MODULE 6

Reliability and Validity

LEARNING OBJECTIVES

- Explain what reliability is and how it is measured.
- Identify and explain the four types of reliability discussed in the text.
- Explain what validity is and how it is measured.
- Identify and explain the four types of validity discussed in the text.

reliability: An indica
of the consistency or
stability of a measuri
instrument.

JLE 6

y

ed in the text.

d in the text.

RELIABILITY

reliability: An indication of the consistency or stability of a measuring instrument.

One means of determining whether the measure you are using is effective is to assess its reliability. **Reliability** refers to the consistency or stability of a measuring instrument. In other words, the measuring instrument must measure exactly the same way every time it is used. This consistency means that individuals should receive a similar score each time they use the measuring instrument. For example, a bathroom scale needs to be reliable, that is, it needs to measure the same way every time an individual uses it, or otherwise it is useless as a measuring instrument.

Error in Measurement

Consider some of the problems with the four types of measures discussed in the previous module (i.e., self-report, tests, behavioral, and physical). Some problems, known as *method errors*, stem from the experimenter and the testing situation. Does the individual taking the measures know how to use the measuring instrument properly? Is the measuring equipment working correctly? Other problems, known as *trait errors*, stem from the participants. Were the participants being truthful? Did they feel well on the day of the test? Both types of problems can lead to measurement error.

In fact, a measurement is a combination of the true score and an error score. The true score is what the score on the measuring instrument would be if there were no error. The error score is any measurement error (method or trait) (Leary, 2001; Salkind, 1997). The following formula represents the *observed score* on a measure, that is, the score recorded for a participant on the measuring instrument used. The observed score is the sum of the true score and the measurement error.

$$\text{Observed score} = \text{True score} + \text{Measurement error}$$

The observed score becomes increasingly reliable (more consistent) as we minimize error and thus have a more accurate true score. True scores should not vary much over time, but error scores can vary tremendously from one testing session to another. How then can we minimize error in measurement? We can make sure that all the problems related to the four types of measures are minimized. These problems include those in recording or scoring data (method error) and those in understanding instructions, motivation, fatigue, and the testing environment (trait error). The conceptual formula for reliability is

$$\text{Reliability} = \frac{\text{True score}}{\text{True score} + \text{Error score}}$$

This conceptual formula indicates that a reduction in error leads to an increase in reliability. If there is no error, reliability is equal to 1.00, the highest possible reliability score. Also, as error increases, reliability drops below 1.00. The greater the error, the lower the reliability of a measure.

How to Measure Reliability: Correlation Coefficients

Reliability is measured using correlation coefficients. We briefly discuss them here; a more comprehensive discussion appears in Chapter Five.

correlation coefficient: A measure of the degree of relationship between two sets of scores. It can vary between −1.00 and +1.00.

A **correlation coefficient** measures the degree of relationship between two sets of scores and can vary between −1.00 and +1.00. The stronger the relationship between the variables, the closer the coefficient is to either −1.00 or +1.00. The weaker the relationship between the variables, the closer the coefficient is to 0. Suppose then that of individuals measured on two variables, the top-scoring individual on variable 1 was also top scoring on variable 2, the second-highest-scoring person on variable 1 was also the second highest on variable 2, and so on down to the lowest-scoring person. In this case there would be a perfect positive correlation (+1.00) between variables 1 and 2. In the case of a perfect negative correlation (−1.00), the person having the highest score on variable 1 would have the lowest score on variable 2, the person with the second-highest score on variable 1 would have the second-lowest score on variable 2, and so on. In reality variables are almost never perfectly correlated. Thus most correlation coefficients are less than 1.

A correlation of 0 between two variables indicates the absence of any relationship as might occur by chance. Suppose we drew a person's scores on variables 1 and 2 out of a hat containing random scores, and suppose we did the same for each person in the group. We would expect no relationship between individuals' scores on the two variables. It would be impossible to predict a person's performance on variable 2 based on the score on variable 1 because there would be no relationship (a correlation of 0) between the variables.

The sign preceding the correlation coefficient indicates whether the observed relationship is positive or negative. However, the terms "positive" and "negative" do not refer to good and bad relationships but rather to how the variables are related. A **positive correlation** indicates a direct relationship between variables: When we see high scores on one variable, we tend to see high scores on the other; when we see low or moderate scores on one variable, we see similar scores on the second. Variables that are positively correlated include height with weight and high school GPA with college GPA.

positive correlation: A direct relationship between two variables in which an increase in one is related to an increase in the other and a decrease in one is related to a decrease in the other.

negative correlation: An inverse relationship between two variables in which an increase in one variable is related to a decrease in the other and vice versa.

A **negative correlation** indicates an inverse, or negative, relationship: High scores on one variable go with low scores on the other and vice versa. Examples of negative relationships are sometimes more difficult for students to generate and to think about. In adults, however, many variables are negatively correlated with age: As age increases, variables such as sight, hearing ability, strength, and energy level tend to decrease.

Correlation coefficients can be weak, moderate, or strong. Table 6.1 gives guidelines for these categories. To establish the reliability (or consistency) of a measure, we expect a strong correlation coefficient—usually in the .80s or .90s—between the two variables or scores being measured (Anastasi & Urbina, 1997). We also expect the coefficient to be positive. A positive coefficient indicates consistency, that is, those who scored high at one time also scored high at another time, those who scored low at one point scored low again, and those with intermediate scores the first time scored similarly the s

riefly discuss them
Five.
ship between two
stronger the rela-
o either −1.00 or
he closer the coef-
on two variables,
ng on variable 2,
he second highest
In this case there
iables 1 and 2. In
having the high-
ble 2, the person
he second-lowest
st never perfectly

bsence of any re-
rson's scores on
and suppose we
t no relationship
be impossible to
re on variable 1
0) between the

whether the ob-
"positive" and
ther to how the
relationship be-
we tend to see
es on one vari-
ositively corre-
llege GPA.
, relationship:
and vice versa.
lt for students
ables are nega-
sight, hearing

ng. Table 6.1
or consistency)
lly in the .80s
l (Anastasi &
ositive coeffi-
one time also
nt scored low
l similarly the

TABLE **6.1**

Values for Weak, Moderate, and Strong Correlation Coefficients

Correlation Coefficient	Strength of Relationship
±.70–1.00	Strong
±.30–.69	Moderate
±.00–.29	None (.00) to Weak

second time. A negative coefficient indicates an inverse relationship between the scores taken at two different times, and it is hardly consistent (i.e., reliable) for a person to score very high at one time and very low at another. Thus to establish that a measure is reliable, we need a positive correlation coefficient of around .80 or higher.

Types of Reliability

There are four types of reliability: test/retest reliability, alternate-forms reliability, split-half reliability, and interrater reliability. Each type provides a measure of consistency, but they are used in different situations.

Test/Retest Reliability

One of the most often used and obvious ways of establishing reliability is to repeat the same test on a second occasion, a process known as **test/retest reliability**. The correlation coefficient obtained is between the two scores of an individual on the same test administered on two occasions. If the test is reliable, we expect the two scores for each individual to be similar, and thus the resulting correlation coefficient will be high (close to +1.00). This measure of reliability assesses the stability of a test over time. Naturally some error will be present in each measurement (for example, an individual may not feel well at one testing or may have problems during the testing session such as with a broken pencil). Therefore it is unusual for the correlation coefficient to be +1.00, but we expect it to be +.80 or higher. A problem related to test/retest measures is that on many tests there are *practice effects*, that is, some people get better at the second testing, and this "practice" lowers the observed correlation. A second problem may occur if the interval between test times is short: Individuals may remember how they answered previously, both correctly and incorrectly. In this case we may be testing their memories and not the reliability of the testing instrument, and we may observe a spuriously high correlation.

test/retest reliability: A reliability coefficient determined by assessing the degree of relationship between scores on the same test administered on two different occasions.

Alternate-Forms Reliability

One means of controlling for test/retest problems is to use **alternate-forms reliability**, that is, using alternate forms of the testing instrument and correlating the performance of individuals on the two different forms. In this case the tests taken at times 1 and 2 are different but equivalent or parallel (hence the terms *equivalent-forms reliability* and *parallel-forms reliability* are also used).

alternate-forms reliability: A reliability coefficient determined by assessing the degree of relationship between scores on two equivalent tests.

As with test/retest reliability alternate-forms reliability establishes the stability of the test over time. In addition, it also establishes the equivalency of the items from one test to another. One problem with alternate-forms reliability is making sure that the tests are truly parallel. To help ensure equivalency, the tests should have the same number of items, the items should be of the same difficulty level, and instructions, time limits, examples, and format should all be equal—often difficult, if not impossible, to accomplish. Further, if the tests are truly equivalent, there is the potential for practice, although not to the same extent as when exactly the same test is administered twice.

Split-Half Reliability

split-half reliability: A reliability coefficient determined by correlating scores on one half of a measure with scores on the other half of the measure.

A third means of establishing reliability is by splitting the items on the test into equivalent halves and correlating scores on one half of the items with scores on the other half. This **split-half reliability** gives a measure of the equivalence of the content of the test but not of its stability over time as test/retest and alternate-forms reliability do. The biggest problem with split-half reliability is determining how to divide the items so that the two halves are in fact equivalent. For example, it would not be advisable to correlate scores on multiple-choice questions with scores on short-answer or essay questions. What is typically recommended is to correlate scores on even-numbered items with scores on odd-numbered items. Thus if the items at the beginning of the test are easier or harder than those at the end of the test, the half scores are still equivalent.

Interrater Reliability

interrater reliability: A reliability coefficient that assesses the agreement of observations made by two or more raters or judges.

Finally, to measure the reliability of observers rather than tests, you can use **interrater reliability**, which is a measure of consistency that assesses the agreement of observations made by two or more raters or judges. Let's say that you are observing play behavior in children. Rather than simply making observations on your own, it is advisable to have several independent observers collect data. The observers all watch the children playing but independently count the number and types of play behaviors they observe. Once the data are collected, interrater reliability needs to be established by examining the percentage of agreement among the raters. If the raters' data are reliable, then the percentage of agreement should be high. If the raters are not paying close attention to what they are doing or if the measuring scale devised for the various play behaviors is unclear, the percentage of agreement among observers will not be high. Although interrater reliability is measured using a correlation coefficient, the following formula offers a quick means of estimating interrater reliability:

$$\text{Interrater reliability} = \frac{\text{Number of agreements}}{\text{Number of possible agreements}} \times 100$$

Thus, if your observers agree 45 times out of a possible 50, the interrater reliability is 90%—fairly high. However, if they agree only 20 times out of 50, then the interrater reliability is 40%—low. Such a low level of agreement indicates a problem with the measuring instrument or with the individuals using the instrument and should be of great concern to a researcher.

IN REVIEW

What it measures

How it is accomplish

CRITI
THINK
CHEC

VALIDITY

validity: An in whether the ins measures what measure.

content validit extent to which measuring instru covers a represe sample of the d behaviors to be

lishes the stability
quivalency of the
e-forms reliability
isure equivalency,
should be of the
ples, and format
complish. Further,
practice, although
inistered twice.

ns on the test into
ms with scores on
he equivalence of
as test/retest and
t-half reliability is
ire in fact equiva-
ores on multiple-
ons. What is typi-
items with scores
the test are easier
still equivalent.

ests, you can use
ssesses the agree-
es. Let's say that
mply making ob-
endent observers
ut independently
e. Once the data
y examining the
data are reliable,
rs are not paying
scale devised for
ement among ob-
neasured using a
neans of estimat-

– × 100
s

he interrater reli-
times out of 50,
of agreement in-
the individuals
earcher.

IN REVIEW **Features of Types of Reliability**

	Types of Reliability			
	Test/Retest	**Alternate-Forms**	**Split-Half**	**Interrater**
What it measures	Stability over time	Stability over time and equivalency of items	Equivalency of items	Agreement between raters
How it is accomplished	Administer the same test to the same people at two different times	Administer alternate but equivalent forms of the test to the same people at two different times	Correlate performance for a group of people on two equivalent halves of the same test	Have at least two people count or rate behaviors and determine the percentage of agreement among them

CRITICAL THINKING CHECK 6.1

1. Why does alternate-forms reliability provide a measure of both equivalency of items and stability over time?
2. Two people observe whether or not vehicles stop at a stop sign. They make 250 observations and disagree 38 times. What is the interrater reliability? Is this good, or should it be of concern to the researchers?

VALIDITY

validity: An indication of whether the instrument measures what it claims to measure.

In addition to being reliable, measures must also be valid. **Validity** refers to whether a measuring instrument measures what it claims to measure. There are several types of validity; we will discuss four. As with reliability, validity is measured by the use of correlation coefficients. For instance, if researchers developed a new test to measure depression, they might establish the validity of the test by correlating scores on the new test with scores on an already established measure of depression, and as with reliability we would expect the correlation to be positive. Unlike reliability coefficients, however, there is no established criterion for the strength of the validity coefficient. Coefficients as low as .20 or .30 may establish the validity of a measure (Anastasi & Urbina, 1997). What is important for validity coefficients is that they are *statistically significant* at the .05 or .01 level. We explain this term in a later module, but in brief it means that the results are most likely not due to chance.

Content Validity

content validity: The extent to which a measuring instrument covers a representative sample of the domain of behaviors to be measured.

A systematic examination of the test content to determine whether it covers a representative sample of the domain of behaviors to be measured assesses **content validity**. In other words, a test with content validity has items that satisfactorily assess the content being examined. To determine whether a test has content validity, researchers consult experts in the area being tested.

As an example, when designers of the GRE generate a subject exam for psychology, they ask professors of psychology to examine the questions to establish that they represent relevant information from the entire discipline of psychology as we know it today.

face validity: The extent to which a measuring instrument appears valid on its surface.

Sometimes face validity is confused with content validity. **Face validity** simply addresses whether or not a test looks valid on its surface. Does it appear to be an adequate measure of the conceptual variable? Face validity is not really validity in the technical sense because it refers not to what the test actually measures but to what it appears to measure. Face validity relates to whether the test looks valid to those who selected it and to those who take it. For instance, does the test selected by the school board to measure student achievement "appear" to be an actual measure of achievement? Face validity has more to do with rapport and public relations than with actual validity (Anastasi & Urbina, 1997).

Criterion Validity

criterion validity: The extent to which a measuring instrument accurately predicts behavior or ability in a given area.

The extent to which a measuring instrument accurately predicts behavior or ability in a given area establishes **criterion validity**. Two types of criterion validity may be used, depending on whether the test is used to estimate present performance (*concurrent validity*) or to predict future performance (*predictive validity*). The SAT and GRE are examples of tests that have predictive validity because performance on the tests correlates with later performance in college and graduate school, respectively. The tests can be used with some degree of accuracy to "predict" future behavior. A test used to determine whether someone qualifies as a pilot is a measure of concurrent validity. The test is estimating the person's ability at the present time, not attempting to predict future outcomes. Thus concurrent validation is used for the diagnosis of existing status rather than the prediction of future outcomes.

Construct Validity

construct validity: The degree to which a measuring instrument accurately measures a theoretical construct or trait that it is designed to measure.

Construct validity is considered by many to be the most important type of validity. The **construct validity** of a test assesses the extent to which a measuring instrument accurately measures a theoretical construct or trait that it is designed to measure. Some examples of theoretical constructs or traits are verbal fluency, neuroticism, depression, anxiety, intelligence, and scholastic aptitude. One means of establishing construct validity is by correlating performance on the test with performance on a test for which construct validity has already been determined. Thus performance on a newly developed intelligence test might be correlated with performance on an existing intelligence test for which construct validity has been previously established. Another means of establishing construct validity is to show that the scores on the new test differ across people with different levels of the trait being measured. For example, if a new test is designed to measure depression, you can compare scores on the test for those known to be suffering from depression with scores for those not suffering from depression. The new measure has construct validity if it measures the construct of depression accurately.

The Relationship between Reliability and Validity

Obviously a measure should be both reliable and valid. It is possible, however, to have a test or measure that meets one of these criteria and not the other. Think for a moment about how this situation might occur. Can a test be reliable without being valid? Can a test be valid without being reliable? To answer these questions, suppose we are going to measure intelligence in a group of individuals with a "new" intelligence test. The test is based on a rather ridiculous theory of intelligence, which states that the larger your brain is, the more intelligent you are. The assumption is that the larger your brain is, the larger your head is. Thus the test is going to measure intelligence by measuring head circumference; so we gather a sample of individuals and measure the circumference of each person's head.

Is this a reliable measure? Many people immediately say no because head circumference seems like such a laughable way to measure intelligence. But reliability is a measure of consistency, not truthfulness. Is this test going to consistently measure the same thing? Yes, it is consistently measuring head circumference, and this measurement is not likely to change over time. Thus each person's score at one time will be the same or very close to the same as the person's score at a later time. The test is therefore very reliable.

Is the test a valid measure of intelligence? No, it in no way measures the construct of intelligence. Thus we have established that a test can be reliable without being valid, and because the test lacks validity, it is not a good measure.

Can the reverse be true? That is, can a test be valid (it truly measures what it claims to measure) and not be reliable? If a test truly measures intelligence, individuals would score about the same each time they took it because intelligence does not vary much over time. Thus if the test is valid, it must be reliable. Therefore a test can be reliable and not valid, but if it is valid, it is necessarily reliable.

IN REVIEW Features of Types of Validity

	Types of Validity			
	Content	Criterion/ Concurrent	Criterion/ Predictive	Construct
What it measures	Whether the test covers a representative sample of the domain of behaviors to be measured	The ability of the test to estimate present performance	The ability of the test to predict future performance	The extent to which the test measures a theoretical construct or trait
How it is accomplished	Ask experts to assess the test to establish that the items are representative of the trait being measured	Correlate performance on the test with a concurrent behavior	Correlate performance on the test with a behavior in the future	Correlate performance on the test with performance on an established test or with people who have different levels of the trait the test claims to measure