td>Grass</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td>Hart</td>
<td>55</td>
<td>1</td>
</tr>
<tr>
<td>Isaacs</td>
<td>90</td>
<td>2</td>
</tr>
<tr>
<td>Jensen</td>
<td>60</td>
<td>1</td>
</tr>
</tbody>
</table>
8. Sharon wanted to show evidence of validity for a test that was designed to predict reading readiness for kindergarten children. She chose as her criterion the overall score on a published standardized test of academic performance that was administered to the children after they completed first grade. What is the problem with her criterion?
   a. Low reliability
   b. Low validity
   c. Low face validity
   d. Criterion contamination

9. When we ask the question, “What is the probability that our study would have yielded the validity coefficient we are evaluating by chance alone?” we are conducting a
   a. validation study.
   b. reliability study.
   c. test of significance.
   d. linear regression.

10. Which one of the following helps us interpret a validity coefficient by telling us how much variance the predictor and the criterion share?
    a. Reliability coefficient
    b. Test of significance
    c. Content validity ratio
    d. Coefficient of determination

11. The difference between linear regression and multiple regression is the number of
    a. predictors.
    b. criteria.
    c. coefficients of determination.
    d. participants.

12. What does the linear regression formula \( Y' = a + bX \) allow us to do?
    a. Predict the value of the criterion measure associated with any test score
    b. Calculate the predictive validity of a test
    c. Provide evidence of validity based on test content
    d. Estimate the accuracy of any test score

13. When using test scores for decision making, the test user is ethically and morally responsible for ascertaining that the test shows acceptable evidence of
    a. the use of both predictive and concurrent studies.
    b. face validity and test taker acceptance.
    c. reliability and validity.
    d. reliability and face validity.

14. Who has the responsibility for preventing test misuse by making test manuals and validity information available before purchase?
    a. Test users
    b. Test takers
    c. Test publishers
    d. Test developers

15. When assessing groups that include minorities, it is preferable for the test user to
    a. always refuse to test minorities with standardized tests.
    b. not allow minority job candidates to participate in employment testing.
    c. use tests that have norms that include the minority groups being tested.
    d. warn minorities that their scores on the test may be invalid or uninterpretable.

---

**Short Answer/Essay**

Read each of the following, and consider your response carefully based on the information presented in this chapter. Write your answer to each question in two or three paragraphs.

1. What is meant by evidence of validity based on test-criterion relationships? Describe two research methods for obtaining it.

2. What is the difference between evidence of validity based on test content and evidence of validity based on test-criterion relationships? Give examples.

3. What are some challenges that organizations face when seeking to obtain predictive evidence of validity for a test?

4. Why is it important to evaluate the quality of the criterion when gathering evidence of validity of a test using the concurrent method?

5. What is the difference between reliability and validity? Give an example.

6. Discuss the difference between objective criteria and subjective criteria. Give examples of each.

7. Why do we conduct tests of significance and calculate the coefficient of determination? Explain the role of each.
8. What is the relation between correlation and linear regression? How does the process of linear regression help us make predictions?

9. Discuss ethical issues that arise when assessing test takers with diverse cultures, primary languages, races, and/or religions.

**Answer Keys**

**Additional Learning Activities**

1. Interpreting Validity Studies

<table>
<thead>
<tr>
<th>Exercise</th>
<th>(1) Predictor</th>
<th>(2) Criterion</th>
<th>(3) Validity Study Method Used</th>
<th>(4) Validity Coefficient</th>
<th>(5) Type of Reliability</th>
<th>(6) Reliability Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Inventory of college students’ experiences</td>
<td>Measure of daily hassles</td>
<td>Concurrent</td>
<td>Strong .76 ($p &lt; .001$)</td>
<td>Internal consistency, coefficient alpha</td>
<td>Inventory = .92; daily hassles = .96</td>
</tr>
<tr>
<td>B</td>
<td>Eysenck Personality Inventory and 16 Personality Factor Questionnaire</td>
<td>Training outcome (pass/fail)</td>
<td>Predictive</td>
<td>Very weak; average of coefficients approximately .20</td>
<td>Test–retest</td>
<td>Ranges between .39 and .85</td>
</tr>
<tr>
<td>C</td>
<td>Computer anxiety and computer aptitude tests</td>
<td>Course grades</td>
<td>Predictive</td>
<td>Computer aptitude = .41 ($p &lt; .01$; moderate) and .13 (ns); Computer anxiety = .01 and .16 (ns)</td>
<td>Not given</td>
<td>Not given</td>
</tr>
</tbody>
</table>

2. Objective and Subjective Criteria

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratings of training success</td>
<td>Subjective</td>
</tr>
<tr>
<td>Letters of reference</td>
<td>Subjective</td>
</tr>
<tr>
<td>Completion of a work sample (pass/fail)</td>
<td>Objective</td>
</tr>
<tr>
<td>Ratings based on a work sample</td>
<td>Subjective</td>
</tr>
<tr>
<td>Annual salary</td>
<td>Objective</td>
</tr>
<tr>
<td>Number of alcoholic drinks</td>
<td>Objective</td>
</tr>
<tr>
<td>Self-ratings of drug use</td>
<td>Subjective</td>
</tr>
<tr>
<td>Course grade</td>
<td>Subjective</td>
</tr>
<tr>
<td>Number of weeks in therapy</td>
<td>Objective</td>
</tr>
<tr>
<td>Therapist’s estimate of weekly progress</td>
<td>Subjective</td>
</tr>
</tbody>
</table>
3. Interpreting Statistics

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$R$</td>
<td>Multiple regression coefficient</td>
</tr>
<tr>
<td>$r^2$</td>
<td>Coefficient of determination for linear relationship</td>
</tr>
<tr>
<td>$R^2$</td>
<td>Coefficient of determination for multiple regression coefficient</td>
</tr>
<tr>
<td>$Y$</td>
<td>The number on the y axis that a linear regression coefficient predicts</td>
</tr>
<tr>
<td>$\beta$</td>
<td>Beta for multiple regression equation</td>
</tr>
<tr>
<td>$X$</td>
<td>Raw score</td>
</tr>
<tr>
<td>$p$</td>
<td>Probability</td>
</tr>
<tr>
<td>$df$</td>
<td>Degrees of freedom</td>
</tr>
</tbody>
</table>

4. A Case for Incremental Validity

The predictors in the Chibnall and Detrick (2003) study were the Minnesota Multiphasic Personality Inventory-2, the Inwald Personality Inventory, and the Revised NEO Personality Inventory, and the criteria were the academic performance of police officers. The information in the tables describes the prediction of the first criterion, namely, academic performance.

At first glance, the predictors are well-known personality tests. From the data in Table 8.3, we can tell that the predictors share variance among themselves and with the predictor. We know this because the contributions of the predictors vary depending on how they are arranged in the multiple regression equation.

The study's strength is that the researchers used well-known predictors with evidence of validity. The study's weaknesses are that the sample contained only 79 participants and that the study, as reported in this textbook, does not provide information on reliability and evidence of validity of the criteria.

However, we would be comfortable using the predictors in this study to select men and women for a police academy in the area in which the study was conducted. The study suggests that the three predictors and demographic variables account for 55% of variance of the predictors and the criterion—assuming the criterion scores are reliable.

5. Testing the Significance of Validity Coefficients

<table>
<thead>
<tr>
<th>Validity</th>
<th>Size of Sample</th>
<th>df</th>
<th>Critical $r$ From Appendix D</th>
<th>Significant at $p &lt; .05$?</th>
<th>Coefficient of Determination</th>
</tr>
</thead>
<tbody>
<tr>
<td>.23</td>
<td>37</td>
<td>35</td>
<td>.3246</td>
<td>No</td>
<td>.0529</td>
</tr>
<tr>
<td>.43</td>
<td>10</td>
<td>8</td>
<td>.6319</td>
<td>No</td>
<td>.1849</td>
</tr>
<tr>
<td>.33</td>
<td>52</td>
<td>50</td>
<td>.2732</td>
<td>Yes</td>
<td>.1089</td>
</tr>
<tr>
<td>.22</td>
<td>1,000</td>
<td>998</td>
<td>.1946</td>
<td>Yes</td>
<td>.0484</td>
</tr>
<tr>
<td>.50</td>
<td>6</td>
<td>4</td>
<td>.8114</td>
<td>No</td>
<td>.2500</td>
</tr>
</tbody>
</table>

The best evidence of validity is when the coefficient is statistically significant and also accounts for a substantial amount of shared variance. In this case, $r = .33$ for a sample of 52 provides the best evidence of criterion-related validity.
### 6. Calculating a Validity Coefficient

<table>
<thead>
<tr>
<th>Test Taker</th>
<th>X</th>
<th>Y</th>
<th>$D_x$</th>
<th>$D_x^2$</th>
<th>$D_y$</th>
<th>$D_y^2$</th>
<th>$D_xD_y$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abel</td>
<td>80</td>
<td>2</td>
<td>7.2</td>
<td>51.84</td>
<td>−0.5</td>
<td>0.25</td>
<td>−3.6</td>
</tr>
<tr>
<td>Bartmann</td>
<td>98</td>
<td>3</td>
<td>25.2</td>
<td>635.04</td>
<td>0.5</td>
<td>0.25</td>
<td>12.6</td>
</tr>
<tr>
<td>Cardoza</td>
<td>95</td>
<td>5</td>
<td>22.2</td>
<td>492.84</td>
<td>2.5</td>
<td>6.25</td>
<td>55.5</td>
</tr>
<tr>
<td>Dixon</td>
<td>55</td>
<td>3</td>
<td>−17.8</td>
<td>316.84</td>
<td>0.5</td>
<td>0.25</td>
<td>−8.9</td>
</tr>
<tr>
<td>Everett</td>
<td>70</td>
<td>2</td>
<td>−2.8</td>
<td>7.84</td>
<td>−0.5</td>
<td>0.25</td>
<td>1.4</td>
</tr>
<tr>
<td>Friedman</td>
<td>75</td>
<td>4</td>
<td>2.2</td>
<td>4.84</td>
<td>1.5</td>
<td>2.25</td>
<td>3.3</td>
</tr>
<tr>
<td>Grass</td>
<td>50</td>
<td>2</td>
<td>−22.8</td>
<td>519.84</td>
<td>−0.5</td>
<td>0.25</td>
<td>11.4</td>
</tr>
<tr>
<td>Hart</td>
<td>55</td>
<td>1</td>
<td>−17.8</td>
<td>316.84</td>
<td>−1.5</td>
<td>2.25</td>
<td>26.7</td>
</tr>
<tr>
<td>Isaacs</td>
<td>90</td>
<td>2</td>
<td>17.2</td>
<td>295.84</td>
<td>−0.5</td>
<td>0.25</td>
<td>−8.6</td>
</tr>
<tr>
<td>Jensen</td>
<td>60</td>
<td>1</td>
<td>−12.8</td>
<td>163.84</td>
<td>−1.5</td>
<td>2.25</td>
<td>19.2</td>
</tr>
<tr>
<td>Sum</td>
<td>728</td>
<td>25</td>
<td>2805.60</td>
<td>14.50</td>
<td>109.0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Mean:** 72.8  
**Standard deviation (calculated using the formula presented in Chapter 5):** 16.75

| $r$ | .54 |

| Abel   | 80 | 2  | −4.67 | 21.78 | −1.00 | 1.00 | 4.67  |
| Bartmann | 98 | 3  | 13.33 | 177.78| 0.00  | 0.00 | 0.00  |
| Cardoza | 95 | 5  | 10.33 | 106.78| 2.00  | 4.00 | 20.67 |
| Everett | 70 | 2  | −14.67| 215.11| −1.00 | 1.00 | 14.67 |
| Friedman | 75 | 4  | −9.67 | 93.44 | 1.00  | 1.00 | −9.67 |
| Isaacs  | 90 | 2  | 5.33  | 28.44 | −1.00 | 1.00 | −5.33 |
| Sum     | 508| 18 | 643.33| 8     | 25    |

**Mean:** 84.67  
**Standard deviation (calculated using the formula presented in Chapter 5):** 10.35

| $r$ | .35 |
The validity coefficient changed because of restriction of range.

Multiple Choice

<table>
<thead>
<tr>
<th>1. c</th>
<th>2. d</th>
<th>3. c</th>
<th>4. c</th>
<th>5. a</th>
</tr>
</thead>
<tbody>
<tr>
<td>6. c</td>
<td>7. a</td>
<td>8. d</td>
<td>9. c</td>
<td>10. d</td>
</tr>
<tr>
<td>11. a</td>
<td>12. a</td>
<td>13. c</td>
<td>14. c</td>
<td>15. c</td>
</tr>
</tbody>
</table>

Short Answer/Essay

Refer to your textbook for answers. If you are unsure of an answer and cannot generate the answer after reviewing your book, ask your professor for clarification.