1-1c Time in Research

Time is an important element of any research design, and here I want to introduce one of the most fundamental distinctions in research design nomenclature: cross-sectional versus longitudinal studies. A cross-sectional study is one that takes place at a single point in time. In effect, you are taking a slice or cross-section of whatever it is you’re observing or measuring. A longitudinal study is one that takes place over time. In a longitudinal study, you measure your research participants on at least two separate occasions or at least two points in time. When you measure at different time points, we often say that you are measuring multiple waves of measurement. Just as with the repeated motion of the waves in the ocean or of waving with your hand, multiple waves of measurement refers to taking measurements on a variable several times.

A further distinction is made between two types of longitudinal designs: repeated measures and time series. There is no universally agreed-upon rule for distinguishing between these two terms, but in general, if you have two or a few waves of measurement, you are using a repeated measures design. If you have many waves of measurement over time, you have a time series (Box & Jenkins, 1976). How many is many? Usually, you wouldn’t use the term time series unless you had at least 20 waves of measurement. With fewer waves than that, you would usually call it a repeated measures design.

1-1d Variables

You won’t be able to do much in research unless you know how to talk about variables. A variable is any entity that can take on different values (Marriott, 1990). Okay, so what does that mean? Anything that can vary can be considered a variable. For instance, age can be considered a variable because age can take different values for different people or for the same person at different times. Similarly, country can be considered a variable because a person’s country can be assigned a value.

Variables aren’t always quantitative or numerical. The variable gender consists of two text values: male and female. If it is useful, quantitative values can be assigned instead of (or in place of) the text values, but it’s not necessary to assign numbers for something to be a variable. It’s also important to realize that variables aren’t the only things measured in the traditional sense. For instance, in much social research and in program evaluation, the treatment or program (that is, the “cause”) is considered to be a variable. An educational program can have varying amounts of time on task, classroom settings, student-teacher ratios, and so on. Therefore, even the program can be considered a variable, which can be made up of a number of subvariables.

An attribute is a specific value on a variable. For instance, the variable sex or gender has two attributes (male and female), or the variable agreement might be defined as having five attributes:

1 = strongly disagree
2 = disagree
3 = neutral
4 = agree
5 = strongly agree

Another important distinction having to do with the term variable is the distinction between an independent and dependent variable (Marriott, 1990). This distinction is particularly relevant when you are investigating cause-effect relationships. It took me the longest time to learn this distinction. (Of course, I’m someone who gets confused about the signs for arrivals and departures at airports—do I go to arrivals because I’m arriving at the airport, or does the person I’m picking up go to arrivals because he or she is arriving on the plane?) I originally thought that an independent variable was one that would be free to vary or respond to some program or treatment, and that a dependent variable must be one that depends on my efforts (that is, it’s the treatment). However, this is entirely backward! In fact, the independent variable is what you (or nature) manipulates—a treatment or program or cause. The dependent variable is what you presume to be affected by the independent variable that you manipulate. For instance, a program or treatment is usually an independent variable.

dependent variable The variable that you manipulate. For instance, a program or treatment is typically an independent variable.

variable Any entity that can take on different values. For instance, age can be considered a variable because age can take on different values for different people at different times.

attribute A specific value of a variable. For instance, the variable sex or gender has two attributes: male and female.

independent variable The variable that you manipulate. For instance, a program or treatment is typically an independent variable.
by the independent variable—your effects or outcomes. For example, if you are studying the effects of a new educational program on student achievement, the program is the independent variable and your measures of achievement are the dependent ones. Or if you are looking at the effects of a new surgical treatment for cancer on rates of mortality for that cancer, the independent variable would be the surgical treatment and the dependent variable would be the mortality rates. The independent variable is what you (or nature) do, and the dependent variable is what results from that.

Finally, the attributes of a variable should be both exhaustive and mutually exclusive. Each variable's attributes should be exhaustive, meaning that they should include all possible answerable responses. For instance, if the variable is religion and the only options are Protestant, Jewish, and Muslim, there are quite a few religions I can think of that haven't been included. The list does not exhaust all possibilities. On the other hand, if you exhaust all the possibilities with some variables—religion being one of them—you would simply have too many responses. The way to deal with this is to list the most common attributes and then use a general category like Other to account for all remaining ones. In addition to being exhaustive, the attributes of a variable should be mutually exclusive, meaning that no respondent should be able to have two attributes simultaneously. Although this might seem obvious, it is often rather tricky in practice. For instance, you might be tempted to represent the variable Educational Status by asking the respondent to check one of the following response attributes: High School Degree, Some College, Two-Year College Degree, Four-Year College Degree, and Graduate Degree. However, these attributes are not mutually exclusive—a person who has a two-year or four-year college degree also could correctly check Some College! In fact, if someone went to college, got a two-year degree and then got a four-year degree, he or she could check all three. The problem here is that you ask the respondent to provide a single response to a set of attributes that are not mutually exclusive. But don't researchers often use questions on surveys that ask the respondent to check all that apply and then list a series of categories? Yes, but technically speaking, each of the categories in a question like that is its own variable and is treated dichotomously as either checked or unchecked—as attributes that are mutually exclusive.

1-1e Types of Relationships

A relationship refers to the correspondence between two variables (see Section 1-1d, "Variables"). When you talk about types of relationships, you can mean that in at least two ways: the nature of the relationship or the pattern of it.

The Nature of a Relationship We start by making a distinction between two types of relationships: a correlational relationship and a causal relationship. A correlational relationship simply says that two things perform in a synchronized manner. For instance, economists often talk of a correlation between inflation and unemployment. When inflation is high, unemployment also tends to be high. When inflation is low, unemployment also tends to be low. The two variables are correlated, but knowing that two variables are correlated does not tell whether one causes the other. For instance, there is a correlation between the number of roads built in Europe and the number of children born in the United States. Does that mean that if fewer children are desired in the United States, there should be a cessation of road building in Europe? Or does it mean that if there aren't enough roads in Europe, U.S. citizens should be encouraged to have more babies? Of course not. (At least, I hope not.) Although there is a relationship between the number of roads built and the number of babies, it's not likely that the relationship is a causal one. A causal relationship is a synchronized relationship between two variables just as a correlational relationship is, but in a causal relationship, we say that one variable causes the other to occur.

This leads to consideration of what is often termed the third-variable problem. In the previous example, it may be that a third variable is causing both the building of roads and the birth rate, and causing the correlation that is observed. For instance, perhaps the
general world economy is responsible for both. When the economy is good, more roads are built in Europe, and more children are born in the United States. The key lesson here is that you have to be careful when you interpret correlations. If you observe a correlation between the number of hours students use the computer to study and their grade point averages (with high computer users getting higher grades), you cannot assume that the relationship is causal—that computer use improves grades. In this case, the third variable might be socioeconomic status—richer students, who have greater resources at their disposal, tend to both use computers and make better grades. Resources drive both use and grades; computer use doesn’t cause the change in the grade point averages.

Several terms describe the major different types of patterns one might find in a relationship. First, there is the case of no relationship at all where, if you know the values on one variable, you don’t know anything about the values on the other. For instance, I suspect that there is no relationship between the length of the lifeline on your hand and your grade point average. If I know your GPA, I don’t have any idea how long your lifeline is. Figure 1-1a shows the case where there is no relationship.

Then, there is the positive relationship. In a positive relationship, high values on one variable are associated with high values on the other, and low values on one are associated with low values on the other. Figure 1-1b shows an idealized positive relationship between years of education and the salary one might expect to be making.

On the other hand, a negative relationship implies that high values on one variable are associated with low values on the other. This is also sometimes termed an inverse relationship. Figure 1-1c shows an idealized negative relationship between a measure of self-esteem and a measure of paranoia in psychiatric patients.

Positive relationship
A relationship between variables in which high values for one variable are associated with high values on another variable, and low values are associated with low values.

Negative relationship
A relationship between variables in which high values for one variable are associated with low values on another variable.

Four Types of Possible Relationships
(a) No relationship. (b) A positive relationship. (c) A negative relationship. (d) A curvilinear relationship.
These are the simplest patterns of relationships that might typically be estimated in research. However, the pattern of a relationship can be more complex than this. For instance, Figure 1-1d (see page 7) shows a relationship that changes over the range of both variables, a curvilinear relationship. In this example, the horizontal axis represents dosage of a drug for an illness, and the vertical axis represents a severity of illness measure. As the dosage rises, the severity of illness goes down. But at some point, the patient begins to experience negative side effects associated with too high a dosage, and the severity of illness begins to increase again.

### 1-1f Hypotheses

An hypothesis is a specific statement of prediction. It describes in concrete (rather than theoretical) terms what you expect to happen in your study. Not all studies have hypotheses. Sometimes, a study is designed to be exploratory (see Section 1-2b, “Deduction and Induction,” later in this chapter). There is no formal hypothesis, and perhaps the purpose of the study is to explore some area more thoroughly to develop some specific hypothesis or prediction that can be tested in future research. A single study may have one or many hypotheses.

Actually, whenever I talk about an hypothesis, I am really thinking simultaneously about two hypotheses. Let’s say that you predict that there will be a relationship between two variables in your study. The way to set up the hypothesis test is to formulate two hypothesis statements: one that describes your prediction and one that describes all the other possible outcomes with respect to the hypothesized relationship. Your prediction might be that variable A and variable B will be related. (You don’t care whether it’s a positive or negative relationship.) Then the only other possible outcome would be that variable A and variable B are not related. Usually, the hypothesis that you support (your prediction) is called the alternative hypothesis, and the hypothesis that describes the remaining possible outcomes is termed the null hypothesis. Sometimes, a notation like \( H_A \) or \( H_1 \) is used to represent the alternative hypothesis or your prediction, and \( H_0 \) or \( H_0 \) to represent the null case. You have to be careful here, though. In some studies, your prediction might well be that there will be no difference or change. In this case, you are essentially trying to find support for the null hypothesis, and you are opposed to the alternative (Marriott, 1990).

If your prediction specifies a direction, the null hypothesis automatically includes both the no-difference prediction and the prediction that would be opposite in direction to yours. This is called a one-tailed hypothesis. For instance, let’s imagine that you are investigating the effects of a new employee-training program and that you believe one of the outcomes will be that there will be less employee absenteeism. Your two hypotheses might be stated like this:

The null hypothesis for this study is

\[
H_0: \text{As a result of the XYZ company employee-training program, there will either be no significant difference in employee absenteeism or there will be a significant increase,}
\]

which is tested against the alternative hypothesis:

\[
H_A: \text{As a result of the XYZ company employee-training program, there will be a significant decrease in employee absenteeism.}
\]

In Figure 1-2, this situation is illustrated graphically. The alternative hypothesis—your prediction that the program will decrease absenteeism—is shown there. The null must account for the other two possible conditions: no difference, or an increase in absenteeism. The figure shows a hypothetical distribution of absenteeism differences. That is, a value of zero means that there has been no difference in absenteeism observed, a positive value means that absenteeism has increased, and a negative value means it has decreased. The term one-tailed refers to the tail of the distribution on the outcome variable.
When your prediction does not specify a direction, you have a two-tailed hypothesis. For instance, let's assume you are studying a new drug treatment for depression. The drug has gone through some initial animal trials, but it has not yet been tested on humans. You believe (based on theory and the previous research) that the drug will have an effect, but you are not confident enough to hypothesize a direction and say the drug will reduce depression. (After all, you've seen more than enough promising drug treatments come along that eventually were shown to have severe side effects that actually worsened symptoms.) In this case, you might state the two hypotheses like this:

The null hypothesis for this study is:

H₀: As a result of 300mg/day of the ABC drug, there will be no significant difference in depression,

which is tested against the alternative hypothesis:

H₁: As a result of 300mg/day of the ABC drug, there will be a significant difference in depression.

Figure 1-3 illustrates this two-tailed prediction for this case. Again, notice that the term two-tailed refers to the tails of the distribution for your outcome variable.

The important thing to remember about stating hypotheses is that you formulate your prediction (directional or not), and then you formulate a second hypothesis that is
two-tailed hypothesis
A hypothesis that does not specify a direction. For example, if your hypothesis is that your program or intervention will have an effect on an outcome, but you are unwilling to specify whether that effect will be positive or negative, you are using a two-tailed hypothesis.
9. Which sampling process begins with the selection of a single random number and assumes that the characteristics being measured are randomly distributed in the population?
   a. simple random sampling
   b. stratified random sampling
   c. systematic random sampling
   d. cluster random sampling

10. What is the sampling technique that is best used when there is a large geographical area to cover?
    a. simple random sampling
    b. stratified random sampling
    c. systematic random sampling
    d. cluster (area) random sampling

11. The normal distribution is often referred to as the "bell curve" because it provides the "ring of truth."
    a. True
    b. False

12. Surveys reported in the media almost always mention that the numbers presented are accurate within a few percentage points. The statistic used to determine the accuracy of such results is called the standard error.
    a. True
    b. False

13. A study based on a nonprobability sampling method can never be considered representative of the population.
    a. True
    b. False

14. A researcher was trying to study a hard-to-research population (for example, homeless adolescents, migrant workers, cocaine dealers, etc.). The researcher decided to try sampling by tapping the social network of the local population, beginning with the first person she could find and then asking that person to help identify others, who would then be asked to further identify possible participants. This researcher is using a sampling technique known as "avalanche sampling" because pretty soon she could expect to have a very large number of participants.
    a. True
    b. False

15. The main advantage of multistage sampling is that it combines sophistication with efficiency, while the main disadvantage is that it can be complex and difficult to explain to nontechnical audiences.
    a. True
    b. False
MEASUREMENT

The Theory of Measurement

Survey Research

Scales and Indexes

Qualitative and Unobtrusive Measures
Chapter Outline

3-1 Construct Validity
   3-1a Measurement Validity Types
      Translation Validity
      Criterion-Related Validity
   3-1b Idea of Construct Validity
   3-1c Convergent and Disciminant Validity
      Convergent Validity
      Discriminant Validity
      Putting It All Together
   3-1d Threats to Construct Validity
      Inadequate Preoperational Explication of Constructs
      Mono-Operation Bias
      Mono-Method Bias
      Interaction of Different Treatments
      Interaction of Testing and Treatment
      Restricted Generalizability across Constructs
      Confounding Constructs and Levels of Constructs
      The Social Threats to Construct Validity

3-2 Reliability
   3-2a True Score Theory
   3-2b Measurement Error
      What Is Random Error?
      What Is Systematic Error?
      Reducing Measurement Error
   3-2c Theory of Reliability
   3-2d Types of Reliability
      Inter-Rater or Inter-Observer Reliability
      Test-Retest Reliability
      Parallel-Forms Reliability
      Internal Consistency Reliability
   3-2e Reliability and Validity

3-3 Levels of Measurement
   3-3a Why Is Level of Measurement Important?

Summary

Suggested Websites

Key Terms

Review Questions
measurement is the process of observing and recording the observations that are collected as part of a research effort. In this chapter, I focus on how we think about and assess quality of measurement. In the section on construct validity, I present the theory of what constitutes a good measure. In the section on reliability of measurement, I consider the consistency or dependability of measurement, including the idea of true score theory and several ways to estimate reliability. In the section on levels of measurement, I describe the four major levels of measurement: nominal, ordinal, interval, and ratio.

3.1 Construct Validity

Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations are based (Cook & Campbell, 1979; Cronbach & Meehl, 1955; Shadish, Cook, & Campbell, 2002). Whoa! Can you believe that the term operationalization has eight syllables? That’s a mouthful. What does it mean? The key is in the root term “operation.” In everyday language, an operation is a specific procedure, the steps you follow to accomplish something. In research, you “operationalize” a concept you want to measure when you describe exactly how you’ll measure it. An operationalization is your translation of the idea or construct into something real and concrete. Let’s say you have an idea for a treatment or program you would like to create. The operationalization is the program or treatment itself, as it exists after you create it. Construct validity is the degree to which the actual (operationalized) program accurately reflects the ideal (the program as you conceptualize or envision it). Imagine that you want to measure the construct of self-esteem. You have an idea of what self-esteem means. Perhaps you construct a 10-item paper-and-pencil instrument to measure that self-esteem concept. The instrument is the operationalization; it’s the translation of the idea of self-esteem into something concrete. The construct validity question here would be how well the 10-item instrument (the operationalization) reflects the idea you had of self-esteem. I’ll cover this in more detail later in this chapter, but I didn’t want to start the chapter by confusing you with an 8-syllable word.

Like external validity (see the discussion in Chapter 2, “Sampling”), construct validity is related to generalizing. However, where external validity involves generalizing from your study context to other people, places, or times, construct validity involves generalizing from your program or measures to the concept or idea of your program or measures. You might think of construct validity as a labeling issue. When you implement a program that you call a Head Start program, is your label an accurate one? When you measure what you call self-esteem is that what you are really measuring?

I would like to address two major issues here. The first is the more straightforward one. I’ll discuss several ways of thinking about the idea of construct validity, and several metaphors that might provide you with a sense of how profound and rich this idea is. Then, I’ll discuss the major construct validity threats, the kinds of arguments your critics are likely to raise when you make a claim that your program or measure is valid.

In this text, as in most research methods texts, construct validity is presented in the section on measurement; it is typically presented as one of many different types of validity (for example, face validity, predictive validity, or concurrent validity) that you might
want to be sure your measures have. I don’t see it that way at all. I see construct validity as the overarching quality of measurement with all of the other measurement validity labels falling beneath it. And construct validity is not limited only to measurement. As I’ve already suggested, it is as relevant to the independent variable—the program or treatment—as it is to the dependent variable. So, I’ll try to make some sense of the various measurement validity types in this chapter and try to move you to think instead of the validity of any operationalization as falling within the general category of construct validity, with a variety of subcategories and subtypes.

### 3-1a Measurement Validity Types

There’s an awful lot of confusion in the methodological literature that stems from the wide variety of labels used to describe the validity of measures. I want to make two cases here. First, it’s dumb to limit our scope only to the validity of measures. I really want to talk about the validity of any operationalization. That is, any time you translate a concept or construct into a functioning and operating reality (the operationalization), you need to be concerned about how well you performed the translation. This issue is as relevant when talking about treatments or programs as it is when talking about measures. In fact, come to think of it, you could also think of sampling in this way. The population of interest in your study is the construct, and the sample is your operationalization. If you think of it this way, you are essentially talking about the construct validity of the sampling, and construct validity merges with the idea of external validity as discussed in Chapter 2, “Sampling.” The construct validity question, “How well does my sample represent the idea of the population?” merges with the external validity question, “How well can I generalize from my sample to the population?” Second, I want to use the term construct validity to refer to the general case of translating any construct into an operationalization. Let’s use all of the other validity types to reflect different ways you can demonstrate different aspects of construct validity.

With all that in mind, here’s a list of the validity types that are typically mentioned in texts and research papers when talking about the quality of measurement and how I would organize and categorize them:

**Construct Validity**

- Translation validity
  - Face validity
  - Content validity
- Criterion-related validity
  - Predictive validity
  - Concurrent validity
  - Convergent validity
  - Discriminant validity

I have to warn you here that I made this list up. I’d never heard of translation validity before, but I needed a good name to summarize what both face and content validity are getting at, and that one seemed sensible. (See how easy it is to be a methodologist?) All of the other labels are commonly known, but the way I’ve organized them is different than I’ve seen elsewhere.

Let’s see if I can make some sense out of this list. First, as mentioned previously, I would like to use the term construct validity to be the overarching category. Construct validity is the approximate truth of the conclusion that your operationalization accurately reflects its construct. All of the other validity types essentially address some aspect of this general issue (which is why I’ve subsumed them under the general category of construct validity). Second, I make a distinction between two broad types of construct validity: translation validity and criterion-related validity. That’s because I think these correspond to the two major ways you can assure/assess the validity of an operationalization.
In translation validity, you focus on whether the operationalization is a good translation of the construct. This approach is definitional in nature; it assumes you have a good, detailed definition of the construct and that you can check the operationalization against it. In criterion-related validity, you examine whether the operationalization behaves the way it should according to some criteria based on your understanding of the construct. This type of validity is a more relational approach to construct validity. It assumes that your operationalization should function in predictable ways in relation to other operationalizations based upon your theory of the construct. (If all this seems a bit dense, hang in there until you’ve gone through the following discussion and then come back and reread this paragraph.) Let’s go through the specific validity types.

Translation validity In essence, both of the translation validity types (face and content validity) attempt to assess the degree to which you accurately translated your construct into the operationalization. Let’s look at the two types of translation validity.

Face Validity In face validity, you look at the operationalization and see whether on its face it seems like a good translation of the construct. In other words, does the way you are measuring the construct appear to measure what you want it to? This is probably the weakest way to try to demonstrate construct validity. For instance, you might look at a measure of math ability, read through the questions, and decide it seems like this is a good measure of math ability (the label math ability seems appropriate for this measure). Or you might observe a teenage pregnancy-prevention program and conclude that it is indeed a teenage pregnancy-prevention program. Of course, if this were all you did to assess face validity, it would clearly be weak evidence because it is essentially a subjective judgment call. (Note that just because it is weak evidence doesn’t mean that it is wrong. You need to rely on your subjective judgment throughout the research process. It’s just that this form of judgment won’t be especially convincing to others.) You can improve the quality of a face-validity assessment considerably by making it more systematic. For instance, if you were trying to assess the face validity of a math-ability measure, it would be more convincing if you sent the test to a carefully selected sample of experts on math-ability testing and they all reported back with the judgment that your measure appears to be a good measure of math ability.

Content Validity In content validity, you essentially check the operationalization against the relevant content domain for the construct. The content domain is like a comprehensive checklist of the traits of your construct. This approach assumes that you have a good detailed description of the content domain, something that’s not always true. Let’s look at an example where it is true. You might lay out all of the characteristics of a teenage pregnancy-prevention program. You would probably include in this domain specification the definition of the target group, a description of whether the program is preventive in nature (as opposed to treatment-oriented), and the content that should be included, such as basic information on pregnancy, the use of abstinence, birth control methods, and so on. Then, armed with these characteristics, you create a type of checklist when examining your program. Only programs that have these characteristics can legitimately be defined as teenage pregnancy-prevention programs. This all sounds fairly straightforward, and for many operationalizations, it may be. However, for other constructs (such as self-esteem or intelligence), it will not be easy to decide which characteristics constitute the content domain.

Criterion-Related Validity In criterion-related validity, you check the performance of your operationalization against some criterion. How is this different from translation validity? In translation validity, the question is, how well did you translate the idea of the construct into its manifestation? No other measure comes into play. In criterion-related validity, you usually make a prediction about how the operationalization will perform on some other measure based on your theory of the construct. The differences among the criterion-related validity types are in the criteria they use as the standard for judgment.

Translation validity A type of construct validity related to how well you translated the idea of your measure into its operationalization.

Criterion-related validity The validation of a measure based on its relationship to another independent measure as predicted by your theory of how the measures should behave.

Face validity A type of validity that assures that "on its face" the operationalization seems like a good translation of the construct.

Content validity A check of the operationalization against the relevant content domain for the construct.
For example, think again about measuring self-esteem. For content validity, you would try to describe all the things that self-esteem is in your mind and translate that into a measure. You might say that self-esteem involves how good you feel about yourself, that it includes things like your self-confidence and the degree to which you think positively about yourself. You could translate these notions into specific questions, a translation validity approach. On the other hand, you might reasonably expect that people with high self-esteem, as you interpret it, would tend to act in certain ways. You might expect that you could distinguish them from people with low self-esteem. For instance, you might argue that high self-esteem people will volunteer for a task that requires self-confidence (such as speaking in public). Notice that in this case, you validate your self-esteem measure by demonstrating that it is correlated with some other independent indicator (raising hands to volunteer) that you theoretically expect high self-esteem people to exhibit. This is the essential idea of criterion-related validity: validating a measure based on its relationship to another independent measure.

**Predictive Validity**  In predictive validity, you assess the operationalization's ability to predict something it should theoretically be able to predict. For instance, you might theorize that a measure of math ability should be able to predict how well a person will do in an engineering-based profession. You could give your measure to experienced engineers and see whether there is a high correlation between scores on the measure and their salaries as engineers. A high correlation would provide evidence for predictive validity; it would show that your measure can correctly predict something that you theoretically think it should be able to predict.

**Concurrent Validity**  In concurrent validity, you assess the operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between. For example, if you come up with a way of assessing depression, your measure should be able to distinguish between people who are diagnosed as depressed and those diagnosed paranoid schizophrenic. If you want to assess the concurrent validity of a new measure of empowerment, you might give the measure to both migrant farm workers and to the farm owners, theorizing that your measure should show that the farm owners are higher in empowerment. As in any discriminating test, the results are more powerful if you are able to show that you can discriminate between two similar groups than if you can show that you can discriminate between two groups that are very different.

**Convergent Validity**  In convergent validity, you examine the degree to which the operationalization is similar to (converges on) other operationalizations to which it theoretically should be similar. For instance, to show the convergent validity of a Head Start program, you might gather evidence that shows that the program is similar to other Head Start programs. To show the convergent validity of a test of arithmetic skills, you might correlate the scores on your test with scores on other tests that purport to measure basic math ability, where high correlations would be evidence of convergent validity.

**Discriminant Validity**  In discriminant validity, you examine the degree to which the operationalization is not similar to (diverges from) other operationalizations that it theoretically should not be similar to. For instance, to show the discriminant validity of a Head Start program, you might gather evidence that shows that the program is not similar to other early childhood programs that don't label themselves as Head Start programs. To show the discriminant validity of a test of arithmetic skills, you might correlate the scores on your test with scores on tests of verbal ability, where low correlations would be evidence of discriminant validity.

### 3-1b Idea of Construct Validity

Construct validity refers to the degree to which inferences can legitimately be made from the operationalizations in your study to the theoretical constructs on which those operationalizations were based. (I know I've said this before, but it never hurts to repeat some-
thing, especially when it sounds complicated.) I find that it helps me when thinking about construct validity to make a distinction between two broad territories that I call the land of theory and the land of observation, as illustrated in Figure 3-1. The land of theory is what goes on inside your mind, and your attempt to explain or articulate this to others. It is all of the ideas, theories, hunches, and hypotheses you have about the world. In the land of theory, you think of the program or treatment as it should be. In the land of theory, you have the idea or construct of the outcomes or measures you are trying to affect. The land of observation, on the other hand, consists of what you actually see or measure happening in the world around you and the public manifestations of that world. In the land of observation, you find your actual program or treatment, and your actual measures or observational procedures. If you have construct validity, then you have constructed the land of observation based on your ideas in the land of theory. You developed the program to reflect the kind of program you had in mind. You created the measures to get at what you wanted to get at.

Construct validity is an assessment of how well your actual programs or measures reflect your ideas or theories—how well the bottom of Figure 3-1 reflects the top. Why is this important? Because when you think about the world or talk about it with others (land of theory), you are using words that represent concepts. If you tell parents that a special type of math tutoring will help their child do better in math, you are communicating at the level of concepts or constructs. You aren’t describing in operational detail the specific things that the tutor will do with their child. You aren’t describing the specific questions that will be on the math test on which their child will excel. You are talking in general terms, using constructs. If you based your recommendation on research that showed that the special type of tutoring improved children’s math scores, you would want to be sure that the type of tutoring you are referring to is the same as what that study implemented and that the type of outcome you’re saying should occur was the type the study measured. Otherwise, you would be mislabeling or misrepresenting the research. In this sense, construct validity can be viewed as a truth in labeling issue.

3-1c Convergent and Discriminant Validity

Convergent and discriminant validity are both considered subcategories or subtypes of construct validity (Campbell & Fiske, 1959). The important thing to recognize is that they work together; if you can demonstrate that you have evidence for both convergent and discriminant validity, you have by definition demonstrated that you have evidence for construct validity. However, neither one alone is sufficient for establishing construct validity.

I find it easiest to think about convergent and discriminant validity as two interlocking propositions. In simple words, I would describe what they are doing as follows:
Measures of constructs that theoretically should be related to each other are, in fact, observed to be related to each other (that is, you should be able to show a correspondence or convergence between similar constructs).

Measures of constructs that theoretically should not be related to each other are, in fact, observed not to be related to each other (that is, you should be able to discriminate between dissimilar constructs).

To estimate the degree to which any two measures are related to each other you would typically use the correlation coefficient discussed in Chapter 11, “Analysis.” That is, you look at the patterns of intercorrelations among the measures. Correlations between theoretically similar measures should be “high,” whereas correlations between theoretically dissimilar measures should be “low.”

The main problem that I have with this convergent-discriminant idea has to do with my use of the quotations around the terms high and low in the previous sentence. The problem is simple: How high do correlations need to be to provide evidence for convergence, and how low do they need to be to provide evidence for discrimination? The answer is that nobody knows! In general, convergent correlations should be as high as possible, and discriminant ones should be as low as possible, but there is no hard and fast rule. Well, let’s not let that stop us. One thing you can assume to be true is that the convergent correlations should always be higher than the discriminant ones. At least that helps a bit.

Before we get too deep into the idea of convergence and discrimination, let’s take a look at each one by using a simple example.

**Convergent Validity** To establish convergent validity, you need to show that measures that should be related are in reality related. In Figure 3-2, you see four measures (each is an item on a scale) that supposedly reflect the construct of self-esteem. For instance, Item 1 might be the statement, “I feel good about myself,” rated using a 1-to-5 scale. You theorize that all four items reflect the idea of self-esteem (which is why I labeled the top part of the figure “Theory”). On the bottom part of the figure (“Observation”), you see the intercorrelations of the four scale items. This might be based on give-
ing your scale out to a sample of respondents. You should readily see that the item intercorrelations for all item pairings are extremely high. (Remember that correlations range from −1.00 to +1.00.) The correlations provide support for your theory that all four items are related to the same construct.

Notice, however, that whereas the high intercorrelations demonstrate that the four items are probably related to the same construct, that doesn’t automatically mean that the construct is self-esteem. Maybe there’s some other construct to which all four items are related (see “Putting It All Together” in Section 3.1c).

However, at least, you can assume from the pattern of correlations that the four items are converging on the same thing, whatever it might be called.

**Discriminant Validity** To establish discriminant validity, you need to show that measures that should not be related are in reality not related. In Figure 3-3, you again see four measures (each is an item on a scale). Here, however, two of the items are thought to reflect the construct of self-esteem, whereas the other two are thought to reflect a different construct called locus of control. The top part of the figure shows the theoretically expected relationships among the four items. If you have discriminant validity, the relationship between measures from different constructs should be low. (Again, nobody knows how low they should be, but I’ll deal with that later.) There are four correlations between measures that reflect different constructs, and these are shown on the bottom of the figure (“Observation”). You should see immediately that these four cross-construct correlations are low (near zero) and certainly much lower than the convergent correlations in Figure 3-2.

Just because there is evidence that the two sets of two measures seem to be related to different constructs (because their intercorrelations are so low) doesn’t mean that the constructs they’re related to are self-esteem and locus of control. However, the correlations do provide evidence that the two sets of measures are discriminated from each other.

**Putting It All Together** Okay, so where does this leave us? I’ve shown how to provide evidence for convergent and discriminant validity separately, but as I said at the outset, to argue for construct validity, you really need to be able to show that both of these
Convergent and Discriminant Validity
Correlations in a Single Table or Correlation Matrix

The correlations support both convergence (black numbers) and discrimination (blue numbers), and therefore, construct validity.

Types of validity are supported. Given the discussions of convergent and discriminant validity, you should be able to see that you could put both principles together into a single analysis to examine both at the same time. This is illustrated in Figure 3-4.

Figure 3-4 shows six measures: three that are theoretically related to the construct of self-esteem and three that are thought to be related to locus of control. The top part of the figure shows this theoretical arrangement. The bottom of the figure shows what a correlation matrix based on a pilot sample might show. To understand this table, first you need to be able to identify the convergent correlations and the discriminant ones. The two sets or blocks of convergent coefficients appear in regular type: one 3 × 3 block for the self-esteem intercorrelations in the upper left of the table, and one 3 × 3 block for the locus-of-control correlations in the lower right. Additionally, two 3 × 3 blocks of discriminant coefficients appear in italics, although if you’re really sharp you’ll recognize that they are the same values in mirror image. (Do you know why? You might want to read up on correlations in Chapter 11, “Analysis.”)

How do you make sense of the correlations' patterns? There are no firm rules for how high or low the correlations need to be to provide evidence for either type of validity, but that the convergent correlations should always be higher than the discriminant ones. Take a good look at the table and you will see that in this example the convergent correlations are always higher than the discriminant ones. I would conclude from this that the correlation matrix provides evidence for both convergent and discriminant validity, all in one table!

It's true the pattern supports discriminant and convergent validity, but does it show that the three self-esteem measures actually measure self-esteem or that the three locus-of-control measures actually measure locus of control? Of course not. That would be much too easy.
So, what good is this analysis? It does show that, as you predicted, the three self-esteem measures seem to reflect the same construct (whatever that might be). The three locus-of-control measures also seem to reflect the same construct (again, whatever that is), and the two sets of measures seem to reflect two different constructs (whatever they are). That's not bad for one simple analysis.

Okay, so how do you get to the really interesting question? How do you show that your measures are actually measuring self-esteem or locus of control? I hate to disappoint you, but there is no simple answer to that. (I bet you knew that was coming.) You can do several things to address this question. First, you can use other ways to address construct validity to help provide further evidence that you're measuring what you say you're measuring. For instance, you might use a face validity or content validity approach to demonstrate that the measures reflect the constructs you say they are. (See Section 3-1a, “Measurement Validity Types,” earlier in this chapter for more information.)

One of the most powerful approaches is to include even more constructs and measures. The more complex your theoretical model (if you find confirmation of the correct pattern in the correlations), the more evidence you are providing that you know what you're talking about (theoretically speaking). Of course, it's also harder to get all the correlations to give you the exact right pattern as you add more measures. In many studies, you simply don't have the luxury of adding more and more measures because it's too costly or demanding. Despite the impracticality, if you can afford to do it, adding more constructs and measures enhances your ability to assess construct validity.

3-1d Threats to Construct Validity

Before I launch into a discussion of the most common threats to construct validity, take a moment to recall what a threat to validity is. In a research study, you are likely to reach a conclusion that your program was a good operationalization of what you wanted and that your measures reflected what you wanted them to reflect. Would you be correct? How will you be criticized if you make these types of claims? How might you strengthen your claims? The kinds of questions and issues your critics will raise are what I mean by threats to construct validity.

The authoritative list of threats to construct validity (Cook & Campbell, 1979) that I follow tends to be a bit academic in its language. Although I love their discussion, I do find some of their terminology a bit cumbersome. Much of what I'll do here is try to translate their terms into words that are more understandable to us normal human beings. Here, then, are the major threats to construct validity (drum roll, please):

- Threat to construct validity
  Any factor that causes you to make an incorrect conclusion about whether your operationalized variables (for example, your program or outcome) reflect well the construct they are intended to represent.

- Mono-operation bias
  A threat to construct validity that occurs when you rely on only a single implementation of your independent variable, cause, program, or treatment in your study.

Okay, this is a fairly daunting phrase, but don’t panic. Breathe in. Breathe out. Let’s decipher this phrase one part at a time. It isn’t nearly as complicated as it sounds. You know what inadequate means, don’t you? (If you don’t, I’d say you’re pretty inadequate!) Here, preoperational is what you were thinking about before you developed your measures or treatments. Explanation is just a fancy word for explanation. Put it all together and what this phrase means is that you didn’t do a good enough job defining what you meant by the construct before you tried to translate it into a measure or program. In other words, you weren’t thinking carefully. How is this a threat? Imagine that your program consisted of a new type of approach to rehabilitation. A critic comes along and claims that, in fact, your program is neither new nor a true rehabilitation program. You are being accused of doing a poor job of thinking through your constructs. Here are some possible solutions:

- Think through your concepts better.
- Use structured methods (for example, concept mapping) to articulate your concepts.
- Get experts to critique your operationalizations.

Mono-operation bias occurs when you use only one version of the treatment or program in your study. Note that it is only relevant to the independent variable, cause, program, or treatment in your study. It does not pertain to
measures or outcomes (see the "Mono-Method Bias" in Section 3-1d). If you only use a single version of a program in a single place at a single point in time, you may not be capturing the full breadth of the concept of the program. Every operationalization is a flawed imperfect reflection of the construct on which it is based. If you conclude that your program reflects the construct of the program, your critics are likely to argue that the results of your study reflect only the peculiar version of the program that you implemented, and not the full breadth of the construct you had in mind. Solution: Try to implement multiple versions of your program.

**Mono-Method Bias** Mono-method bias occurs when you only use one measure of a construct. Note that it is only relevant to your measures or observations, not to your programs or causes. Otherwise, it's essentially the same issue as mono-operation bias. With only a single version of a self-esteem measure, you can't provide much evidence that you're really measuring self-esteem. Your critics will suggest that you aren't measuring self-esteem, that you're only measuring part of it, for instance. Solution: Try to implement multiple measures of key constructs and try to demonstrate (perhaps through a pilot or side study) that the measures you use behave as you theoretically expect them to behave.

**Interaction of Different Treatments** You give a new program designed to encourage high-risk teenage girls to go to school and not become pregnant. The results of your study show that the girls in your treatment group have higher school attendance and lower pregnancy rates. You're feeling pretty good about your program until your critics point out that the targeted at-risk treatment group in your study is also likely to be involved simultaneously in several other programs designed to have similar effects. Can you really claim that the program effect is a consequence of your program? The real program that the girls received may actually be the combination of the separate programs in which they participated. What can you do about this threat? One approach is to try to isolate the effects of your program from the effects of any other treatments. You could do this by creating a research design that uses a control group (This is discussed in detail in Chapter 7, "Design."). In this case, you could randomly assign some high-risk girls to receive your program and some to a no-program control group. Even if girls in both groups receive some other treatment or program, the only systematic difference between the groups is your program. If you observe differences between them on outcome measures, the differences must be due to the program. By using a control group that makes your program the only thing that differentiates the two groups, you control for the potential confusion or "confounding" of multiple treatments.

**Interaction of Testing and Treatment** Does testing or measurement itself make the groups more sensitive or receptive to the treatment? If it does, the testing is essentially a part of the treatment; it's inseparable from the effect of the treatment. This is a labeling issue (and, hence, a concern of construct validity) because you want to use the label program to refer to the program alone, but in fact, it also includes the testing. As in the previous threat, one way to control for this is through research design. If you are worried that a pretest makes your program participants more sensitive or receptive to the treatment, randomly assign your program participants into two groups, where one group gets the treatment and the other doesn't. If there are differences on outcomes between these groups, you have evidence that there is an effect of the testing. If not, the testing doesn't matter. In fact, there is a research design known as the Solomon Four-Group Design that was created explicitly to control for this.

**Restricted Generalizability across Constructs** This is what I like to refer to as the unintended consequences threat to construct validity. You do a study and conclude that Treatment X is effective. In fact, Treatment X is effective, but only on the outcome you measured. What you failed to anticipate is that the treatment may have drastic negative consequences or side effects on other outcomes. When you say that Treatment X is effective, you have defined effective in regards to only the outcomes you measured. But,
in fact, significant unintended consequences might affect constructs you did not measure and cannot generalize to. This threat should remind you that you have to be careful about whether your observed effects (Treatment X is effective) would generalize to other potential outcomes. How can you deal with this threat? The critical issue here is to try to anticipate the unintended and measure a broad range of potential relevant outcomes.

Confounding Constructs and Levels of Constructs Imagine a study to test the effect of a new drug treatment for cancer. A fixed dose of drug X is given to a randomly assigned treatment group and a placebo to the other group. No treatment effects are detected. But perhaps the observed result is only true for a certain dosage level. Slight increases or decreases of the dosage of drug X may radically change the results. In this context, it is not fair for you to label the treatment as “drug X” because you only looked at a narrow range of dose of the drug. Like the other construct validity threats, this threat is essentially a labeling issue; your label is not a good description for what you implemented. What can you do about it? If you find a treatment effect at a specific dosage, be sure to conduct subsequent studies that explore the range of effective doses. Note that, although I use the term dose here, you shouldn’t limit the idea to medical studies. If you find an educational program effective at a particular dose—say one hour of tutoring a week—conduct subsequent studies to see if dose responses change as you increase or decrease from there. Similarly, if you don’t find an effect with an initial dose, don’t automatically give up. It may be that at a higher dose the desired outcome will occur.

The Social Threats to Construct Validity The remaining major threats to construct validity can be distinguished from the ones I discussed so far because they are all related to the social and human nature of research.

Hypothesis Guessing Most people don’t just participate passively in a research project. They guess at what the real purpose of the study is. Therefore, they are likely to base their behavior on what they guess, not just on your treatment. In an educational study conducted in a classroom, students might guess that the key dependent variable has to do with class participation levels. If they increase their participation not because of your program but because they think that’s what you’re studying, you cannot label the outcome as an effect of the program. It is this labeling issue that makes this a construct validity threat. This is a difficult threat to eliminate. In some studies, researchers try to hide the real purpose of the study, but this may be unethical, depending on the circumstances. In some instances, they eliminate the need for participants to guess by telling them the real purpose (although who’s to say that participants will believe them). If this is a potentially serious threat, you may think about trying to control for it explicitly through your research design. For instance, you might have multiple program groups and give each one slightly different explanations about the nature of the study, even though they all get exactly the same treatment or program. If they perform differently, it may be evidence that they were guessing differently and that this was influencing the results.

Evaluation apprehension Many people are anxious about being evaluated. Some are even phobic about testing and measurement situations. If their apprehension (and not your program conditions) makes them perform poorly, you certainly can’t label that as a treatment effect. Another form of evaluation apprehension concerns the human tendency to want to look good or look smart, and so on. If, in their desire to look good, participants perform better (and not as a result of your program), you would be wrong to label this as a treatment effect. In both cases, the apprehension becomes confounded with the treatment itself, and you have to be careful about how you label the outcomes. Researchers take a variety of steps to reduce apprehension. In any testing or measurement situation, it is probably a good idea to give participants some time to get comfortable and adjusted to their surroundings. You might ask a few warm-up questions, knowing that you are not going to use the answers and trying to encourage the participant to get comfortable responding. In many research projects, people misunderstand what you are
measuring. If it is appropriate, you may want to tell them that there are no right or wrong answers and that they aren’t being judged or evaluated based on what they say or do.

**Experimenter Expectancies** These days, where we engage in lots of nonlaboratory applied social research, we generally don’t use the term experimenter to describe the person in charge of the research. So, let’s relabel this threat researcher expectancies (Rosenthal, 1966). The researcher can bias the results of a study in countless ways, both consciously or subconsciously. Sometimes, the researcher can communicate what the desired outcome for a study might be (and the participants’ desire to look good leads them to react that way). For instance, the researcher might look pleased when participants give a desired answer. If researcher feedback causes the response, it would be wrong to label the response a treatment effect. As in many of the previous threats, probably the most effective way to address this threat is to control for it through your research design. For instance, if resources allow, you can have multiple experimenters who differ in their characteristics. Or you can address the threat through measurement; you can measure expectations prior to the study and use the information in that analysis to attempt to adjust for expectations.

### 3-2 RELIABILITY

Reliability is part of the quality of measurement. In its everyday sense, reliability is the consistency or repeatability of your measures. Before I can define reliability precisely, I have to lay the groundwork. First, you have to learn about the foundation of reliability, the true score theory of measurement. Along with that, you need to understand the different types of measurement error because errors in measures play a key role in degrading reliability. With this foundation, you can then consider the basic theory of reliability, including a precise definition of reliability. There you will find out that you cannot calculate reliability—you can only estimate it. Because of this, there are a variety of different types of reliability and multiple ways to estimate reliability for each type. In the end, it’s important to integrate the idea of reliability with the other major criteria for the quality of measurement—validity—and develop an understanding of the relationships between reliability and validity in measurement.

#### 3-2a True Score Theory

True score theory is a theory about measurement (Lord & Novick, 1968). Like all theories, you need to recognize that it is not proven; it is postulated as a model of how the world operates. Like many powerful models, true score theory is a simple one. Essentially, true score theory maintains that every measurement is the sum of two components: true ability (or the true level) of the respondent on that measure and random error. This is illustrated in Figure 3-5. You observe the measurement: a score on a test, the total for a self-esteem instrument, or the scale value for a person’s weight. You don’t observe what’s on the right side of the equation. In true score theory, you assume that there are only the two components to the right side of the equal sign in the equation.

**Why is true score theory important?** For one thing, it is a simple yet powerful model for measurement. It is a reminder that most measurement has an error component. Second, true score theory is the foundation of reliability theory. A measure that has no ran-
dom error (is all true score) is perfectly reliable; a measure that has no true score (is nothing but random error) has zero reliability. Third, true score theory can be used in computer simulations as the basis for generating observed scores with certain known properties.

You should know that the true score model is not the only measurement model available. Measurement theorists continue to come up with more and more complex models that they think represent reality even better. However, these models are complicated enough that they lie outside the boundaries of this book. In any event, true score theory should give you an idea of why measurement models are important at all and how they can be used as the basis for defining key research ideas.

3-2b Measurement Error

True score theory is a good simple model for measurement, but it may not always be an accurate reflection of reality. In particular, it assumes that any observation is composed of the true value plus some random error value, but is that reasonable? What if all error is not random? Isn’t it possible that some errors are systematic, that they hold across most or all of the members of a group? One way to deal with this notion is to revise the simple true score model by dividing the error component into two subcomponents: random error and systematic error. Figure 3-6 shows these two components of measurement error, what the difference between them is, and how they affect research.

What Is Random Error? Random error is caused by any factors that randomly affect measurement of the variable across the sample. For instance, people’s moods can inflate or deflate their performance on any occasion. In a particular testing, some children may be in a good mood, and others may be depressed. If mood affects the children’s performance on the measure, it might artificially inflate the observed scores for some children and artificially deflate them for others. The important thing about random error is that it does not have any consistent effects across the entire sample. Instead, it pushes observed scores up or down randomly. This means that if you could see all the random errors in a distribution, they would have to sum to 0. There would be as many negative errors as positive ones. (Of course you can’t see the random errors because all you see is the observed score X.) The important property of random error is that it adds variability to the data but does not affect average performance for the group (see Figure 3-7). Because of this, random error is sometimes considered noise.

What Is Systematic Error? Systematic error is caused by any factors that systematically affect measurement of the variable across the sample. For instance, if there is
Systematic Error

Systematic error affects the central tendency of a distribution.

Notice that systematic error does affect the average—we call this a bias.

loud traffic going by just outside of a classroom where students are taking a test, this noise is likely to affect all of the children's scores—in this case, systematically lowering them. Unlike random error, systematic errors tend to be either positive or negative consistently; because of this, systematic error is sometimes considered to be bias in measurement (see Figure 3-8).

Reducing Measurement Error So, how can you reduce measurement errors, random or systematic? One thing you can do is to pilot-test your instruments to get feedback from your respondents regarding how easy or hard the measure was and how the testing environment affected their performance. Second, if you are gathering measures using people to collect the data (as interviewers or observers), you should make sure you train them thoroughly so that they aren't inadvertently introducing error. Third, when you collect the data for your study, you should double-check the data thoroughly. All data entry for computer analysis should be double-entered and verified. Ideally, you enter the data twice, the second time having the computer check that you are typing the exact same data you typed the first time. Fourth, you can use statistical procedures to adjust for measurement error. These range from rather simple formulas you can apply directly to your data to complex procedures for modeling the error and its effects. Finally, one of the best things you can do to deal with measurement errors, especially systematic errors, is to use multiple measures of the same construct. Especially if the different measures don't share the same systematic errors, you will be able to triangulate across the multiple measures and get a more accurate sense of what's happening.

3-2c Theory of Reliability

What is reliability? We hear the term used a lot in research contexts, but what does it really mean? If you think about how we use the word reliable in everyday language, you might get a hint. For instance, we often speak about a machine as reliable: "I have a reliable car." Or newspaper talk about a "usually reliable source." In both cases, the word reliable usually means dependable or trustworthy. In research, the term reliable also means dependable in a general sense, but that's not a precise enough definition. What does it mean to have a dependable measure or observation in a research context? The reason dependable is not a good enough description is that it can be confused too easily with the idea of a valid measure (see Section 3-1, "Construct Validity," earlier in this chapter). Certainly, when researchers speak of a dependable measure, we mean one that is both reliable and valid. So, we have to be a little more precise when we try to define reliability.
In research, the term *reliability* means repeatability or consistency. A measure is considered reliable if it would give you the same result over and over again (assuming that what you are measuring isn’t changing).

Let’s explore in more detail what it means to say that a measure is repeatable or consistent. I’ll begin by defining a measure that I’ll arbitrarily label $X$. It might be a person’s score on a math achievement test or a measure of severity of illness. It is the value (numerical or otherwise) that you observe in your study. Now, to see how repeatable or consistent an observation is, you can measure it twice. You use subscripts to indicate the first and second observation of the same measure, as shown in Figure 3-9. If you assume that what you’re measuring doesn’t change between the time of the first and second observation, you can begin to understand how you get at reliability. Although you observe a single score for what you’re measuring, you usually think of that score as consisting of two parts: the true score or actual level for the person on that measure, and the error in measuring it (Lord & Novick, 1968) (see Section 3-2a, “True Score Theory,” earlier in this chapter).

It’s important to keep in mind that you observe the $X$ score; you never actually see the true ($T$) or error ($e$) scores. For instance, a student may get a score of 85 on a math achievement test. That’s the score you observe, an $X$ of 85. However, the reality might be that the student is actually better at math than that score indicates. Let’s say the student’s true math ability is 89 ($T = 89$). That means that the error for that student is $-4$. What does this mean? Well, while the student’s true math ability may be 89, he or she may have had a bad day, may not have had breakfast, may have had an argument with someone, or may have been distracted while taking the test. Factors like these can contribute to errors in measurement that make students’ observed abilities appear lower than their true or actual abilities.

Okay, back to reliability. Let’s make a simple assumption to make the discussion more straightforward. Assume that you are repeatedly measuring some characteristic that is not changing over the period you are measuring it. For example, perhaps you are measuring heights of adults over a several day period, or intelligence in the same children within the same week. It’s not likely that these characteristics would really change considerably in that time. If your measure is reliable, you should pretty much get the same result each time you measure it. Why would the results be essentially the same in these circumstances? If you look at Figure 3-9, you should see that the only thing that the two observations have in common is their true scores, $T$. How do you know that? Because the error scores ($e_1$ and $e_2$) have different subscripts, indicating that they are different values. (You are likely to have different errors on different occasions.) However, the true score symbol ($T$) is the same for both observations. What does this mean? The two observed scores, $X_1$ and $X_2$, are related only to the degree that the observations share a true score. You should remember that the error score is assumed to be random (see Section 3-2a, “True Score Theory,” earlier in this chapter). Sometimes, errors will lead you to perform better on a test than your true ability (you had a good day guessing!), while other times, they will lead you to score worse. The true score—your true ability on that measure—would be the
The Reliability Ratio
Expressed in Terms of Variances

\[ \frac{\text{var}(T)}{\text{var}(X)} \]

same on both observations (assuming, of course, that your true ability didn't change between the two measurement occasions).

With this in mind, I can now define reliability more precisely. Reliability is a ratio or fraction. In layperson terms, you might define this ratio as shown in Figure 3-10.

You might think of reliability as the proportion of truth in your measure. Now, it makes no sense to speak of the reliability of a measure for an individual; reliability is a characteristic of a measure that's taken across individuals. So, to get closer to a more formal definition, I'll restate the definition of reliability in terms of a set of observations. The easiest way to do this is to speak of the variance of the scores. The variance is a measure of the spread or distribution of a set of scores (see Chapter 11, "Analysis"). So, I can now state the definition as shown in Figure 3-11.

I might put this into slightly more technical terms by using the abbreviated name for the variance and our variable names (see Figure 3-12).

We're getting to the critical part now. If you look at the equation in Figure 3-12, you should recognize that you can easily determine or calculate the bottom part of the reliability ratio; it's just the variance of the set of observed scores. (You remember how to calculate the variance, don't you? It's the sum of the squared deviations of the scores from their mean, divided by the number of scores. If you're still not sure, see Chapter 11, "Analysis").) So, how do you calculate the variance of the true scores? You can't see the true scores. (You only see X!) Therefore, if you can't calculate the variance of the true scores, you can't compute the ratio, which means you can't compute reliability! Everybody got that? Here's the bottom line:

You can't compute reliability because you can't calculate the variance of the true scores!

Great. So, where does that leave you? If you can't compute reliability, perhaps the best you can do is to estimate it. Maybe you can get an estimate of the variability of the true scores. How do you do that? Remember your two observations, X₁ and X₂? You assume (using true score theory) that these two observations would be related to each other to the degree that they share true scores. So, let's calculate the correlation between X₁ and X₂. Figure 3-13 shows a simple formula for the correlation.

\[ \frac{\text{covariance}(X₁, X₂)}{\text{sd}(X₁) \times \text{sd}(X₂)} \]
In Figure 3-13, the sd stands for the standard deviation (which is the square root of the variance). If you look carefully at this equation, you can see that the covariance, which simply measures the shared variance between measures, must be an indicator of the variability of the true scores because the true scores in \( X_1 \) and \( X_2 \) are the only things the two observations share! So, the top part is essentially an estimate of \( \text{var}(T) \) in this context. Additionally, since the bottom part of the equation multiplies the standard deviation of one observation with the standard deviation of the same measure at another time, you would expect that these two values would be the same (it is the same measure we're taking) and that this is essentially the same thing as squaring the standard deviation for either observation. However, the square of the standard deviation is the same thing as the variance of the measure. So, the bottom part of the equation becomes the variance of the measure (or \( \text{var}(X) \)). If you read this paragraph carefully, you should see that the correlation between two observations of the same measure is an estimate of reliability. Got that? I've just shown that a simple and straightforward way to estimate the reliability of a measure is to compute the correlation of two administrations of the measure!

It's time to reach some conclusions. You know from this discussion that you cannot calculate reliability because you cannot measure the true score component of an observation. You also know that you can estimate the true score component as the covariance between two observations of the same measure. With that in mind, you can estimate the reliability as the correlation between two observations of the same measure. It turns out that there are several ways to estimate this reliability correlation. These are discussed in Section 3-2d, "Types of Reliability," later in this chapter.

There's only one other issue I want to address here. What is the range of a reliability estimate? What is the largest or smallest value a reliability can be? To figure this out, let's go back to the equation given earlier (see Figure 3-14).

Remember, because \( X = T + e \), you can substitute it in the bottom of the ratio as shown in Figure 3-15.

With this slight change, you can easily determine the range of a reliability estimate. If a measure is perfectly reliable, there is no error in measurement; everything you observe is true score. Therefore, for a perfectly reliable measure, \( \text{var}(e) \) is zero and the equation would reduce to the equation shown in Figure 3-16.

Therefore, reliability = 1. Now, if you have a perfectly unreliable measure, there is no true score; the measure is entirely error. In this case, the equation would reduce to the equation shown in Figure 3-17.
even be better if you randomly assign individuals to receive Form A or B on the pretest and then switch them on the posttest. With split-half reliability, you have an instrument to use as a single-measurement instrument and only develop randomly split halves for purposes of estimating reliability.

**Internal Consistency Reliability**  In internal consistency reliability estimation, you use a single measurement instrument administered to a group of people on one occasion to estimate reliability. In effect, you judge the reliability of the instrument by estimating how well the items on the instrument that reflect the same construct yield similar results. You are looking at how consistent the results are for different items for the same construct within the measure. There are a wide variety of internal-consistency measures you can use.

**Average Inter-Item Correlation**  The average inter-item correlation uses all of the items on your instrument that are designed to measure the same construct. You first compute the correlation between each pair of items, as illustrated in Figure 3-21. For example, if you have six items, you will have 15 different item pairings (15 correlations). The average inter-item correlation is simply the average of all of these correlations. In the example, you find an average inter-item correlation of .90 with the individual correlations ranging from .84 to .95.

**Average Item-Total Correlation**  This approach also uses the inter-item correlations. In addition, you compute a total score for the six items and treat that in the analysis like an additional variable. Figure 3-22 shows the six item-to-total correlations at the bottom of the correlation matrix. They range from .82 to .88 in this sample analysis, with the average of these at .85.

**Split-Half Reliability**  In split-half reliability, you randomly divide into two sets all items that measure the same construct. You administer the entire instrument to a sample and calculate the total score for each randomly divided half of the measure. The split-half reliability estimate, as shown Figure 3-23, is simply the correlation between these two total scores. In the example, it is .87.

**Cronbach's Alpha (α)**  Imagine that you compute one split-half reliability and then randomly divide the items into another set of split halves and recompute, and keep doing this until you have computed all possible split-half estimates of reliability.

---

**Figure 3-21**

The Average Inter-Item Correlation

```
<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.89</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>.91</td>
<td>.92</td>
<td>1.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>.88</td>
<td>.93</td>
<td>.95</td>
<td>1.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>.84</td>
<td>.86</td>
<td>.92</td>
<td>.85</td>
<td>1.00</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.88</td>
<td>.91</td>
<td>.95</td>
<td>.87</td>
<td>.85</td>
<td>1.00</td>
</tr>
</tbody>
</table>
```

\[
.89 + .91 + .88 + .84 + .88 + .92 + .93 + .86 + .91 + .95 + .92 + .95 + .85 + .87 + .85 = 13.41 \\
13.41 / 15 = .90
\]
Cronbach's alpha (Miller, 1995) is mathematically equivalent to the average of all possible split-half estimates. Thankfully, that's not how we typically compute it. Some clever mathematician (Cronbach, no doubt!) figured out a way to get the mathematical equivalent a lot more quickly. One way is with the simple formula shown in Figure 3-24. This formula only requires that you have the average inter-item correlation and the number of items. The example shown in Figure 3-24 shows that if you had six items and an average inter-item correlation of .85, Cronbach's alpha would equal .97.

**Comparison of Reliability Estimators** Each of the reliability estimators has certain advantages and disadvantages. Inter-rater reliability is one of the best ways to estimate reliability when your measure is an observation. However, it requires multiple raters or observers. As an alternative, you could look at the correlation of ratings of the same single observer repeated on two different occasions. For example, let's say you collected videotapes of child-mother interactions and had a rater code the videos for how often the mother smiled at the child. To establish inter-rater reliability, you could take a sample of videos and have two raters code them independently. You might use the inter-rater approach especially if you were interested in using a team of raters and you wanted to establish that they yielded consistent results. If you get a suitably high inter-rater reliability, you could then justify allowing them to work independently on coding different videos. (When you only have a single rater and cannot easily train others, you might alternatively use a test-retest approach by having this person rate the same sample of
Alpha = \frac{K \cdot \bar{r}}{1 + (K-1) \cdot \bar{r}}

where:
K = Number of items on measure
\bar{r} = Average inter-item correlation

Example:
If K = 6 and \bar{r} = .85:
\begin{align*}
\text{Alpha} &= \frac{6 \cdot .85}{1 + (6 - 1) \cdot .85} \\
&= \frac{5.1}{1 + 4.25} \\
&= .97
\end{align*}

3-2e Reliability and Validity

We often think of reliability and validity as separate ideas, but in fact, they're related to each other. One of my favorite metaphors for the relationship between reliability and validity is that of a target. Think of the center of the target as the concept you are trying to measure. Imagine that for each person you are measuring, you are taking a shot at the target. If you measure the concept perfectly for a person, you are hitting the center of the target. If you don't, you are missing the center. The more you are off for that person, the further you are from the center (see Figure 3-25).

Figure 3-25 shows four possible situations. In the first one, you are hitting the target consistently, but you are missing the center of the target. That is, you are consistently and systematically measuring the wrong value for all respondents. This measure is reliable but not valid. (It's consistent but wrong.) The second shows hits that are randomly spread across the target. You seldom hit the center of the target, but on average, you are getting the right answer for the group (but not very well for individuals). In this case, you get a valid group average, but you are inconsistent. Here, you can clearly see that reliability is directly related to the variability of your measure. The third scenario shows a case where
your hits are spread across the target and you are consistently missing the center. Your measure in this case is neither reliable nor valid. Finally, the figure shows the Robin Hood scenario; you consistently hit the center of the target. Your measure is both reliable and valid. (I bet you never thought of Robin Hood in those terms before.)

3-3 LEVELS OF MEASUREMENT

The level of measurement refers to the relationship among the values that are assigned to the attributes for a variable. What does that mean? Begin with the idea of the variable, for example, party affiliation (see Figure 3-26). That variable has a number of attributes. Let's assume that in this particular election context, the only relevant attributes are Republican, Democrat, and Independent. For purposes of analyzing the results of this variable, we arbitrarily assign the values 1, 2, and 3 to the three attributes. The level of measurement describes the relationship among these three values. In this case, the numbers function as shorter placeholders for the lengthier text terms. Higher values don't mean more of something, and lower numbers don't signify less. The value of 2 doesn't mean that Democrats are twice something than Republicans. Assigning a 1 for Republicans doesn't mean they are in first place or have the highest priority. In this case, the level of measurement can be described as nominal.

3-3a Why Is Level of Measurement Important?

First, knowing the level of measurement helps you decide how to interpret the data from that variable. When you know that a measure is nominal (like the one just described), you know that the numerical values are short codes for the longer names. Second, knowing the level of measurement helps you decide what statistical analysis is appropriate on the values that were assigned. If you know that a measure is nominal, then you would automatically know that you don't average the data values (except in certain circumstances like the use of "dummy" variables, described in Chapter 12, "Analysis for Research Design"). Why? Because it makes no sense to add "names" and then divide them by the number of names, which is how an average is calculated (see Chapter 12). And it also means that all statistical analyses that depend on the average or use it as part of their calculation (for example, the t-test as described in Chapter 12) would also not be appropriate.

There are four levels of measurement that are most commonly discussed (Stevens, 1946) (see Figure 3-27):

- Nominal—In nominal measurement, the numerical values simply name the attribute uniquely. No ordering of the cases is implied. For example, jersey numbers in basketball are measures at the nominal level. A player with number 30 is not more of anything than a player with number 15 and is certainly not twice whatever number 15 is.

![Figure 3-26](image)

**FIGURE 3-26**

Level of Measurement

The level of measurement describes the relationship among the values associated with the attributes of a variable.