Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2013). How to design and evaluate research in education (Vol. 7). New York: McGraw-Hill.

8 Validity and Reliability

OBJECTIVES **Studying this chapter should enable you to:**

- Explain what is meant by the term "validity" as it applies to the use of instruments in educational research.
- Name three types of evidence of validity that can be obtained, and give an example of each type.
- Explain what is meant by the term "correlation coefficient" and describe briefly the difference between positive and negative correlation coefficients.
- Explain what is meant by the terms "validity coefficient" and "reliability coefficient."
- Explain what is meant by the term "reliability" as it applies to the use of instruments in educational research.
- Explain what is meant by the term "errors of measurement."
- Explain briefly the meaning and use of the term "standard error of measurement."
- Describe briefly three ways to estimate the reliability of the scores obtained using a particular instrument.
- Describe how to obtain and evaluate scoring agreement.

INTERACTIVE AND APPLIED LEARNING After, or while, reading this chapter:

Go to the Online Learning Center at www.mhhe.com/fraenkel8e to:

- Learn More About Validity and Reliability

Go to your online Student Mastery Activities book to do the following activities:

- Activity 8.1: Instrument Validity
- Activity 8.2: Instrument Reliability (1)
- Activity 8.3: Instrument Reliability (2)
- Activity 8.4: What Kind of Evidence: Content-Related, Criterion-Related, or Construct-Related?
- Activity 8.5: What Constitutes Construct-Related Evidence of Validity?

It isn't fair, Tony!"

"What isn't, Lily?"

"Those tests that Mrs. Leonard gives. Grrr!"

"What about them?"

"Well, take this last test we had on the Civil War. All during her lectures and the class discussions over the last few weeks, we've been talking about the causes and effects of the War."

"So?"

"Well, then on this test, she asked a lot about battles and generals and other stuff that we didn't study."

"Did you ask her how come?"

"Yeah, I did. She said she wanted to test our thinking ability. But she was asking us to think about material she hadn't even gone over or discussed in class. That's why I think she isn't fair."

Lily is correct. Her teacher, in this instance, isn't being fair. Although she isn't using the term, what Lily is talking about is a matter of *validity*. It appears Mrs. Leonard is giving an *invalid* test. What this means, and why it isn't a good thing for a teacher (or any researcher) to do, is largely what this chapter is about.

The Importance of Valid Instrumentation

The quality of the instruments used in research is very important, for the conclusions researchers draw are based on the information they obtain using these instruments. Accordingly, researchers use a number of procedures to ensure that the inferences they draw, based on the data they collect, are valid and reliable.

Validity refers to the appropriateness, meaningfulness, correctness, and usefulness of the inferences a researcher makes. *Reliability* refers to the consistency of scores or answers from one administration of an instrument to another, and from one set of items to another. Both concepts are important to consider when it comes to the selection or design of the instruments a researcher intends to use. In this chapter, therefore, we shall discuss both validity and reliability in some detail.

Validity

Validity is the most important idea to consider when preparing or selecting an instrument for use. More than anything else, researchers want the information they obtain through the use of an instrument to serve their purposes. For example, to find out what teachers in a particular school district think about a recent policy passed by the school board, researchers need both an instrument to record the data and some sort of assurance that the information obtained will enable them to

draw correct conclusions about teacher opinions. The drawing of correct conclusions based on the data obtained from an assessment is what validity is all about. Though not essential, some kind of score that summarizes the information for each person greatly simplifies the comprehension and use of data, and because most instruments provide such scores, we present the following discussion in this context.

In recent years, **validity** has been defined as referring to the *appropriateness, correctness, meaningfulness,* and *usefulness* of the specific *inferences* researchers make based on the data they collect. *Validation* is the process of collecting and analyzing evidence to support such inferences. There are many ways to collect evidence, and we will discuss some of them shortly. The important point here is to realize that validity refers to the degree to which evidence supports any inferences a researcher makes based on the data he or she collects using a particular instrument. It is the inferences about the specific uses of an instrument that are validated, not the instrument itself.* These inferences should be appropriate, meaningful, correct, and useful.

One interpretation of this conceptualization of validity has been that test publishers no longer have a responsibility to provide evidence of validity. We do not agree; publishers have an obligation to state what an instrument is intended to measure and to provide evidence that it does. Nonetheless, researchers must still give attention to the way in which *they* intend to interpret the information.

An appropriate inference would be one that is relevant—that is, related—to the purposes of the study. If the purpose of a study were to determine what students know about African culture, for example, it would make no sense to make inferences about this from their scores on a test about the physical geography of Africa.

A meaningful inference is one that says something about the *meaning* of the information (such as test scores) obtained through the use of an instrument. What exactly does a high score on a particular test mean? What does such a score allow us to say about the individual who received it? In what way is an individual who receives a high score different from one who receives a low score? And so forth. It is one thing to collect information from people. We do this all the time—names, addresses, birth dates, shoe sizes, car license numbers, and so on. But unless we can make inferences that mean something from the information we obtain, it is of little use. The purpose of research is not merely to collect data but to use such data to draw warranted conclusions about the people (and others like them) on whom the data were collected.

A useful inference is one that helps researchers make a decision related to what they were trying to find out. Researchers interested in the effects of inquiry-related teaching materials on student achievement, for example, need information that will enable them to infer whether achievement is affected by such materials and, if so, how.

Validity, therefore, depends on the amount and type of evidence there is to support the interpretations researchers wish to make concerning data they have collected. The crucial question is: Do the results of the assessment provide useful information about the topic or variable being measured?

What kinds of evidence might a researcher collect? Essentially, there are three main types.

Content-related evidence of validity refers to the content and format of the instrument. How appropriate is the content? How comprehensive? Does it logically get at the intended variable? How adequately does the sample of items or questions represent the content to be assessed? Is the format appropriate? The content and format must be consistent with the definition of the variable and the sample of subjects to be measured.

Criterion-related evidence of validity refers to the relationship between scores obtained using the instrument and scores obtained using one or more other instruments or measures (often called a *criterion*). How strong is this relationship? How well do such scores estimate, present, or predict future performance of a certain type?

Construct-related evidence of validity refers to the nature of the psychological construct or characteristic being measured by the instrument. How well does a measure of the construct explain differences in the behavior of individuals or their performance on certain tasks? We provide further explanation of this rather complex concept later in the chapter.

Figure 8.1 illustrates these three types of evidence.

*This is somewhat of a change from past interpretations. It is based on the set of standards prepared by a joint committee consisting of members of the American Educational Research Association, the American Psychological Association, and the National Council on Measurement in Education. See American Psychological Association (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association, pp. 9–18, 19–23.

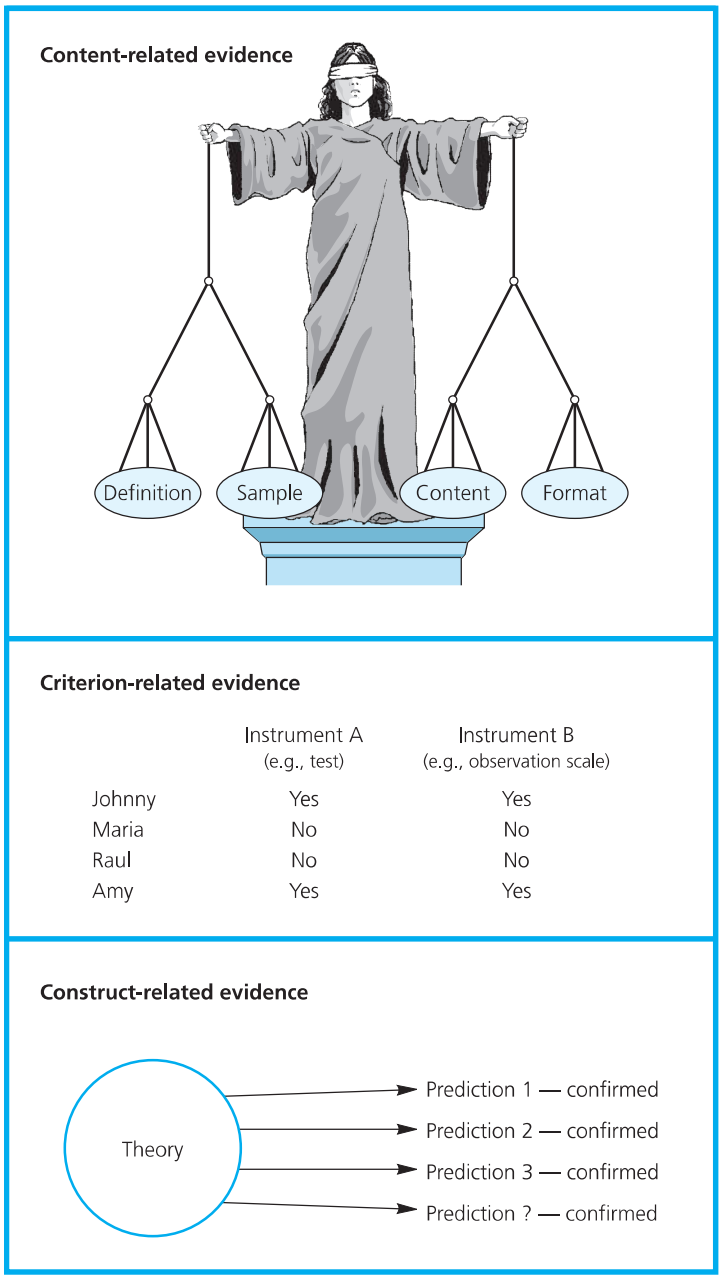


Figure 8.1 *Types of Evidence of Validity*

CONTENT-RELATED EVIDENCE

Suppose a researcher is interested in the effects of a new math program on the mathematics ability of fifth-graders. The researcher expects that students who complete the program will be able to solve a number of different types of word problems correctly. To assess their mathematics ability, the researcher plans to give them a math test containing about 15 such problems. The performance of the students on this test is important only to the degree that it provides evidence of their ability to solve these kinds of problems. Hence, performance on the instrument in this case (the math test) will provide valid evidence of the mathematics ability of these students *if* the instrument provides an adequate sample of the types of word problems covered in the program. If only easy problems are included on the test,

CONTROVERSIES IN RESEARCH

High-Stakes Testing

High-stakes testing refers to the use of tests (often only a single achievement test) as the primary, or only basis for decisions having major consequences. For students, such consequences include retention in grade and/or the denial of diplomas and awards. For schools, they include public praise or condemnation, sanctions, and financial rewards or punishments. "In state after state, legislatures, governors, and state boards, supported by business leaders, have imposed tougher requirements in mathematics, English, science, and other fields, together with new tests by which the performance of both students and schools is to be judged."*

For years, tests had been used as one indicator of performance; what was new was exclusive reliance on them. "The backlash, touching virtually every state that has instituted high-stakes testing, arises from a spectrum of complaints. A major complaint is that the focus on testing and obsessive test preparation, sometimes beginning in kindergarten, is killing innovative teaching and curricula and driving out good teachers. Other complaints are that (conversely) the standards on which the tests are based are too vague and that students have not been taught the material on which the tests are based; or that the tests are unfair to poor and minority students or to those who lack test-taking skills; or that the tests put too much stress on young children. And some argue that they are too long (in Massachusetts they can take up to 13 hours!) or too tough or simply not good enough."†

In response, the American Educational Research Association developed a position statement of "conditions essential to sound implementation of high-stakes educational testing programs."‡ It contained 14 specific points, 4 of the most important being that (1) such decisions about students should not be based on test scores alone; (2) tests should be made fairer to all students; (3) tests should match the curriculum; and (4) the reliability and validity of tests should continually be evaluated.

Two examples of responses to the guidelines were the following:

- "In the face of too much testing with far too severe consequences, the AERA positions, if implemented, would be a step forward relative to current practice."§
- "The statement reflects what is desired for all state tests and assessments. But, just as all students have not yet met the standards, not all state tests and assessments will immediately meet the goals contained in this statement."‖

What do you think? Are the complaints about high-stakes tests warranted?

*P. Schrag (2000). High stakes are for tomatoes. *Atlantic Monthly, 286* (August): 19.

†Ibid.

‡American Educational Research Association (2000). Position statement of the American Educational Research Association concerning high-stakes testing in pre-12 education. *Educational Researcher, 29* (11): 24–35.

§M. Neill (2000). Quoted in Initial responses to AERA's position statement concerning high-stakes testing. *Educational Researcher, 29* (11): 28.

‖W. Martin (2000). Quoted in Ibid., p. 27.

or only very difficult or lengthy ones, or only problems involving subtraction, the test will be unrepresentative and hence not provide information from which valid inferences can be made.

One key element in **content-related evidence of validity**, then, concerns the adequacy of the sampling. Most instruments (and especially achievement tests) provide only a sample of the kinds of problems that might be solved or questions that might be asked. Content validation, therefore, is partly a matter of determining if the content that the instrument contains is an adequate sample of the domain of content it is supposed to represent.

The other aspect of content validation has to do with the format of the instrument. This includes such things as the clarity of printing, size of type, adequacy of work space (if needed), appropriateness of language, clarity of directions, and so on. Regardless of the adequacy of the questions in an instrument, if they are presented in an inappropriate format (such as giving a test written in English to children whose English is minimal), valid results cannot be obtained. For this reason, it is important that the characteristics of the intended sample be kept in mind.

How does one obtain content-related evidence of validity? A common way to do this is to have someone look at the content and format of the instrument and judge whether or not it is appropriate. The "someone," of course, should not be just anyone, but rather an individual who can be expected to render an intelligent judgment about the adequacy of the instrument—in other words, someone who knows enough about what is to be measured to be a competent judge.

The usual procedure is somewhat as follows. The researcher writes out the definition of what he or she wants to measure and then gives this definition, along with the instrument and a description of the intended sample, to one or more judges. The judges look at the definition, read over the items or questions in the instrument, and place a check mark in front of each question or item that they feel does not measure one or more aspects of the definition (objectives, for example) or other criteria. They also place a check mark in front of each aspect not assessed by any of the items. In addition, the judges evaluate the appropriateness of the instrument format. The researcher then rewrites any item or question so checked and resubmits it to the judges, and/or writes new items for criteria not adequately covered. This continues until the judges approve all the items or questions in the instrument and also indicate that they feel the total number of items is an adequate representation of the total domain of content covered by the variable being measured.

To illustrate how a researcher might go about trying to establish content-related validity, let us consider two examples.

Example 1. Suppose a researcher desires to measure students' ability to *use information that they have previously acquired.* When asked what she means by this phrase, she offers the following definition.

> As evidence that students can use previously acquired information, they should be able to:
>
> 1. Draw a correct conclusion (verbally or in writing) that is based on information they are given.
> 2. Identify one or more logical implications that follow from a given point of view.
> 3. State (orally or in writing) whether two ideas are identical, similar, unrelated, or contradictory.

How might the researcher obtain such evidence? She decides to prepare a written test that will contain various questions. Students' answers will constitute the evidence she seeks. Here are three examples of the kinds of questions she has in mind, designed to produce each of the three types of evidence listed above.

1. If A is greater than B, and B is greater than C, then:
   a. A must be greater than C.
   b. C must be smaller than A.
   c. B must be smaller than A.
   d. All of the above are true.
2. Those who believe that increasing consumer expenditures would be the best way to stimulate the economy would advocate
   a. an increase in interest rates.
   b. an increase in depletion allowances.
   c. tax reductions in the lower income brackets.
   d. a reduction in government expenditures.
3. Compare the dollar amounts spent by the U.S. government during the past 10 years for (*a*) debt payments, (*b*) defense, and (*c*) social services.

Now, look at each of the questions and the corresponding objective they are supposed to measure. Do you think each question measures the objective it was designed for? If not, why not?*

Example 2. Here is what another researcher designed as an attempt to measure (at least in part) the ability of students to *explain why events occur.*

> Read the directions that follow, and then answer the question.
>
> *Directions:* Here are some facts.
>
> *Fact W:* A camper started a fire to cook food on a windy day in a forest.
>
> *Fact X:* A fire started in some dry grass near a campfire in a forest.
>
> Here is another fact that happened later the same day in the same forest.
>
> *Fact Y:* A house in the forest burned down.
>
> You are to explain what might have caused the house to burn down (Fact Y). Would Fact W and X be useful as parts of your explanation?
>
> a. Yes, both W and X and the possible cause-and-effect relationship between them would be useful.
> b. Yes, both W and X would be useful, even though neither was likely a cause of the other.
> c. No, because only one of Facts W and X was likely a cause of Y.
> d. No, because neither W or X was likely a cause of Y.[1]

*We would rate correct answers to questions 1 (choice *d*) and 2 (choice *c*) as valid evidence, although 1 could be considered questionable, since students might view it as somewhat tricky. We would not rate the answers to 3 as valid, since students are not asked to contrast ideas, only facts.

Once again, look at the question and the objective it was designed to measure. Does it measure this objective? If not, why not?*

Attempts like these to obtain evidence of some sort (in the above instances, the support of independent judges that the items measure what they are supposed to measure) typify the process of obtaining content-related evidence of validity. As we mentioned previously, however, the qualifications of the judges are always an important consideration, and the judges must keep in mind the characteristics of the intended sample.

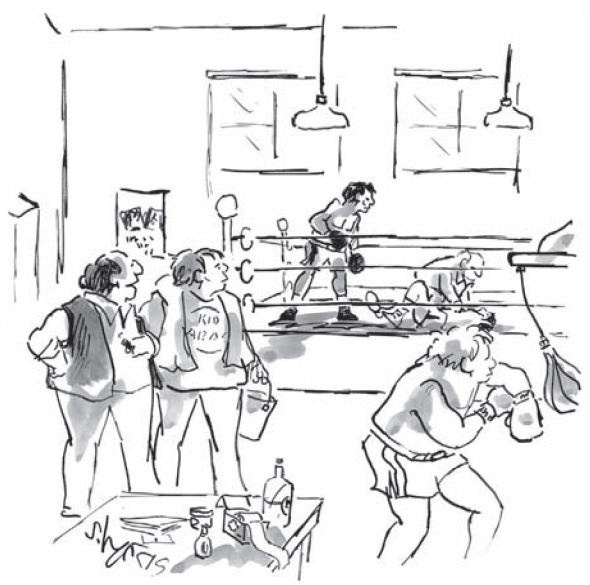
CRITERION-RELATED EVIDENCE

To obtain **criterion-related evidence of validity**, researchers usually compare performance on one instrument (the one being validated) with performance on some other, independent criterion. A **criterion** is a second test or other assessment procedure presumed to measure the same variable. For example, if an instrument has been designed to measure academic ability, student scores on the instrument might be compared with their grade-point averages (the external criterion). If the instrument does indeed measure academic ability, then students who score high on the test would also be expected to have high grade-point averages. Can you see why?

There are two forms of criterion-related validity—predictive and concurrent. To obtain evidence of **predictive validity**, researchers allow a time interval to elapse between administration of the instrument and obtaining the criterion scores. For example, a researcher might administer a science aptitude test to a group of high school students and later compare their scores on the test with their end-of-semester grades in science courses.

On the other hand, when instrument data and criterion data are gathered at nearly the same time, and the results are compared, this is an attempt by researchers to obtain evidence of **concurrent validity**. An example is when a researcher administers a self-esteem inventory to a group of eighth-graders and compares their scores on it with their teachers' ratings of student self-esteem obtained at about the same time.

A key index in both forms of criterion-related validity is the correlation coefficient.† A **correlation coefficient**, symbolized by the letter *r*, indicates the degree of relationship that exists between the scores individuals obtain on two instruments. A positive relationship is indicated when a high score on one of the instruments is accompanied by a high score on the other or when a low score on one is accompanied by a low score on the other. A negative relationship is indicated when a high score on one instrument is accompanied by a low score on the other, and vice versa. All correlation coefficients fall somewhere between +1.00 and −1.00. An *r* of .00 indicates that no relationship exists.



"He looks very promising—but let's see how he does on the written test."

When a correlation coefficient is used to describe the relationship between a set of scores obtained by the same group of individuals on a particular instrument and their scores on some criterion measure, it is called a **validity coefficient**. For example, a validity coefficient of +1.00 obtained by correlating a set of scores on a mathematics aptitude test (the predictor) and another set of scores, this time on a mathematics achievement test (the criterion), for the same individuals would indicate that each individual in the group had exactly the same relative standing on both measures. Such a correlation, if obtained, would allow the researcher to predict perfectly math achievement based on aptitude test scores. Although this correlation coefficient would be very unlikely, it illustrates what such coefficients mean. The higher the validity coefficient obtained, the more accurate a researcher's predictions are likely to be.

*We would rate a correct answer to this question as valid evidence of student ability to explain why events occur.

†The correlation coefficient, explained in detail in Chapter 10, is an extremely useful statistic. This is one of its many applications or uses.

Gronlund suggests the use of an expectancy table as another way to depict criterion-related evidence.[2] An **expectancy table** is nothing more than a two-way chart, with the predictor categories listed down the left-hand side of the chart and the criterion categories listed horizontally along the top of the chart. For each category of scores on the predictor, the researcher then indicates the percentage of individuals who fall within each of the categories on the criterion.

Table 8.1 presents an example. As you can see from the table, 51 percent of the students who were classified outstanding by these judges received a grade of A in orchestra, 35 percent received a B, and 14 percent received a C. Although this table refers only to this particular group, it could be used to predict the scores of other aspiring music students who were evaluated by these same judges. If a student obtained an evaluation of "outstanding," we might predict (approximately) that he or she would have a 51 percent chance of receiving an A, a 35 percent chance of receiving a B, and a 14 percent chance of receiving a C.

Expectancy tables are particularly useful devices for researchers to use with data collected in schools. They are simple to construct, easily understood, and clearly show the relationship between two measures.

It is important to realize that the nature of the criterion is the most important factor in gathering criterion-related evidence. High positive correlations do not mean much if the criterion measure does not make logical sense. For example, a high correlation between scores on an instrument designed to measure aptitude for science and scores on a physical fitness test would not be relevant criterion-related evidence for either instrument. Think back to the example we presented earlier of the questions designed to measure student ability to explain why events occur. What sort of criteria could be used to establish criterion-referenced validity for those items?

TABLE 8.1 *Example of an Expectancy Table*

Judges' Classification of Music Aptitude	Course Grades in Orchestra (Percentage Receiving Each Grade)			
	A	B	C	D
Outstanding	51	35	14	0
Above average	20	43	37	0
Average	0	6	83	11
Below average	0	0	13	87

CONSTRUCT-RELATED EVIDENCE

Construct-related evidence of validity is the broadest of the three categories of evidence for validity that we are considering. There is no single piece of evidence that satisfies construct-related validity. Rather, researchers attempt to collect a variety of *different* types of evidence (the more and the more varied the better) that will allow them to make warranted inferences—to assert, for example, that the scores obtained from administering a self-esteem inventory permit accurate inferences about the degree of self-esteem that people who receive those scores possess.

Usually, there are three steps involved in obtaining construct-related evidence of validity: (1) the variable being measured is clearly defined; (2) hypotheses, based on a theory underlying the variable, are formed about how people who possess a lot versus a little of the variable will behave in a particular situation; and (3) the hypotheses are tested both logically and empirically.

To make the process clearer, let us consider an example. Suppose a researcher interested in developing a pencil-and-paper test to measure honesty wants to use a construct-validity approach. First, he defines *honesty*. Next he formulates a theory about how "honest" people behave as compared to "dishonest" people. For example, he might theorize that honest individuals, if they find an object that does not belong to them, will make a reasonable effort to locate the individual to whom the object belongs. Based on this theory, the researcher might hypothesize that individuals who score high on his honesty test will be more likely to attempt to locate the owner of an object they find than individuals who score low on the test. The researcher then administers the honesty test, separates the names of those who score high and those who score low, and gives all of them an opportunity to be honest. He might, for example, leave a wallet with $5 in it lying just outside the test-taking room so that the individuals taking the test can easily see it and pick it up. The wallet displays the name and phone number of the owner in plain view. If the researcher's hypothesis is substantiated, more of the high scorers than the low scorers on the honesty test will attempt to call the owner of the wallet. (This could be checked by having the number answered by a recording machine asking the caller to leave his or her name and number.) This is one piece of evidence that could be used to support inferences about the honesty of individuals, based on the scores they receive on this test.

We must stress, however, that a researcher must carry out a series of studies to obtain a *variety* of evidence

suggesting that the scores from a particular instrument can be used to draw correct inferences about the variable that the instrument purports to measure. It is a broad array of evidence, rather than any one particular type of evidence, that is desired.

Consider a second example. Some evidence that might be considered to support a claim for construct validity in connection with a test designed to measure mathematical reasoning ability might be as follows:

- Independent judges all indicate that all items on the test require mathematical reasoning.
- Independent judges all indicate that the features of the test itself (such as test format, directions, scoring, and reading level) would not in any way prevent students from engaging in mathematical reasoning.
- Independent judges all indicate that the sample of tasks included in the test is relevant and representative of mathematical reasoning tasks.
- A high correlation exists between scores on the test and grades in mathematics.
- High scores have been made on the test by students who have had specific training in mathematical reasoning.
- Students actually engage in mathematical reasoning when they are asked to "think aloud" as they go about trying to solve the problems on the test.
- A high correlation exists between scores on the test and teacher ratings of competence in mathematical reasoning.
- Higher scores are obtained on the test by mathematics majors than by general science majors.

Other types of evidence might be listed for the above task (perhaps you can think of some), but we hope this is enough to make clear that it is not just one type, but many types, of evidence that a researcher seeks to obtain. Determining whether the scores obtained through the use of a particular instrument measure a particular variable involves a study of how the test was developed, the theory underlying the test, how the test functions with a variety of people and in a variety of situations, and how scores on the test relate to scores on other appropriate instruments. Construct validation involves, then, a wide variety of procedures and many different types of evidence, including both content-related and criterion-related evidence. The more evidence researchers have from many different sources, the more confident they become about interpreting the scores obtained from a particular instrument.

Reliability

Reliability refers to the consistency of the scores obtained—how consistent they are for each individual from one administration of an instrument to another and from one set of items to another. Consider, for example, a test designed to measure typing ability. If the test is reliable, we would expect a student who receives a high score the first time he takes the test to receive a high score the next time he takes the test. The scores would probably not be identical, but they should be close.

The scores obtained from an instrument can be quite reliable but not valid. Suppose a researcher gave a group of eighth-graders two forms of a test designed to measure their knowledge of the Constitution of the United States and found their scores to be consistent: those who scored high on form A also scored high on form B; those who scored low on A scored low on B; and so on. We would say that the scores were reliable. But if the researcher then used these same test scores to predict the success of these students in their physical education classes, she would probably be looked at in amazement. Any inferences about success in physical education based on scores on a Constitution test would have no validity. Now, what about the reverse? Can an instrument that yields unreliable scores permit valid inferences? No! If scores are completely inconsistent for a person, they provide no useful information. We have no way of knowing which score to use to infer an individual's ability, attitude, or other characteristic.

The distinction between reliability and validity is shown in Figure 8.2. Reliability and validity always depend on the context in which an instrument is used. Depending on the context, an instrument may or may not yield reliable (consistent) scores. If the data are unreliable, they cannot lead to valid (legitimate) inferences—as shown in target (*a*). As reliability improves, validity may improve, as shown in target (*b*), or it may not, as shown in target (*c*). An instrument may have good reliability but low validity, as shown in target (*d*). What is desired, of course, is both high reliability and high validity, as target (*e*) shows.

ERRORS OF MEASUREMENT

Whenever people take the same test twice, they will seldom perform exactly the same—that is, their scores or answers will not usually be identical. This may be due to a variety of factors (differences in motivation,

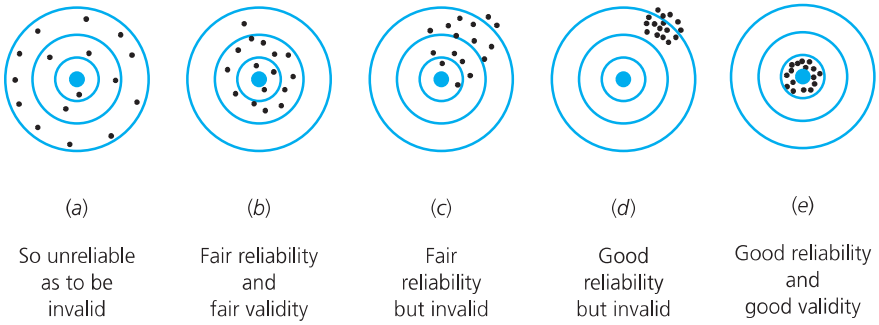


Figure 8.2 *Reliability and Validity*

The bull's-eye in each target represents the information that is desired. Each dot represents a separate score obtained with the instrument. A dot in the bull's-eye indicates that the information obtained (the score) is the information the researcher desires.

energy, anxiety, a different testing situation, and so on), and it is inevitable. Such factors result in **errors of measurement** (Figure 8.3).

Because errors of measurement are always present to some degree, researchers expect some variation in test scores (in answers or ratings, for example) when an instrument is administered to the same group more than once, when two different forms of an instrument are used, or even from one part of an instrument to another. Reliability estimates provide researchers with an idea of how much variation to expect. Such estimates are usually expressed as another application of the correlation coefficient known as a **reliability coefficient**.

As we mentioned earlier, a validity coefficient expresses the relationship between scores of the same individuals on two *different* instruments. A reliability coefficient also expresses a relationship, but this time it is between scores of the same individuals on the same instrument at two different times, or on two parts of the *same* instrument. The three best-known ways to obtain a reliability coefficient are the test-retest method, the equivalent-forms method; and the internal-consistency methods. Unlike other uses of the correlation coefficient, reliability coefficients must range from .00 to 1.00—that is, have no negative values.

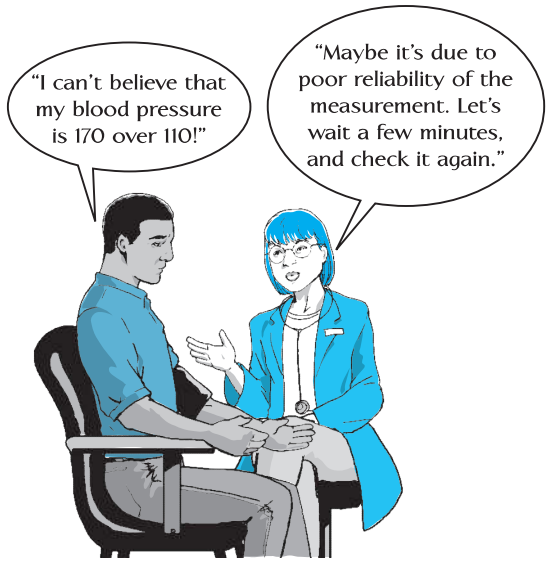


Figure 8.3 *Reliability of a Measurement*

TEST-RETEST METHOD

The **test-retest method** involves administering the same test twice to the *same* group after a certain time interval has elapsed. A reliability coefficient is then calculated to indicate the relationship between the two sets of scores obtained.

Reliability coefficients will be affected by the length of time that elapses between the two administrations of the test. The longer the time interval, the lower the reliability coefficient is likely to be, since there is a greater likelihood of changes in the individuals taking the test. In checking for evidence of test-retest reliability, an appropriate time interval should be selected. This interval should be that during which individuals would be assumed to retain their relative position in a meaningful group.

There is no point in studying, or even conceptualizing, a variable that fluctuates wildly in individuals for whom it is measured. When researchers assess someone as academically talented, for example, or skilled in typing or as having a poor self-concept, they assume that this characteristic will continue to differentiate individuals for some period of time. It is impossible to study a variable that has no stability in the individual.

Researchers do not expect all variables to be equally stable. Experience has shown that some abilities (such as writing) are more subject to change than others (such as abstract reasoning). Some personal characteristics (such as self-esteem) are considered to be more stable than others (such as teenage vocational interests). Mood is a variable that, by definition, is considered to be stable for short periods of time—a matter of minutes or hours. But even here, unless the instrumentation used is reliable, meaningful relationships with other (perhaps causal) variables will not be found. For most educational research, stability of scores over a two- to three-month period is usually viewed as sufficient evidence of test-retest reliability. In reporting test-retest reliability coefficients, therefore, the time interval between the two testings should always be reported.

EQUIVALENT-FORMS METHOD

When the **equivalent-forms method** is used, two different but equivalent (also called *alternate* or *parallel*) forms of an instrument are administered to the *same* group of individuals during the same time period. Although the questions are different, they should sample the same content and they should be constructed separately from each other. A reliability coefficient is then calculated between the two sets of scores obtained. A high coefficient would indicate strong evidence of reliability—that the two forms are measuring the same thing.

It is possible to combine the test-retest and equivalent-forms methods by giving two different forms of the same test with a time interval between the two administrations. A high reliability coefficient would indicate not only that the two forms are measuring the same sort of performance but also what we might expect with regard to consistency over time.

INTERNAL-CONSISTENCY METHODS

The methods mentioned so far all require two administration or testing sessions. There are several **internal-consistency methods** of estimating reliability, however, that require only a single administration of an instrument.

Split-half Procedure. The **split-half procedure** involves scoring two halves (usually odd items versus even items) of a test separately for each person and then calculating a correlation coefficient for the two sets of scores. The coefficient indicates the degree to which the two halves of the test provide the same results and hence describes the internal consistency of the test.

The reliability coefficient is calculated using what is known as the *Spearman-Brown prophecy formula.* A simplified version of this formula is as follows:

$$\text{Reliability of scores on total test} = \frac{2 \times \text{reliability for } \frac{1}{2} \text{ test}}{1 + \text{reliability for } \frac{1}{2} \text{ test}}$$

Thus, if we obtained a correlation coefficient of .56 by comparing one half of the test items to the other half, the reliability of scores for the total test would be:

$$\text{Reliability of scores on total test} = \frac{2 \times .56}{1 + .56} = \frac{1.12}{1.56} = .72$$

This illustrates an important characteristic of reliability. The reliability of a test (or any instrument) can generally be increased by the addition of more items, provided they are similar to the original ones.

Kuder-Richardson Approaches. Perhaps the most frequently employed method for determining internal consistency is the **Kuder-Richardson approach**, particularly formulas KR20 and KR21. The latter formula requires only three pieces of information—the number of items on the test, the mean, and the standard deviation. Note, however, that formula KR21 can be used only if it can be assumed that the items are of equal difficulty.* A frequently used version of the KR21 formula is the following:

$$\text{KR21 reliability coefficient} = \frac{K}{K-1}\left[1 - \frac{M(K-M}{K(SD^2)}\right]$$

where K = number of items on the test, M = mean of the set of test scores, and SD = standard deviation of the set of test scores.†

Although this formula may look somewhat intimidating, its use is actually quite simple. For example, if

*Formula KR20 does not require the assumption that all items are of equal difficulty, although it is harder to calculate. Computer programs for doing so are commonly available, however, and should be used whenever a researcher cannot assume that all items are of equal difficulty.

†See Chapter 10 for an explanation of standard deviation.

MORE ABOUT RESEARCH

Checking Reliability and Validity—An Example

The projective device (Picture Situation Inventory) described on pages 130 and 132 consists of 20 pictures, each scored on the variables *control need* and *communication* according to a point system. For example, here are some illustrative responses to picture 1 of Figure 7.23. The control need variable, defined as "motivated to control moment-to-moment activities of their students," is scored as follows:

- "I thought you would enjoy something special." (1 point)
- "I'd like to see how well you can do it." (2 points)
- "You and Tom are two different children." (3 points)
- "Yes, I would appreciate it if you would finish it." (4 points)
- "Do it quickly please." (5 points)

In addition to the appeal to content validity, there is some evidence in support of these two measures (control and communication).

Rowan studied relationships between the two scores and several other measures with a group of elementary school teachers.* She found that teachers scoring high on control need were more likely to (1) be seen by classroom observers as imposing themselves on situations and having a higher content emphasis, (2) be judged by interviewers as having more rigid attitudes of right and wrong, and (3) score higher on a test of authoritarian tendencies.

In a study of ability to predict success in a program preparing teachers for inner-city classrooms, evidence was found that the Picture Situation Inventory control score had predictive value.†

Correlations existed between the control score obtained on entrance to the program and a variety of measures subsequently obtained through classroom observation in student teaching and subsequent first-year teaching assignments. The most clear-cut finding was that those scoring higher in control need had classrooms observed as less noisy. The finding adds somewhat to the validity of the measurement, since a teacher with higher control need would be expected to have a quieter room.

The reliability of both measures was found to be adequate (.74 and .81) when assessed by the split-half procedure. When assessed by follow-up over a period of eight years, the consistency over time was considerably lower (.61 and .53), as would be expected.

*N. T. Rowan (1967). The relationship of teacher interaction in classroom situations to teacher personality variables. Unpublished doctoral dissertation. Salt Lake City: University of Utah.

†N. E. Wallen (1971). *Evaluation report to Step-TTT Project.* San Francisco, CA: San Francisco State University.

$K = 50$, $M = 40$, and $SD = 4$, the reliability coefficient would be calculated as shown below:

$$\text{Reliability} = \frac{50}{49}\left[1 - \frac{40(50-40)}{50(4^2)}\right]$$

$$= 1.02\left[1 - \frac{40(10)}{50(16)}\right]$$

$$= 1.02\left[1 - \frac{400}{800}\right]$$

$$= (1.02)(1 - .50)$$

$$= (1.02)(.50)$$

$$= .51$$

Thus, the reliability estimate for scores on this test is .51.

Is a reliability estimate of .51 good or bad? high or low? As is frequently the case, there are some benchmarks we can use to evaluate reliability coefficients. First, we can compare a given coefficient with the extremes that are possible. As you will recall, a coefficient of .00 indicates a complete absence of a relationship, hence no reliability at all, whereas 1.00 is the maximum possible coefficient that can be obtained. Second, we can compare a given reliability coefficient with the sorts of coefficients that are usually obtained for measures of the same type. The reported reliability coefficients for many commercially available achievement tests, for example, are typically .90 or higher when Kuder-Richardson formulas are used. Many classroom tests report reliability coefficients of .70 and higher. Compared to these figures, our obtained coefficient must be judged rather low. For research purposes, a useful rule of thumb is that reliability should be at least .70 and preferably higher.

Alpha Coefficient. Another check on the internal consistency of an instrument is to calculate an

TABLE 8.2 *Methods of Checking Validity and Reliability*

Validity ("Truthfulness")

Method	Procedure
Content-related evidence	Obtain expert judgment
Criterion-related evidence	Relate to another measure of the same variable
Construct-related evidence	Assess evidence on predictions made from theory

Reliability ("Consistency")

Method	Content	Time Interval	Procedure
Test-retest	Identical	Varies	Give identical instrument twice
Equivalent forms	Different	None	Give two forms of instrument
Equivalent forms/ retest	Different	Varies	Give two forms of instrument, with time interval between
Internal consistency	Different	None	Divide instrument into halves and score each or use Kuder-Richardson approach
Scoring observer agreement	Identical	None	Compare scores obtained by two or more observers or scorers

alpha coefficient (frequently called **Cronbach alpha** after the man who developed it). This coefficient (α) is a general form of the KR20 formula to be used in calculating the reliability of items that are not scored right versus wrong, as in some essay tests where more than one answer is possible.[3]

Table 8.2 summarizes the methods used in checking the validity and reliability of an instrument.

THE STANDARD ERROR OF MEASUREMENT (SEMeas)

The **standard error of measurement (SEMeas)** is an index that shows the extent to which a measurement would vary under changed circumstances (i.e., the amount of *measurement error*). Because there are many ways in which circumstances can vary, there are many possible standard errors for a given score. For example, the standard error will be smaller if it includes only error due to different content (internal-consistency or equivalent-forms reliability) than if it also includes error due to the passage of time (test-retest reliability). Under the assumption that errors of measurement are normally distributed (see p. 195 in Chapter 10), a range of scores can be determined that shows the amount of error to be expected.

For many IQ tests, the standard error of measurement over a one-year period and with different specific content is about 5 points. Over a 10-year period, it is about 8 points. This means that a score fluctuates considerably more the longer the time between measurements. Thus, a person scoring 110 can expect to have a score between 100 and 120 one year later; five years later, the score can be expected to be between 94 and 126 (see Figure 8.4). Note that we doubled the standard errors of measurement in computing the ranges within which the second score is expected to fall. This was done so we could be 95 percent sure that our estimates were correct.

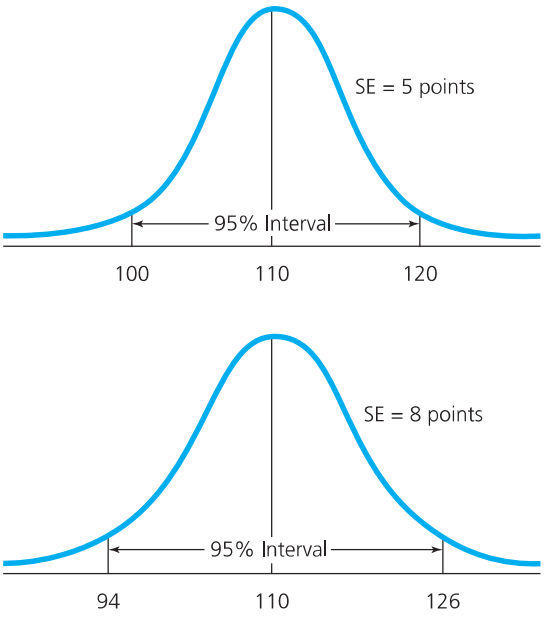


Figure 8.4 *Standard Error of Measurement*

The formula for the standard error of measurement is $SD\sqrt{1 - r_{11}}$ where SD = the standard deviation of scores and r_{11} = the reliability coefficient appropriate to the conditions that vary. In the above example, the standard error (SEMeas) of 5 in the first example was obtained as follows:

$$SD = 16, r_{11} = .90$$

$$SEM = 16\sqrt{1 - .90} = 16\sqrt{.10} = 16(.32) = 5.1$$

SCORING AGREEMENT

Most tests and many other instruments are administered with specific directions and are scored objectively, that is, with a key that requires no judgment on the part of the scorer. Although differences in the resulting scores with different administrators or scorers are still possible, it is generally considered highly unlikely that they would occur. This is not the case with instruments that are susceptible to differences in administration, scoring, or both, such as essay evaluations. In particular, instruments that use direct observation are highly vulnerable to observer differences. Researchers who use such instruments are obliged to investigate and report the degree of **scoring agreement**. Such agreement is enhanced by training the observers and by increasing the number of observation periods.

Instruments differ in the amount of training required for their use. In general, observation techniques require considerable training for optimum use. Such training usually consists of explaining and discussing the procedures involved, followed by trainees using the instruments as they observe videotapes or live situations. All trainees observe the same behaviors and then discuss any differences in scoring. This process, or some variation thereon, is repeated until independent observers reach an acceptable level of agreement. What is desired is a correlation of at least .90 among scorers or agreement of at least 80 percent. Usually, even after such training, 8 to 12 observation periods are required to get evidence of adequate reliability over time.

To further illustrate the concept of reliability, let's take an actual test and calculate the internal consistency of its items. Figure 8.5 presents an example of a

Directions: Read each of the following questions and write your answers on a separate sheet of paper. Suggested time to take the test is ten minutes.

1. There are two people in a room. The first is the son of the second person, but the second person is not the first person's father. How are the two people related?
2. Who is buried in Grant's tomb?
3. Some months have thirty days, some have thirty-one. How many have twenty-eight days?
4. If you had only one match and entered a dark room in which there was an oil lamp, an oil heater, and some firewood, which would you light first?
5. If a physician gave you three pills and told you to take one every half hour, how long would they last?
6. A person builds a house with four sides to it, a rectangular structure, with each side having a southern exposure. A big bear comes wandering by. What color is the bear?
7. A farmer has seventeen sheep. All but nine died. How many did he have left?
8. Divide 30 by $\frac{1}{2}$. Add 10. What is the correct answer?
9. Take two apples from three apples. What do you have?
10. How many animals of each species did Moses take aboard the Ark?

Figure 8.5 *The "Quick and Easy" Intelligence Test*

CONTROVERSIES IN RESEARCH

Is Consequential Validity a Useful Concept?

In recent years, increased attention has been given to a concept called *consequential validity,* originally proposed by Samuel Messick in 1989.* He intended not to change the core meaning of *validity,* but to expand it to include two new ideas: "value implications" and "social consequences."

Paying attention to value implications requires the "appraisal of the value implications of the construct label, of the theory underlying test interpretation, and the ideologies in which the theory is imbedded."† This involves expanding the idea of construct-related evidence of validity that we discussed on pages 153–154. *Social consequences* refers to "the appraisal of both potential and actual social consequences of applied testing."

Disagreement with Messick has been primarily with regard to applying his proposal. Using his experience as a developer of a widely used college admissions test battery (ACT) as an example, Reckase systematically analyzed the feasibility of using this concept. He concluded that, although difficult, the critical analysis of value implications is both feasible and useful.‡

However, he argued that assessing the cause-and-effect relationships implied in determining potential and actual social consequences of the use of a test is difficult or impossible, even with a clear intended use such as determining college admissions. Citing the concern of the National Commission on Testing and Public Policy that such tests often undermine vital social policies,§ he argues that obtaining the necessary data seems unlikely and that, by definition, appraising unintended consequences is not possible ahead of time, because one does not know what they are.

What do you think of Messick's proposal?

*S. Messick (1989). Consequential validity. In R. L. Linn (Ed.). *Educational measurement,* 3rd ed. New York: American Council on Education, pp. 13–103.
†Ibid., p. 20.
‡M. D. Reckase (1998). Consequential validity from the test developer's perspective. *Educational Measurement Issues and Practice, 17* (2): 13–16.
§National Commission on Testing and Public Policy. *From gatekeeper to gateway: Transforming testing in America* (Technical report). Chestnut Hill, MA: Boston College.

non-typical intelligence test that we have adapted. Follow the directions and take the test. Then we will calculate the split-half reliability.

Now look at the answer key in the footnote at the bottom of page 161. Give yourself one point for each correct answer. Assume, for the moment, that a score on this test provides an indication of intelligence. If so, each item on the test should be a partial measure of intelligence. We could, therefore, divide the 10-item test into two 5-item tests. One of these 5-item tests can consist of all the odd-numbered items, and the other 5-item test can consist of all the even-numbered items. Now, record your score on the odd-numbered items and also on the even-numbered items.

We now want to see if the odd-numbered items provide a measure of intelligence similar to that provided by the even-numbered items. If they do, your scores on the odd-numbered items and the even-numbered items should be pretty close. If they are not, then the two 5-item tests do not give consistent results. If this is the case, then the total test (the 10 items) probably does not give consistent results either, in which case the score could not be considered a reliable measure.

Person	Score on five-item test 1 (#1, 3, 5, 7, 9)	Score on five-item test 2 (#2, 4, 6, 8, 10)
You	_____	_____
#1	_____	_____
#2	_____	_____
#3	_____	_____
#4	_____	_____
#5	_____	_____

Figure 8.6 *Reliability Worksheet*

Ask five other people to take the test. Record their scores on the odd and even sets of items, using the worksheet shown in Figure 8.6.

Take a look at the scores on each of the five-item sets for each of the five individuals, and compare them with your own. What would you conclude about the reliability of the scores? What would you say about any

inferences about intelligence a researcher might make based on scores on this test? Could they be valid?*

Note that we have examined only one aspect of reliability (internal consistency) for results of this test. We still do not know how much a person's score might change if we gave the test at two different times (test-retest reliability). We could get a different indication of reliability if we gave one of the five-item tests at one time and the other 5-item test at another time to the same people (equivalent-forms/retest reliability). Try to do this with a few individuals, using a worksheet like the one shown in Figure 8.6.

Researchers typically use the procedures just described to establish reliability. Normally, however, they test many more people (at least 100). You should also realize that most tests would have many more than 10 items, since longer tests are usually more reliable than short ones, presumably because they provide a larger sampling of a person's behavior.

In sum, we hope it is clear that a major aspect of research design is the obtaining of reliable and valid information. Because both reliability and validity depend on the way in which instruments are used and on the inferences researchers wish to make from them, researchers can never simply assume that their instrumentation will provide satisfactory information. They can have more confidence if they use instruments on which there is previous evidence of reliability and validity, provided they use the instruments in the same way—that is, under the same conditions as existed previously. Even then, researchers cannot be sure; even when all else remains the same, the mere passage of time may have impaired the instrument in some way.

What this means is that there is no substitute for checking reliability and validity as a part of the research procedure. There is seldom any excuse for failing to check internal consistency, since the necessary information is at hand and no additional data collection is required. Reliability over time does, in most cases, require an additional administration of an instrument, but this can often be done. In considering this option, it should be noted that not all members of the sample need be retested, though this is desirable. It is better to retest a randomly selected subsample, or even a convenience subsample, than to have no evidence of retest reliability at all. Another option is to test and retest a different, though very similar, sample.

Obtaining evidence on validity is more difficult but seldom prohibitive. Content-related evidence can usually be obtained, since it requires only a few knowledgeable and available judges. It is unreasonable to expect a great deal of construct-related evidence to be obtained, but, in many studies, criterion-related evidence can be obtained. At a minimum, a second instrument should be administered. Locating or developing an additional means of instrumentation is sometimes difficult and occasionally impossible (for example, there is probably no way to validate a self-report questionnaire on sexual behavior), but the results are well worth the time and energy involved. As with retest reliability, a subsample can be used, or both instruments can be given to a different, but similar, sample.

VALIDITY AND RELIABILITY IN QUALITATIVE RESEARCH

While many qualitative researchers use many of the procedures we have described, some take the position that validity and reliability, as we have discussed them, are either irrelevant or not suited to their research efforts because they are attempting to describe a specific situation or event as viewed by a particular individual. They emphasize instead the honesty, believability, expertise, and integrity of the researcher. We maintain that all researchers should ensure that any inferences they draw that are based on data obtained through the use of an instrument are appropriate, credible, and backed up by evidence of the sort we have described in this chapter.

Specific methods for enhancing the validity and reliability of qualitative studies are discussed in Chapters 18, 19, and 21. Moreover, in the next chapter, we discuss the concept of *internal validity* and how it applies both to quantitative and qualitative research.

*You might want to assess the content validity of this test. How would you define *intelligence?* As you define the term, how would you evaluate this test as a measure of intelligence?
Answer Key for Q-E Intelligence Test on page 159. 1. Mother and son; 2. Ulysses S. Grant; 3. All of them; 4. The match; 5. One hour; 6. White; 7. Nine; 8. 70; 9. Two; 10. None. (It wasn't Moses, but Noah who took the animals on the Ark.)

Go back to the **INTERACTIVE AND APPLIED LEARNING** feature at the beginning of the chapter for a listing of interactive and applied activities. Go to the **Online Learning Center** at **www.mhhe.com/fraenkel8e** to take quizzes, practice with key terms, and review chapter content.

Main Points

VALIDITY

- It is important for researchers to use valid instruments, for the conclusions they draw are based on the information they obtain using these instruments.
- The term *validity,* as used in research, refers to the appropriateness, meaningfulness, correctness, and usefulness of any inferences a researcher draws based on data obtained through the use of an instrument.
- Content-related evidence of validity refers to judgments on the content and logical structure of an instrument as it is to be used in a particular study.
- Criterion-related evidence of validity refers to the degree to which information provided by an instrument agrees with information obtained on other, independent instruments.
- A criterion is a standard for judging; with reference to validity, it is a second instrument against which scores on an instrument can be checked.
- Construct-related evidence of validity refers to the degree to which the totality of evidence obtained is consistent with theoretical expectations.
- A validity coefficient is a numerical index representing the degree of correspondence between scores on an instrument and a criterion measure.
- An expectancy table is a two-way chart used to evaluate criterion-related evidence of validity.

RELIABILITY

- The term *reliability,* as used in research, refers to the consistency of scores or answers provided by an instrument.
- Errors of measurement refer to variations in scores obtained by the same individuals on the same instrument.
- The test-retest method of estimating reliability involves administering the same instrument twice to the same group of individuals after a certain time interval has elapsed.
- The equivalent-forms method of estimating reliability involves administering two different, but equivalent, forms of an instrument to the same group of individuals at the same time.
- The internal-consistency method of estimating reliability involves comparing responses to different sets of items that are part of an instrument.
- Scoring agreement requires a demonstration that independent scorers can achieve satisfactory agreement in their scoring.
- The standard error of measurement is a numerical index of measurement error.

Key Terms

alpha coefficient 158
concurrent validity 152
construct-related evidence of validity 153
content-related evidence of validity 150
correlation coefficient 152
criterion 152
criterion-related evidence of validity 152
Cronbach alpha 158
equivalent-forms method 156
errors of measurement 155
expectancy table 153
internal-consistency methods 156
Kuder-Richardson approach 156
predictive validity 152
reliability 154

reliability coefficient 155
scoring agreement 159
split-half procedure 156
standard error of measurement (SEMeas) 158
test-retest method 155
validity 148
validity coefficient 152

For Discussion

1. We point out in the chapter that scores from an instrument may be reliable but not valid, yet not the reverse. Why would this be so?
2. What type of evidence—content-related, criterion-related, or construct-related—do you think is the easiest to obtain? the hardest? Why?
3. In what way(s) might the format of an instrument affect its validity?
4. "There is no single piece of evidence that satisfies construct-related validity." Is this statement true? If so, explain why.
5. Which do you think is harder to obtain, validity or reliability? Why?
6. Might reliability ever be more important than validity? Explain.
7. How would you assess the Q-E Intelligence Test in Figure 8.4 with respect to validity? Explain.
8. The importance of using *valid* instruments in research cannot be overstated. Why?

Notes

1. N. E. Wallen, M. C. Durkin, J. R. Fraenkel, A. J. McNaughton, and E. I. Sawin (1969). *The Taba Curriculum Development Project in Social Studies: Development of a comprehensive curriculum model for social studies for grades one through eight, inclusive of procedures for implementation and dissemination.* Menlo Park, CA: Addison-Wesley, p. 307.

2. N. E. Gronlund (1988). *How to construct achievement tests,* 4th ed. Englewood Cliffs, NJ: Prentice Hall, p. 140.

3. See L. J. Cronbach (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16:* 297–334.

Research Exercise 8: Validity and Reliability

Use Problem Sheet 8 to describe how you plan to check on the validity and reliability of scores obtained with your instruments. If you plan to use an existing instrument, summarize what you have been able to learn about the validity and reliability of results obtained with it. If you plan to develop an instrument, explain how you will attempt to ensure validity and reliability. In either case, explain how you will obtain evidence to check validity and reliability.

Problem Sheet 8

Validity and Reliability

1. If you plan to use an *existing* instrument, describe what you have learned about the validity and reliability of scores obtained with this instrument.

   _____

   _____

2. If you plan to *develop* an instrument, explain how you will try to ensure the validity and reliability of results obtained with this instrument by using one or more of the tips described on page 114 (*specify which*).

   _____

   _____

3. If you have not already indicated so above for each instrument that you plan to use, tell specifically how you will check for:

   a. internal consistency _____

   _____

   b. stability (reliability over time) _____

   _____

   c. validity _____

   _____

An electronic version of this Problem Sheet that you can fill in and print, save, or e-mail is available on the Online Learning Center at www.mhhe.com/fraenkel8e.