CHAPTER 3

The Theory of Measurement

KEY TERMS

- concurrent validity
- construct validity
- content validity
- convergent validity
- criterion-related validity
- Cronbach's alpha
- discriminant validity
- external validity
- face validity
- hypothesis
- mono-method bias
- mono-operation bias
- operationalization
- pattern matching
- predictive validity
- reliability
- standard deviation
- translation validity
- true score theory
- validity

OUTLINE

3-1 Construct Validity, 56
3-1a Measurement Validity Types, 58
3-1b Idea of Construct Validity, 61
3-1c Convergent and Discriminant Validity, 63
3-1d The Nomological Network, 67
3-1e The Multitrait-Multimethod Matrix, 68
3-1f Pattern Matching for Construct Validity, 73
3-1g Structural Equation Modeling, 76
3-1h Threats to Construct Validity, 78
3-2 Reliability, 80
3-2a True Score Theory, 80
3-2b Measurement Error, 81
3-2c Theory of Reliability, 83
3-2d Types of Reliability, 87
3-2e Reliability and Validity, 93
3-3 Levels of Measurement, 95
3-3a Why Is Level of Measurement Important? 95
Summary, 97

Measurement is the process of observing and recording the observations that are collected as part of a research effort. There are two major issues that will be considered here.

First, you have to understand the fundamental ideas or theory involved in measuring. 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 consideration of true score theory and a variety of reliability estimators. In the section on levels of measurement, I explain the meaning of the four major levels of measurement: **nominal**, **ordinal**, **interval**, and **ratio**.

3-1 Construct Validity

In the first chapter of this book, you were introduced to an idea about ideas. When researchers think about what to study, they go through a process of defining the concepts they are interested in and the relationships that might exist between various concepts. The method of concept mapping was introduced as one way to go from an abstract notion about something to a more specific conceptualization of the idea. Once the concept has been defined and differentiated from other concepts, it can be formally studied as a construct. The steps involved in moving toward a concrete representation of the construct and the issues involved in determining how well that process has been conducted are the subject of this chapter. The most important characteristic that a measure of a construct can have is **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. Whoa! Can you believe that the term *operationalization* has eight syllables? That's a mouthful. What does it mean here? An operationalization is your translation of an 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. The construct validity issue is the degree to which the actual (operationalized) program reflects the ideal (the program as you conceptualized or envisioned it). Imagine that you want to measure the construct of self-esteem. You have an idea of what self-esteem means. You construct a ten-item paper-and-pencil instrument to measure self-esteem. The instrument is the operationalization; it's the translation of the idea of self-esteem into something concrete, into specific operations. The construct validity question here would be how well the ten-item instrument (the operationalization) reflects the idea you had of self-esteem. Well, I'll cover this in more detail later, but I didn't want to start the chapter with an eight-syllable word that will confuse you at the outset.

validity
The best available approximation of the truth of a given proposition, inference, or conclusion.

construct validity
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.

operationalization
The act of translating a construct into its manifestation—for example, translating the idea of your treatment or program into the actual program, or translating the idea of what you want to measure into the real measure. The result is also referred to as an *operationalization*; that is, you might describe your actual program as an *operationalized program*.

Like **external validity** (see the discussion in Chapter 2) construct validity is related to generalizing. However, whereas 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 term *self-esteem* is that what you were really measuring?

external validity
The degree to which the conclusions in your study would hold for other persons in other places and at other times.

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 foundation in the richness of this idea. 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. I don't see construct validity as limited only to measurement. As I've already implied, I think it is as much a part of the independent variable—the program or treatment—as it is 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.

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

predictive validity
A type of construct validity based on the idea that your measure is able to predict what it theoretically should be able to predict.

concurrent validity
An operationalization's ability to distinguish between groups that it should theoretically be able to distinguish between.

This view of validity has much in common with the perspective developed by Samuel Messick, who had a very influential career in test development and validation at the Educational Testing Service. Messick (1995) thought of validity as a unified idea with many facets, including the theories that dictate what the structure of constructs should look like and the social consequences of test scores. We will return to consideration of facets of construct validity, but first let's look at some history to see how these ideas evolved in the real world.

During World War II, the U.S. government involved hundreds (and perhaps thousands) of psychologists and psychology graduate students in the development of an array of measures that were relevant to the war effort. They needed personality screening tests for prospective fighter pilots, personnel measures that would enable sensible assignment of people to job skills, psychophysical measures to test reaction times, and so on. After the war, these psychologists needed to find gainful employment outside of the military, and it's not surprising that many of them moved into testing and measurement in a civilian context. During the early 1950s, the American Psychological Association became increasingly concerned with the quality or validity of all of the new measures that were being generated and decided to convene an effort to set standards for psychological measures. The first formal articulation of the idea of construct validity came from this effort and was couched under the somewhat grandiose term of the nomological network (see Section 3-1e, The Nomological Network). The nomological network provided a theoretical basis for the idea of construct validity, but it didn't provide practicing researchers with a way to actually establish whether their measures had construct validity. In 1959, an attempt was made to develop a concrete, practical method for assessing construct validity using what is called a *multitrait-multimethod matrix*, or MTMM for short (see Section 3-1f, The Multitrait-Multimethod Matrix). To argue that your measures had construct validity under the MTMM approach, you had to demonstrate that there was *both convergent* and discriminant validity in your measures. You demonstrated construct validity when you showed that measures that are theoretically supposed to be highly interrelated are, in practice, highly interrelated. You showed discriminant validity when you demonstrated that measures that shouldn't be related to each other in fact were not.

While the MTMM did provide a methodology for assessing construct validity, it was a difficult one to implement well, especially in applied social research contexts and, in fact, has seldom been formally attempted. When the thinking about construct validity that underlies both the nomological network and the MTMM is examined carefully, one of the key themes that can be identified is in the idea of pattern. When you claim that your programs or measures have construct validity, you are essentially claiming that you, as a researcher, understand how your constructs or theories of the programs and measures operate in theory, and you are claiming that you can provide evidence that they behave in practice the way you think they should, that they follow the expected pattern.

The researcher essentially has a theory about how the programs and measures relate to each other (and other theoretical terms), a *theoretical pattern* if you will. The researcher provides evidence through observation that the programs or measures actually behave that way in reality, an *observed pattern.* When you claim construct validity, you're essentially claiming that your observed pattern—how things operate in reality—corresponds with your theoretical pattern—how you think the world works. I call this process **pattern matching**, and I believe that it is the heart of construct validity. It is clearly an underlying theme in both the nomological network and the MTMM ideas. In addition, I think that, as researchers, we can develop concrete and feasible methods that enable practicing researchers to assess pattern matches to assess the construct validity of their research. Section 3-1g, Pattern Matching for Construct Validity, lays out my idea of how you might use this approach to assess construct validity.

pattern matching
The degree of correspondence between two patterns. For instance, you might look at a pattern match of a theoretical expectation pattern with an observed pattern to see if you are getting the outcomes you expect or if your measures intercorrelate the way you would theoretically predict they would.

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. 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 typical measurement-related validity terms to reflect different ways you can demonstrate different aspects of construct validity.

With all that in mind, following is 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've 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.

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

content validity
A check of the operationalization against the relevant content domain for the construct.

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: translation validity and criterion-related validity. That's because I think these correspond to the two major ways you can ensure and assess the validity of an operationalization.

In translation validity, you focus on whether the operationalization is a good reflection 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 given your theory 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 on 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.

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.

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, and hence the choice of name. 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. 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 is all you do 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 are 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. This approach assumes that you have a good detailed description of the content domain, something that's not always true. For instance, you might lay out all of the criteria that should be met in a program that claims to be a teenage pregnancy-prevention program. You would probably include in this domain specification the definition of the target group, criteria for deciding whether the program is preventive in nature (as opposed to treatment-oriented), and criteria that spell out 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 your criteria, you create a type of checklist when examining your program. Only programs that meet the checklist criteria can legitimately be defined as teenage pregnancy-prevention programs. This all sounds fairly straightforward, and for many operationalizations it will be. However, for other constructs (such as self-esteem or intelligence), it will not be easy to decide which criteria 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 in relation to some other measure* based on your theory of the construct. The differences among the criterion-related validity types is in the criteria they use as the standard for judgment.

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 construe 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 evidence. 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 manic-depression, your measure should be able to distinguish between people who are diagnosed manic-depressive 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.

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.

convergent validity
The degree to which the operationalization is similar to (converges on) other operationalizations to which it should be theoretically similar.

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 be 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.

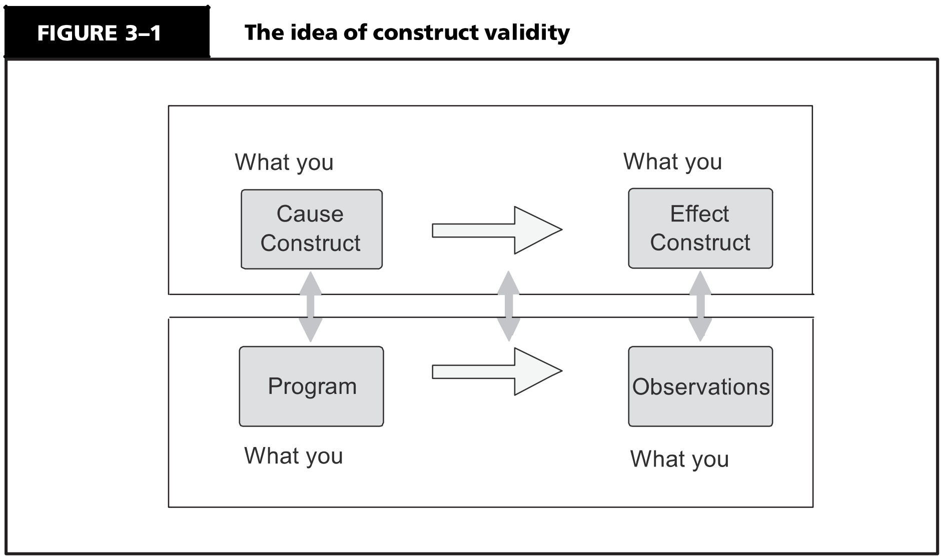
3-1b Idea of Construct Validity

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 something, 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. It includes the idea or construct of the outcomes or measures you believe you are trying to affect. The land of observation consists of what you see 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. Presumably, you have constructed the land of observation based on your theories. 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.

hypothesis
A model in which two mutually exclusive hypotheses that together exhaust all possible outcomes are tested, such that if one hypothesis is accepted, the second must therefore be rejected.

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

FIGURE 3–1 **The idea of construct validity**



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.

The truth in labeling aspect of validity is reminiscent of what Messick (1995) was writing about when describing the *consequential* facet of validity. In addition to focusing on the relative success of a translation of a construct to a measure as many other validity theorists have done, Messick asked us to consider what happens to individuals or groups as a result of a testing process in terms of the effect on people who took the test. Therefore, his model of validity added a kind of ethical dimension because it takes into account the fact that sometimes the consequences of a testing process can be positive, as when a person succeeds or a program gets better, or negative, which is particularly troublesome when invalidity in a test creates systematic bias in scoring or unfairness in application of results.

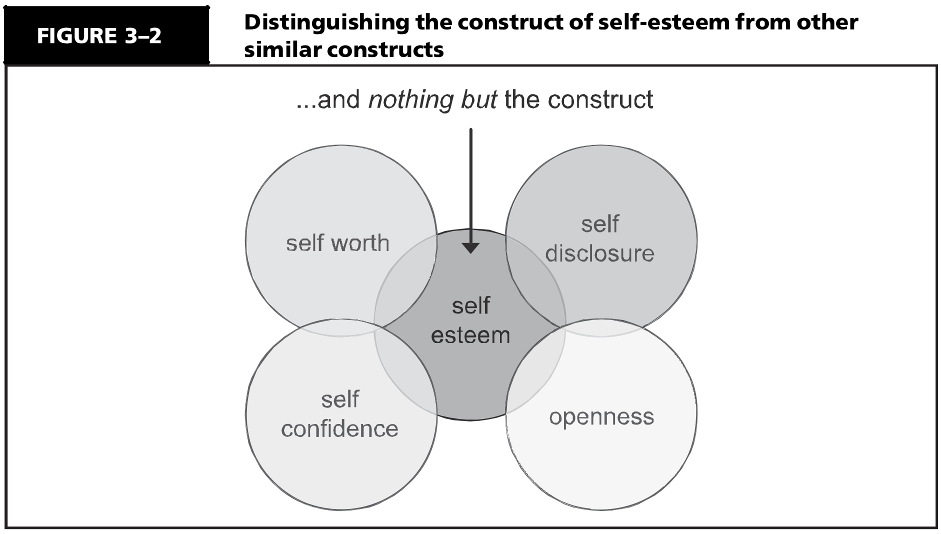
Definitionalist versus Relationalist Views There really are two broad ways of looking at the idea of construct validity. I'll call the first the *definitionalist* perspective because it essentially holds that the way to ensure construct validity is to define the construct so precisely that you can operationalize it in a straightforward manner. In a definitionalist view, either you have operationalized the construct correctly or you haven't; it's either/or type of thinking. Either this program is a "Type A Tutoring Program" or it isn't. Either you're measuring self-esteem or you aren't.

The other perspective I'd call *relationalist.* To a relationalist, things are not either/or or black and white; concepts are more or less related to each other. The meaning of terms or constructs differs relatively, not absolutely. The program in your study might be a "Type A Tutoring Program" in some ways, while in others it is not. It might be more that type of program than another program. Your measure might be capturing some of the construct of self-esteem, but it may not capture all of it. There may be another measure that is closer to the construct of self-esteem than yours is. Relationalism suggests that meaning changes gradually. It rejects the idea that you can rely on operational definitions as the basis for construct definition.

To get a clearer idea of this distinction, you might think about how the law approaches the construct of truth. Most of you have heard the standard oath that witnesses in a U.S. court are expected to swear. They are to tell "the truth, the whole truth and nothing but the truth." What does this mean? If witnesses had to swear only to tell the truth, they might choose to interpret that to mean that they should make sure what they say is true. However, that wouldn't guarantee that they would tell *everything* they knew to be true. They might leave out some important things and still tell the truth. They just wouldn't be telling everything. On the other hand, they are asked to tell "nothing but the truth." This suggests that you can say simply that Statement X is true and Statement Y is not true.

Now, let's see how this oath translates into a measurement and construct validity context. For instance, you might want your measure to reflect the construct, the whole construct, and nothing but the construct. What does this mean? Let's assume, as shown in Figure 3–2, that you have five distinct constructs that are all conceptually related to each other: self-esteem, self-worth, self-disclosure,

FIGURE 3–2 **Distinguishing the construct of self-esteem from other similar constructs**



self-confidence, and openness. Most people would say that these concepts are similar, although they can be distinguished from each other. If you were trying to develop a measure of self-esteem, what would it mean to measure self-esteem, all of self-esteem, and nothing but self-esteem? If the concept of self-esteem overlaps with the others, how could you possibly measure all of it (that would presumably include the part that overlaps with others) *and* nothing but it? You couldn't! If you believe that meaning is relational in nature—that some concepts are closer in meaning than others—the legal model discussed here does not work well as a model for construct validity.

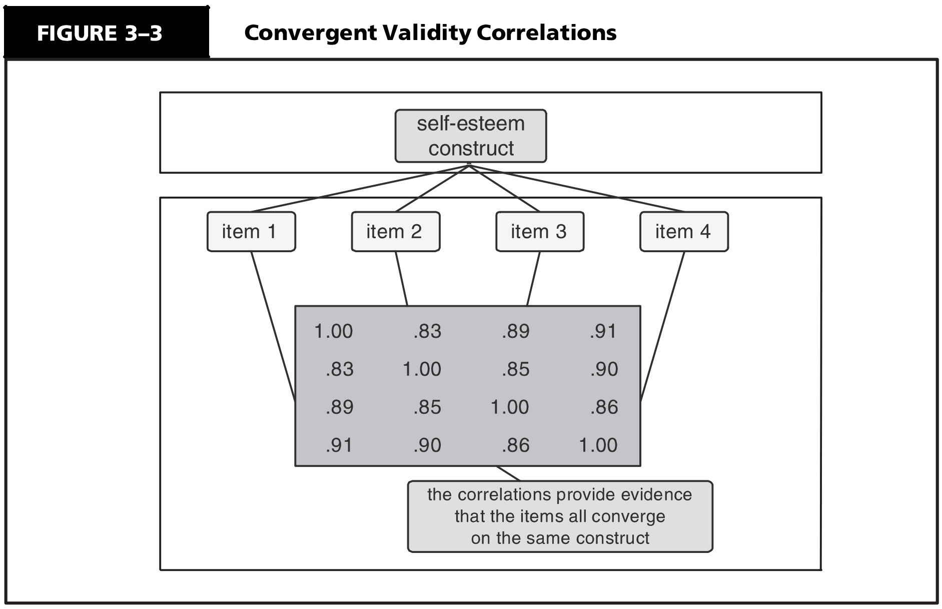
In fact, you will see that most social research methodologists have (whether they've thought about it or not) rejected the definitionalist perspective in favor of a relationalist one. To establish construct validity from a relationalist perspective you have to meet the following conditions:

- You have to set the construct you want to operationalize for example, self-esteem) within a *semantic net* (or net of meaning). This means that you have to tell what your construct is more or less similar to in meaning.
- You need to be able to provide direct evidence that you *control* the operationalization of the construct and that your operationalizations look like what they should theoretically look like. If you are trying to measure self-esteem, you have to be able to explain why you operationalized the questions the way you did. If all your questions are addition problems, how can you argue that your measure reflects self-esteem and not adding ability?
- You have to provide evidence that your data supports your theoretical view of the relations among constructs. If you believe that self-esteem is closer in meaning to self-worth than it is to anxiety, you should be able to show that measures of self-esteem are more highly correlated with measures of self-worth than with ones of anxiety.

3-1c Convergent and Discriminant Validity

Convergent and discriminant validity are both considered subcategories or subtypes of construct validity. 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.

FIGURE 3–3 **Convergent Validity Correlations**



1.00	.83	.89	.91
.83	1.00	.85	.90
.89	.85	1.00	.86
.91	.90	.86	1.00

the correlations provide evidence that the items all converge on the same construct

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).

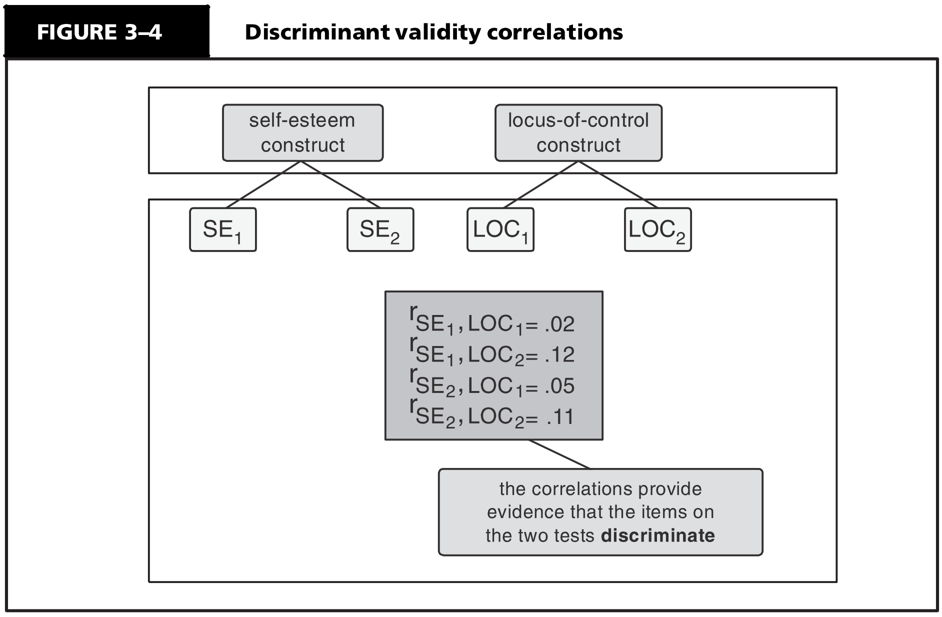
To estimate the degree to which any two measures are related to each other you would typically use the correlation coefficient discussed in Chapter 12. That is, you look at the patterns of intercorrelations among the measures. Correlations between theoretically similar measures should be "high," whereas 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 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 3–3, you see four measures (each is an item on a scale) that all purport to 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 giving your scale out to a sample of respondents. You

FIGURE 3–4 **Discriminant validity correlations**



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 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 (more about this later). 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.

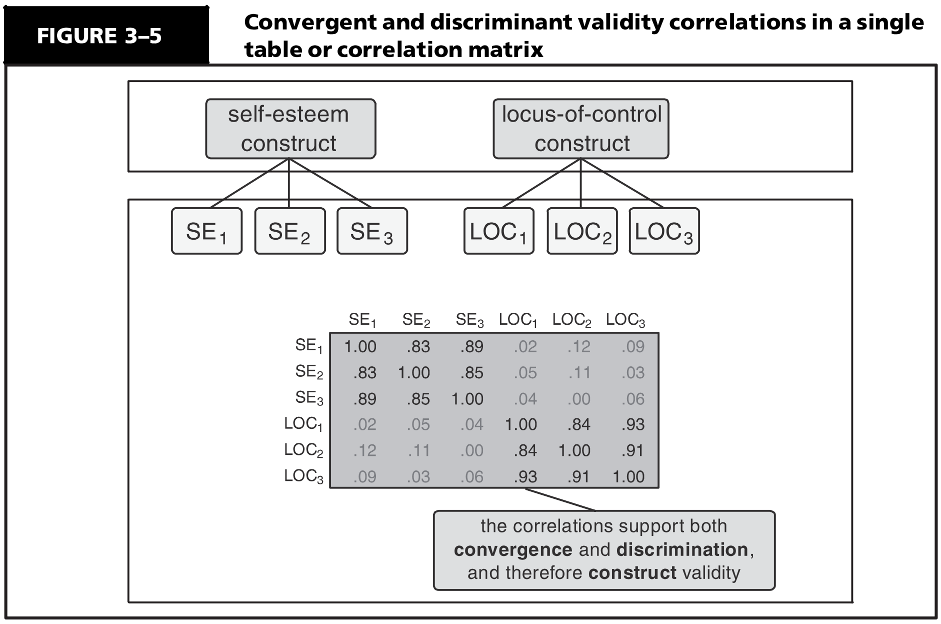
Discriminant Validity To establish **discriminant validity**, you need to show that measures that should not be related are in reality not related. In Figure 3–4, you again see four measures (each is an item on a scale). Here, however, two of the items are thought to reflect the construct of self-esteem, whereas the other two are thought to reflect 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 low 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 3–3.

discriminant validity
The degree to which concepts that should not be related theoretically are, in fact, not interrelated in reality.

As I mentioned previously, just because there is evidence that the two sets of two measures seem to be unrelated 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 types of validity are supported. Given the previous discussions of convergent and discriminant validity, you should be able to see that you could put both

FIGURE 3–5 **Convergent and discriminant validity correlations in a single table or correlation matrix**



	SE_1	SE_2	SE_3	LOC_1	LOC_2	LOC_3
SE_1	1.00	.83	.89	.02	.12	.09
SE_2	.83	1.00	.85	.05	.11	.03
SE_3	.89	.85	1.00	.04	.00	.06
LOC_1	.02	.05	.04	1.00	.84	.93
LOC_2	.12	.11	.00	.84	1.00	.91
LOC_3	.09	.03	.06	.93	.91	1.00

the correlations support both **convergence** and **discrimination**, and therefore **construct** validity

principles together into a single analysis to examine both at the same time. This is illustrated in Figure 3–5.

Figure 3–5 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 a darker font: 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. In addition, two 3 ■ 3 blocks of discriminant coefficients appear in a lighter-shaded font, 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 12.)

How do you make sense of the correlations' patterns? Remember that I said previously that 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 all convergent correlations are always higher than any of 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 the discussion of types of construct validity in this chapter for more information.)

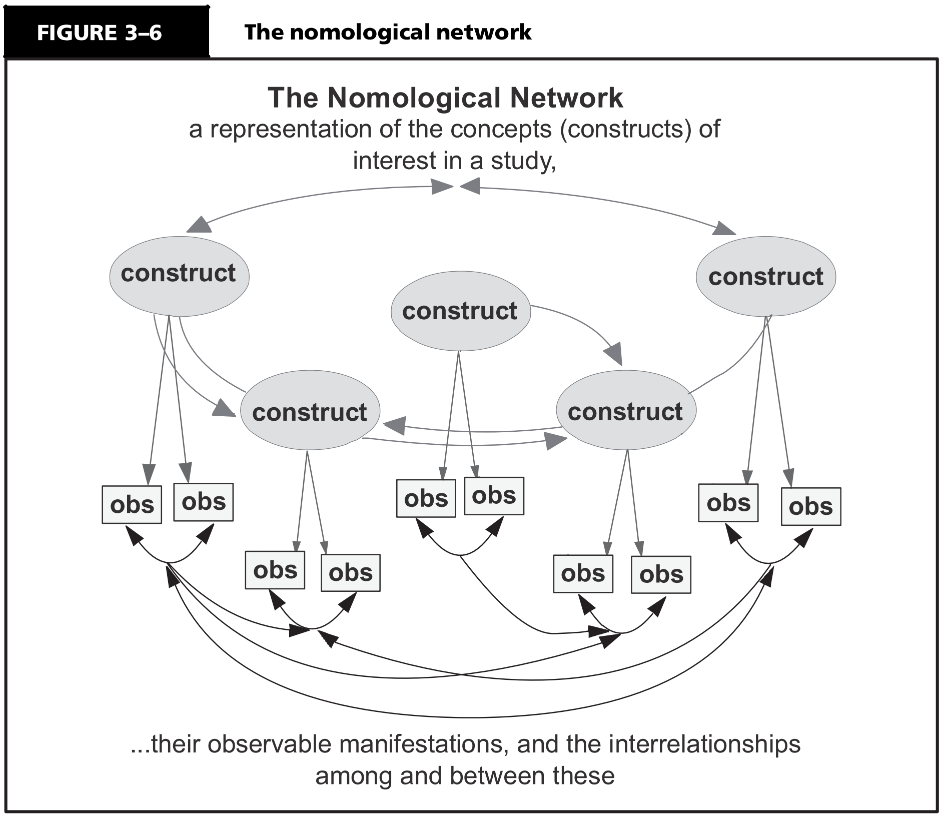
One of the most powerful approaches is to include even more constructs and measures. The more complex your theoretical model (if you find confirmation of the correct pattern in the correlations), the more evidence you are providing that you know what you're talking about (theoretically speaking). Of 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 using approaches like the MTMM and the nomological network described later in this chapter.

Perhaps the most interesting approach to getting at construct validity involves the idea of pattern matching. Instead of viewing convergent and discriminant validity as differences of kind, pattern matching views them as differences in degree. Because of this, pattern matching seems a more reasonable idea when compared with the MTMM and the nomological network, and helps you avoid the problem of how high or low correlations need to be to say that you've established convergence or discrimination.

3-1d The Nomological Network

The nomological network (Figure 3–6) is an idea that was developed by Lee Cronbach and Paul Meehl in 1955 as part of the American Psychological Association's efforts to develop standards for psychological testing. The term *nomological* is derived from Greek and means lawful, so the nomological network can be thought

FIGURE 3–6 The nomological network


The Nomological Network
a representation of the concepts (constructs) of interest in a study,
construct
construct
construct
construct
construct
obs
obs
obs
obs
obs
obs
obs
obs
obs
obs
obs
obs
...their observable manifestations, and the interrelationships among and between these

of as the lawful network. The nomological network was Cronbach and Meehl's view of construct validity. In short, to provide evidence that your measure has construct validity, Cronbach and Meehl argued that you had to develop a nomological network for your measure. This network would include the theoretical framework for what you are trying to measure, an empirical framework for how you are going to measure it, and specification of the linkages among and between these two frameworks.

According to Cronbach and Meehl, the nomological network is founded on the following principles that guide the researcher trying to establish construct validity:

- "Scientifically, to make clear what something is or means, so that laws can be set forth in which that something occurs.
- The laws in a nomological network may relate to:
  - Observable properties or quantities to each other
  - Different theoretical constructs to each other
  - Theoretical constructs to observables
- At least some of the laws in the network must involve observables.
- Learning more about a theoretical construct is a matter of elaborating the nomological network in which it occurs or of increasing the definiteness of its components.
- The basic rule for adding a new construct or relation to a theory is that it must generate laws (nomologicals) confirmed by observation or reduce the number of nomologicals required to predict some observables.
- Operations which are qualitatively different overlap or measure the same thing."

What Cronbach and Meehl were trying to do with this idea is to link the conceptual/theoretical realm with the observable one because this is the central concern of construct validity. Although the nomological network idea may be useful as a philosophical foundation for construct validity, it does not provide a practical and usable methodology for actually assessing construct validity. The next phase in the evolution of the idea of construct validity—the development of the MTMM—moved us a bit further toward a methodological approach to construct validity.

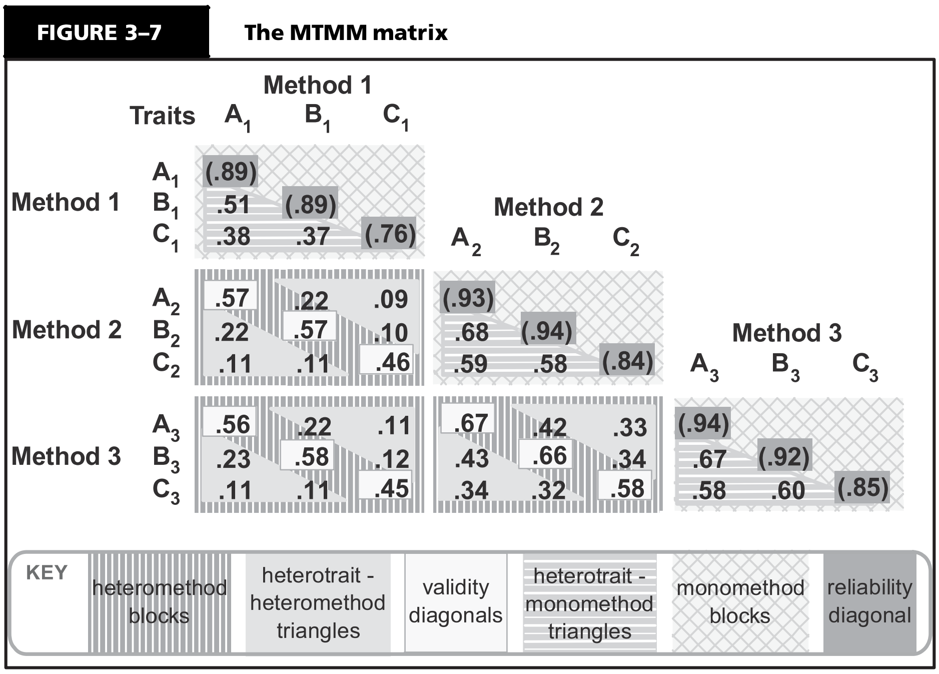
3-1e The Multitrait-Multimethod Matrix

multitrait-multimethod (MTMM) matrix A matrix of correlations arranged to facilitate the assessment of construct validity. The MTMM assumes that you have measured each construct (trait) with different methods in a fully crossed design (traits by methods).

The **multitrait-multimethod matrix** (hereafter labeled **MTMM**) is an approach to assessing the construct validity of a set of measures in a study. It was developed in 1959 by Campbell and Fiske in part as an attempt to provide a practical methodology that researchers could actually use (as opposed to the nomological network idea, which was theoretically useful but did not include a methodology). Along with the MTMM, Campbell and Fiske introduced two new types of validity: convergent and discriminant—as subcategories of construct validity. To recap, convergent validity is the degree to which concepts that should be related theoretically are interrelated in reality. Discriminant validity is the degree to which concepts that should *not* be related theoretically are, in fact, *not* interrelated in reality. You can assess both convergent and discriminant validity using the MTMM. To be able to claim that your measures have construct validity, you have to demonstrate both convergence and discrimination.

The MTMM (Figure 3–7) is simply a matrix or table of correlations arranged to facilitate the assessment of construct validity. The MTMM assumes that you measure each of several concepts (called *traits* by Campbell and Fiske) by each of several methods (such as a paper-and-pencil test, a direct observation, or a performance

FIGURE 3–7 **The MTMM matrix**



	Traits	Method 1 A_1	Method 1 B_1	Method 1 C_1	Method 2 A_2	Method 2 B_2	Method 2 C_2	Method 3 A_3	Method 3 B_3	Method 3 C_3
Method 1	A_1	(.89)								
	B_1	.51	(.89)							
	C_1	.38	.37	(.76)						
Method 2	A_2	.57	.22	.09	(.93)					
	B_2	.22	.57	.10	.68	(.94)				
	C_2	.11	.11	.46	.59	.58	(.84)			
Method 3	A_3	.56	.22	.11	.67	.42	.33	(.94)		
	B_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

KEY
heteromethod blocks
heterotrait - heteromethod triangles
validity diagonals
heterotrait - monomethod triangles
monomethod blocks
reliability diagonal

measure). The MTMM is a restrictive methodology; ideally, you should measure *each* concept by *each* method.

To construct an MTMM, you need to arrange the correlation matrix by methods within concepts. Figure 3–7 shows an MTMM for three concepts (traits 1, 2, and 3), each of which is measured with three different methods (A, B, and C). Note that you lay the matrix out in blocks by *method.* Essentially, the MTMM is just a correlation matrix between your measures, with one exception: instead of 1's along the diagonal (as in the typical correlation matrix) you substitute an estimate of the reliability of each measure as the diagonal (see Section 3-2, Reliability).

Before you can interpret an MTMM, you must understand how to identify the different parts of the matrix. First, you should note that the matrix consists of nothing but correlations. It is a square, symmetric matrix, so you need to look at only half of it. Figure 3–7 shows the lower triangle. Second, these correlations can be grouped into three kinds of shapes: diagonals, triangles, and blocks. The specific shapes are as follows:

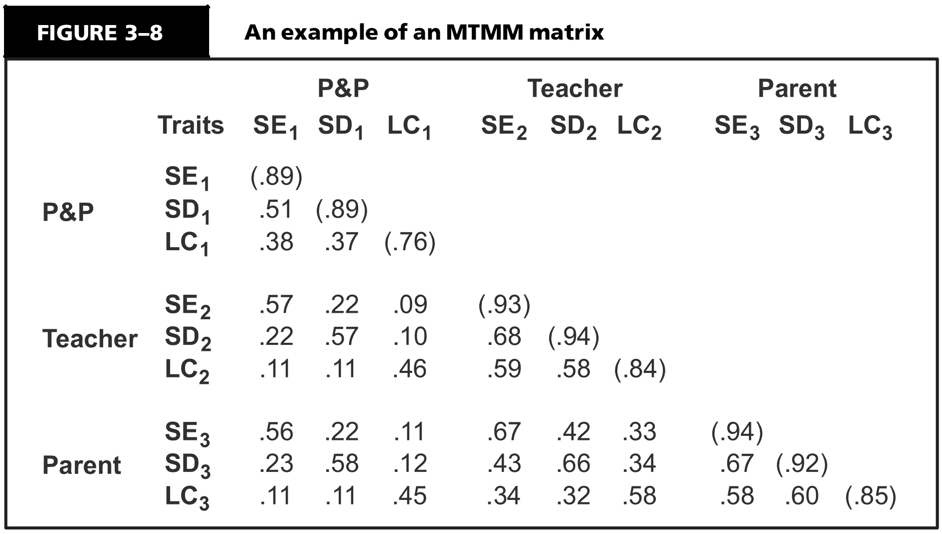
- *The reliability diagonal (monotrait-monomethod).* These are estimates of the reliability of each measure in the matrix. You can estimate reliabilities in different ways (for example, test-retest or internal consistency). There are as many correlations in the reliability diagonal as there are measures; in this example, there are nine measures and nine reliabilities. The first reliability in the example is the correlation of Trait A, Method 1 with Trait A, Method 1. (Hereafter, I'll abbreviate this relationship A1-A1). Notice that this is essentially the correlation of the measure with itself. In fact, such a correlation would always be perfect ($r = 1.0$). Instead, you substitute an estimate of reliability. You could also consider these values to be monotrait-monomethod correlations.
- *The validity diagonals (monotrait-heteromethod).* These are correlations between measures of the same trait measured using different methods. Since the MTMM is organized into method blocks, there is one validity diagonal in each method block. For example, look at the A1-A2 correlation of .57 in Figure 3–7. This is the correlation between two measures of the same trait (A) measured with two different measures (1 and 2). Because the two measures are of the same trait or concept, you would expect them to be strongly correlated. You could also consider these values to be monotrait-heteromethod correlations.

- *The heterotrait-monomethod triangles.* These are the correlations among measures that share the same method of measurement, for instance, A1-B1 = .51 in the upper left heterotrait-monomethod triangle in Figure 3–7. Note that what these correlations share is method, not trait or concept. If these correlations are high, it is because measuring different things with the same method results in correlated measures. Or, in more straightforward terms, you have a strong methods factor.
- *Heterotrait-heteromethod triangles.* These are correlations that differ in both trait and method. For instance, A1-B2 is .22 in the example in Figure 3–7. Generally, because these correlations share neither trait nor method you expect them to be the lowest in the matrix.
- *The monomethod blocks.* These consist of all of the correlations that share the same method of measurement. There are as many blocks as there are methods of measurement.
- *The heteromethod blocks.* These consist of all correlations that do *not* share the same methods. There are [*K*(*K* 1)]/2 such blocks, where *K* = the number of methods. In the example in Figure 3–7, there are three methods, so there are [3(3 1)]/2 = [3(2)]/2 = 6/2 = 3 such blocks.

Principles of Interpretation Now that you can identify the different parts of the MTMM, you can begin to understand the rules for interpreting it. You should realize that MTMM interpretation requires the researcher to use judgment. Even though some of the principles might be violated in a specific MTMM, you might still wind up concluding that you have fairly strong construct validity. In other words, you won't necessarily get *perfect* adherence to these principles in applied research settings, even when you do have evidence to support construct validity. To me, interpreting an MTMM is a lot like a physician's reading of an x-ray. A practiced eye can often spot things that the neophyte misses! A researcher who is experienced with MTMM can use it to identify weaknesses in measurement as well as to assess construct validity.

To help make the principles more concrete, let's make the example a bit more realistic. Imagine that you are going to conduct a study of sixth-grade students and you want to measure three traits or concepts: Self-Esteem (SE), Self-Disclosure (SD), and Locus of Control (LC). Furthermore, you want to measure each of these traits three different ways: a Paper-and-Pencil (P&P) measure, a Teacher rating, and a Parent rating. The results are arrayed in the MTMM as shown in Figure 3–8. As the principles are presented, try to identify the appropriate coefficients in the MTMM and make a judgment yourself about the strength of construct validity claims.

FIGURE 3–8 **An example of an MTMM matrix**



		P&P			Teacher			Parent		
	Traits	SE_1	SD_1	LC_1	SE_2	SD_2	LC_2	SE_3	SD_3	LC_3
P&P	SE_1	(.89)								
	SD_1	.51	(.89)							
	LC_1	.38	.37	(.76)						
Teacher	SE_2	.57	.22	.09	(.93)					
	SD_2	.22	.57	.10	.68	(.94)				
	LC_2	.11	.11	.46	.59	.58	(.84)			
Parent	SE_3	.56	.22	.11	.67	.42	.33	(.94)		
	SD_3	.23	.58	.12	.43	.66	.34	.67	(.92)	
	LC_3	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

The following list contains the basic principles or rules for the MTMM. You use these rules to determine the strength of the construct validity:

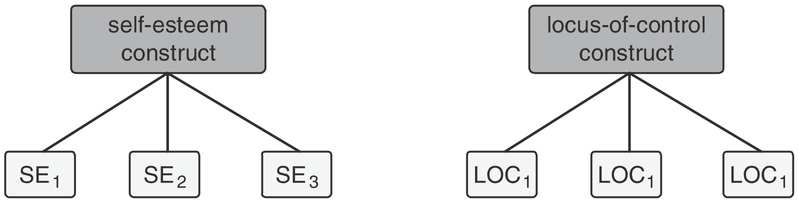
- Coefficients in the reliability diagonal should consistently be the highest in the matrix. That is, a trait should be more highly correlated with itself than with anything else! This rule is uniformly true in the example in Figure 3–8.
- Coefficients in the validity diagonals should be significantly different from zero and high enough to warrant further investigation. This rule is essentially evidence of convergent validity. All of the correlations in Figure 3–8 meet this criterion.
- A validity coefficient should be higher than the values in its column and row in the same heterotrait-heteromethod triangle. In other words, (SE P&P) —(SE Teacher) should be greater than (SE P&P) (SD Teacher), (SE P&P) (LC Teacher), (SE Teacher) (SD P&P), and (SE Teacher) (LC P&P). This is true in all cases in Figure 3–8.
- A validity coefficient should be higher than all coefficients in the heterotrait-monomethod triangles. This rule essentially emphasizes that trait factors should be stronger than methods factors. Note that this is *not* true in all cases in the example in Figure 3–8. For instance, the (LC P&P) —(LC Teacher) correlation of .46 is less than (SE Teacher) (SD Teacher), (SE Teacher) (LC Teacher), and (SD Teacher) (LC Teacher)—evidence that there might be a methods factor, especially on the Teacher observation method.
- The same *pattern* of trait interrelationship should be seen in all triangles. The example in Figure 3–8 clearly meets this criterion. Notice that in all triangles the SE-SD relationship is approximately twice as large as the relationships that involve LC.

Advantages and Disadvantages of MTMM The MTMM idea provided an operational methodology for assessing construct validity. In the one matrix, it was possible to examine both convergent and discriminant validity simultaneously. By including methods on an equal footing with traits, Campbell and Fiske stressed the importance of looking for the effects of how we measure in addition to what we measure. In addition, MTMM provided a rigorous framework for assessing construct validity.

Despite these advantages, MTMM has received little use since its introduction in 1959 for several reasons. First, in its purest form, MTMM requires a fully crossed measurement design; each of several traits is measured by each of several methods. Although Campbell and Fiske explicitly recognized that one could have an incomplete design, they stressed the importance of multiple replication of the same trait across methods. In some applied research contexts, it just isn't possible to measure all traits with all desired methods. (For example, what would you use to obtain multiple observations of weight?) In most applied social research, it isn't feasible to make methods an explicit part of the research design. Second, the judgmental nature of the MTMM may have worked against its wider adoption (although it should actually be perceived as a strength). Many researchers wanted a test for construct validity that would result in a single statistical coefficient that could be tested—the equivalent of a reliability coefficient. It was impossible with MTMM to quantify the *degree* of construct validity in a study. Finally, the judgmental nature of MTMM meant that different researchers could legitimately arrive at different conclusions.

A Modified MTMM —Leaving Out the Methods Factor What if we try to obtain some of the benefits of the MTMM while minimizing some of the disadvantages that have limited its use? One of the major limiting aspects of the MTMM is the requirement that each construct be measured with multiple methods, a requirement that is just not practical in most applied social research. What if we eliminate that requirement? In this case, the MTMM becomes equivalent to a

FIGURE 3–9 **MTMM emphasizes methods as confounding factors**



	SE$_1$	SE$_2$	SE$_3$	LOC$_1$	LOC$_2$	LOC$_3$
SE$_1$	1.00	.83	.89	.02	.12	.09
SE$_2$	.83	1.00	.85	.05	.11	.03
SE$_3$	.89	.85	1.00	.04	.00	.06
LOC$_1$	.02	.05	.04	1.00	.84	.93
LOC$_2$	.12	.11	.00	.84	1.00	.91
LOC$_3$	.09	.03	.06	.93	.91	1.00

multitrait matrix and would look like the matrix shown earlier in Figure 3–5 when describing convergent and discriminant validity.

The important thing to notice about the matrix in Figure 3–5 is that *it does not include a methods factor* as a true MTMM would. The matrix does examine both convergent and discriminant validity (just like the MTMM) but it explicitly looks at only construct intrarelationships and interrelationships. That is, it doesn't look at methods relationships like a full MTMM does.

You can see in Figure 3–9 that the MTMM idea really had two major themes. The first is the idea of looking simultaneously at the pattern of convergence and discrimination. This idea is similar in purpose to the notions implicit in the nomological network described earlier; you are looking at the *pattern* of interrelationships based on your theory of the nomological net. The second idea in MTMM is the emphasis on methods as a potential confounding factor.

Although methods may confound the results, they won't necessarily do so in any given study; and, perhaps it is too much to ask of any single methodology that it simultaneously be able to assess construct validity and address the potential for methods factors in measurement. Perhaps if you split the two agendas, you will find that the feasibility of examining convergent and discriminant validity is greater; but what do you do about methods factors? One way to deal with them is to replicate research projects, rather than try to incorporate a methods test into a single research study. Thus, if you find a particular outcome in a study using several measures, you might see whether that same outcome is obtained when you replicate the study using *different methods of measurement* for the same constructs. The methods issue is considered more as an issue of generalizability (across measurement methods) rather than one of construct validity.

When viewed without the methods component, the idea of a MTMM is a much more realistic and practical approach for assessing convergent and discriminant validity, and hence construct validity. You will see that when you move away from the explicit consideration of methods and when you begin to see convergence and discrimination as differences of degree, you essentially have the foundation for the pattern matching approach to assessing construct validity as discussed in the following section.

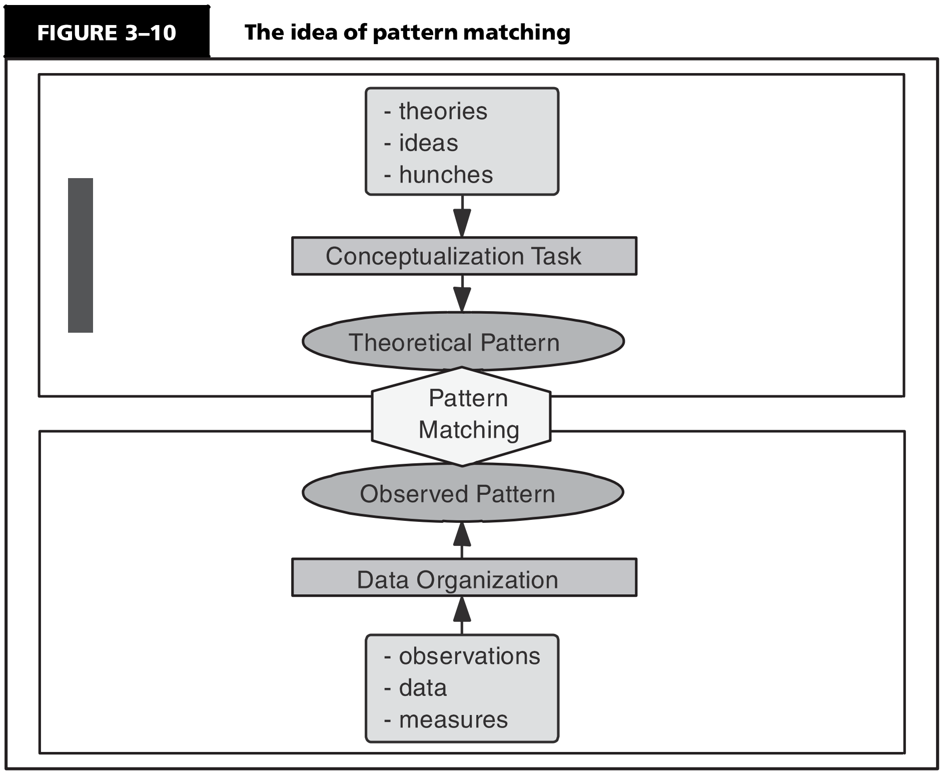
3-1f Pattern Matching for Construct Validity

The idea of using pattern matching as a rubric for assessing construct validity is an area in which I have tried to make a contribution (Trochim, 1985, 1989), although my work was clearly foreshadowed, especially in much of Donald T. Campbell's writings on the MTMM. Here, I'll try to explain what I mean by pattern matching with respect to construct validity.

The Theory of Pattern Matching A pattern is any arrangement of objects or entities. The term *arrangement* is used here to indicate that a pattern is by definition nonrandom and at least potentially describable. All theories imply some pattern, but theories and patterns are not the same thing. In general, a theory postulates structural relationships between key constructs. The theory can be used as the basis for generating patterns of predictions. For instance, E = MC can be considered a theoretical formulation. A pattern of expectations can be developed from this formula by generating predicted values for one of these variables given fixed values of the others. Not all theories are stated in mathematical form, especially in applied social research, but all theories provide information that enables the generation of patterns of predictions.

Pattern matching always involves an attempt to link two patterns where one is a theoretical pattern and the other is an observed or operational one. The top part of Figure 3–10 shows the realm of theory. The theory might originate from a formal tradition of theorizing, might be the investigator's ideas or hunches, or might arise from some combination of these. The conceptualization task involves the translation of these ideas into a specifiable theoretical pattern indicated by the top shape in the figure. The bottom part of the figure indicates the realm of observation. This is broadly meant to include direct observation in the form of impressions, field notes, and the like, as well as more formal objective measures. The collection or organization of relevant operationalizations (relevant to the theoretical pattern) is termed the *observational pattern* and is indicated by the lower shape in the figure.

FIGURE 3–10 **The idea of pattern matching**



The inferential task involves the attempt to relate, link, or match these two patterns as indicated by the double arrows in the center of the figure. To the extent that the patterns match, one can conclude that the theory and any other alternative theories that predict the same observed pattern may be plausible explanations for it.

It is important to demonstrate that no plausible alternative theories can account for the observed pattern and this task is made much easier when the theoretical pattern of interest is a unique one. In effect, a more complex theoretical pattern is like a unique fingerprint that one is seeking in the observed pattern. With more complex theoretical patterns, it is usually more difficult to construe sensible alternative patterns that would also predict the same result. To the extent that theoretical and observed patterns do not match, the theory may be incorrect or poorly formulated, the observations may be inappropriate or inaccurate, or some combination of both states may exist.

All research employs pattern-matching principles, although this is seldom done consciously. In the traditional two-group experimental context (see Chapter 9), for instance, the typical theoretical outcome pattern is the hypothesis that there will be a significant difference between treated and untreated groups. The observed outcome pattern might consist of the averages for the two groups on one or more measures. The pattern match is accomplished by a test of significance such as the *t*-test or ANOVA. In survey research, pattern matching forms the basis of generalizations across different concepts or population subgroups. (This is covered in Chapter 4.) In qualitative research pattern matching lies at the heart of any attempt to conduct thematic analyses. (This is discussed in Chapter 6.)

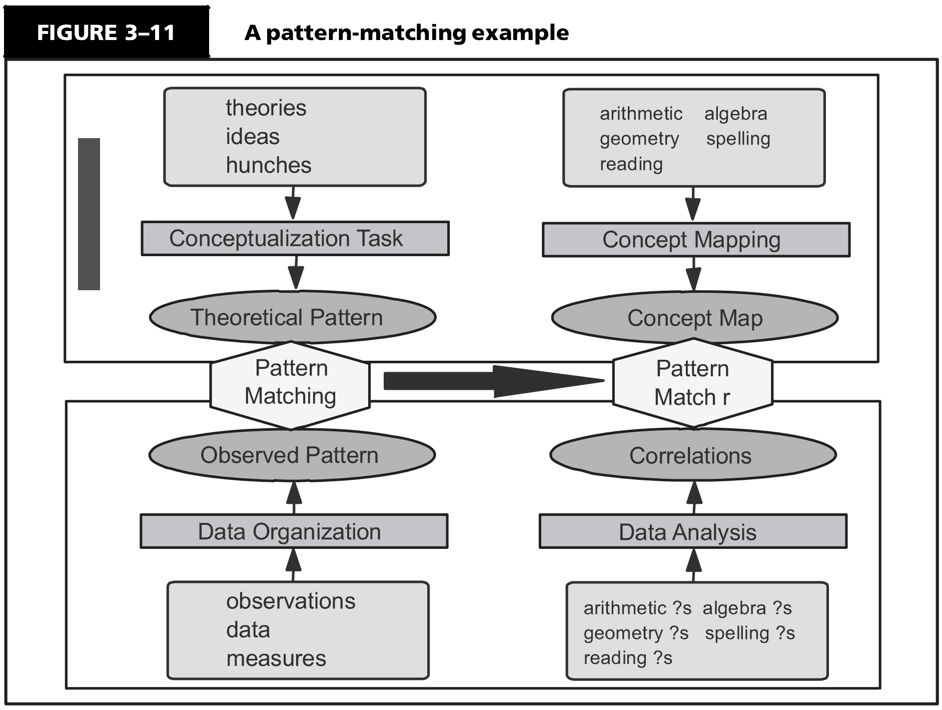
Although current research methods can be described in pattern-matching terms, the idea of pattern matching implies more and suggests how one might improve on these current methods. Specifically, pattern matching implies that *more complex patterns, if matched, yield greater validity for the theory.* Pattern matching does not differ fundamentally from traditional hypothesis testing and model building approaches. A theoretical pattern is a hypothesis about what is expected in the data. The observed pattern consists of the data used to examine the theoretical model. The major differences between pattern matching and more traditional hypothesis-testing approaches are that pattern matching encourages the use of more complex or detailed hypotheses and treats the observations from a multivariate rather than a univariate perspective.

Pattern Matching and Construct Validity Although pattern matching can be used to address a variety of questions in social research, the emphasis here is on its use in assessing construct validity.

Figure 3-11 shows the pattern-matching structure for an example involving five measurement constructs: arithmetic, algebra, geometry, spelling, and reading. This example uses concept mapping (see Chapter 2) to develop the theoretical pattern among these constructs. In concept mapping, you generate a large set of potential arithmetic, algebra, geometry, spelling, and reading questions. You sort them into piles of similar questions and develop a map that shows each question in relation to the others. On the map, questions that are more similar are closer to each other; those that are less similar are more distant. From the map, you can find the straight-line distances between all pair of points (all questions). This mapping is the matrix of interpoint distances. You might use the questions from the map when constructing your measurement instrument, or you might sample from these questions. On the observed side, you have one or more test instruments that contain a number of questions about arithmetic, algebra, geometry, spelling, and reading. You analyze the data and construct a matrix of interitem correlations.

What you want to do is compare the matrix of interpoint distances from your concept map (the theoretical pattern) with the correlation matrix of the questions (the observed pattern). How do you achieve this? Let's assume that you had 100 prospective questions on your concept map, 20 for each construct. Correspondingly, you

FIGURE 3–11 **A pattern-matching example**



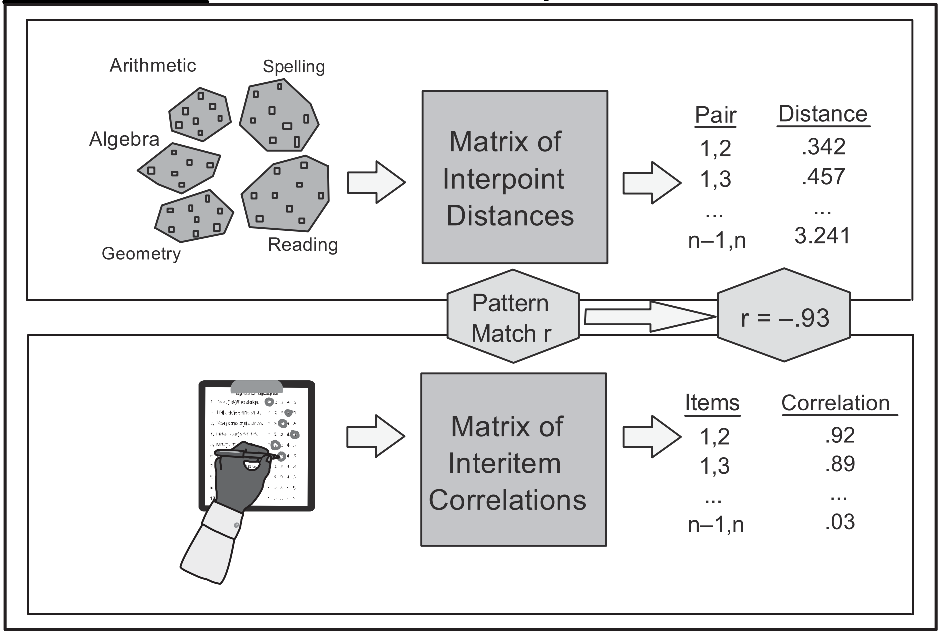
have 100 questions on your measurement instrument, 20 in each area. Thus, both matrices are 100 100 in size. Because both matrices are symmetric, you actually have $[N(N-1)]/2 = [100(99)]/2 = 9900/2 = 4950$ unique pairs (excluding the diagonal). If you string out the values in each matrix, you can construct a vector or column of 4950 numbers for each matrix. The first number is the value comparing pair (1,2); the next is (1,3), and so on, to (N 1,N) or (99,100). This procedure is illustrated in Figure 3–12. Now, you can compute the overall correlation between these two columns, which is the correlation between the theoretical and observed patterns (the pattern matching correlation). In this example, let's assume it is .93. Why would it be a *negative* correlation? Because you are correlating *distances* on the map with the *similarities* in the correlations and you expect that *greater* distance on the map should be associated with *lower* correlation and *less* distance with *greater* correlation.

The pattern matching correlation is the overall estimate of the degree of construct validity in this example because it estimates the degree to which the operational measures reflect your theoretical expectations.

Advantages and Disadvantages of Pattern Matching The pattern-matching approach to construct validity has several disadvantages. The most obvious is that pattern matching requires that you specify your theory of the constructs rather precisely. This is typically not done in applied social research, at least not to the level of specificity implied here; but perhaps it *should* be done. Perhaps the more restrictive assumption in pattern matching is that you are able to structure the theoretical and observed patterns the same way so that you can directly correlate them. This method requires you to quantify both patterns and, ultimately, describe them in matrices that have the same dimensions. In most research as it is currently done, it is relatively easy to construct a matrix of the interitem correlations from the data. However, researchers seldom currently use methods like concept mapping to estimate theoretical patterns that can be linked with the observed ones. Again, perhaps this ought to be done more frequently.

The pattern-matching approach has a number of advantages, especially relative to the MTMM. First, it is more *general* and *flexible* than MTMM. It does not require

FIGURE 3–12 Linking expected relationships as measured through concept mapping to observed intercorrelations to achieve a pattern match for construct validity



that you measure each construct with multiple methods. Second, it treats convergence and discrimination as a *continuum*. Concepts are more or less similar and so their interrelations would be more or less convergent or discriminant, which moves the convergent/discriminant distinction away from the simplistic dichotomous categorical notion to one that is more suitably post-positivist and continuous in nature. Third, the pattern-matching approach does make it possible to estimate the overall construct validity for a set of measures in a specific context—it is the correlation of the theoretical expectations with the observed relationships. Notice that you don't estimate construct validity for a single measure because construct validity, like discrimination, is always a relative metric. Just as you can ask only whether you have distinguished something if there is something to distinguish it from, you can assess construct validity only in terms of a theoretical semantic or nomological net, the conceptual context within which it resides. The pattern-matching correlation tells you, for your particular study, whether there is a demonstrable relationship between how you theoretically expect your measures to interrelate and how they interrelate in practice. Finally, because pattern matching requires a more specific theoretical pattern than you typically articulate, it *requires* you to specify what you think about the constructs. That's got to be a good thing.

Social research has long been criticized for conceptual sloppiness, for repackaging old constructs in new terminology and failing to develop an evolution of research around key theoretical constructs. Perhaps the emphasis on theory articulation in pattern matching will encourage researchers to be more careful about the conceptual underpinnings of their empirical work, and, after all, isn't that what construct validity is all about?

3-1g Structural Equation Modeling

Like pattern matching, structural equation modeling (SEM) is a method that allows researchers to compare the theoretical with the actual, that is, models based on unobserved ideas with models based on data from observed measurements. In fact, we might even describe it as a very good and general way to do MTMM analysis, pattern-matching analysis, or practically any sort of problem that involves an assessment of construct validity. It is a method that is flexible enough to fit specific construct validity questions, whether having to do with the structural validity of a single measure or of an entire system of related constructs as in MTMM analysis. In a sense, we could say that the mathematics and technology needed to examine a model like the one Cronbach and Meehl proposed in their 1955 article (for example, the nomological network in Figure 3–6) have caught up with the theory a mere 50 years later! As software to conduct SEM studies has improved, these kinds of studies have become common in the social science literature. Graphic user interfaces (GUIs) make it possible to draw the model like the one in Figure 3–6 as you would on paper, and then submit the model for testing based on the data you have collected to represent the constructs. With SEM, we can now estimate the overall goodness of fit of the theory to the data as well as study component relationships within the model. As a final note, we would do well to be mindful that construct validation is a "never-ending procedure" (Fiske, 2002, p. 175), even with the sophistication of SEM computer programs that can solve a large and complex set of equations simultaneously. Just as methods of evaluating aspects of validity will evolve, so will theories that generate the constructs we seek to understand.

3-1h 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.

I take the list of threats from the discussion in Cook and Campbell (1979). Although I love their discussion, I do find some of their terminology less than straightforward; much of what I'll do here is try to explain this stuff in terms that the rest of us might hope to understand. One way we can sort the threats to validity is into two major categories: those resulting implicitly from the study design and those arising from the behavior or participants.

Design Threats to Validity

Inadequate Preoperational Explication of Constructs. This section title isn't nearly as ponderous as it sounds. Here, *preoperational* means before translating constructs into measures or treatments, and *explication* means explanation; in other words, *you didn't do a good enough job of defining (operationally) what you mean by the construct.* 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 methods (for example, concept mapping) to articulate your concepts.
- Get experts to critique your operationalizations.

Mono-Operation Bias. **Mono-operation bias** pertains to the independent variable, cause, program, or treatment in your study: it does not pertain to measures or outcomes (see mono-method bias in the following section). *If you use only 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 flawed relative to 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

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.

study reflect only the peculiar version of the program that you implemented, and not the actual construct you had in mind. Solution: try to implement multiple versions of your program.

mono-method bias
A threat to construct validity that occurs because you use only a single method of measurement.

Mono-Method Bias. **Mono-method bias** refers 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 measuring only 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. *In the real world, you cannot control or even know about all of the possible experiences that participants might have had that could possibly interact with the things your study has included.* For example, let's say 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 birth 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.) In this case, you might 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 confound 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 in effect 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 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, one of which gets the pretest and the other not. 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. (This is discussed in Section 9-6a, The Solomon Four-Group Design.)

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 does cause a reduction in symptoms, but what you failed to anticipate was the drastic negative consequences of the side effects of the treatment. When you say that Treatment X is effective, you have defined *effective* in regard to only the directly targeted symptom. 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 any potential outcomes. For instance, the drug Viagra was not originally developed to help erectile dysfunction. It was created as a drug for hypertension. When that didn't pan out, it was tried as an anti-angina medicine. (The chemists had reason to think a drug designed for hypertension might work on angina.) It was only a chance observation, when the drug was being tested in Wales and men were reporting penile erections, that led the pharmaceutical company to investigate that potential outcome. This is an example of an unintended positive outcome (although there is more recent evidence on Viagra to suggest that the initial enthusiasm needs to be tempered by the potential for its own unanticipated negative side effects).

Confounding Constructs and Levels of Constructs. *This issue has to do with the decisions you make about how frequently or how intensely your study participants are exposed to the independent variable (the program, treatment, or intervention).* Imagine a study to test the effect of a new drug treatment for cancer. A fixed dose of the drug is given to a randomly assigned treatment group and a placebo to the other group. No treatment effects are detected, or perhaps the observed result is true only for a certain dosage level. Slight increases or decreases of the dosage may radically change the results. In this context, it is not fair for you to use the label for the drug as a description for your treatment because you looked only at a narrow range of dose. 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 1 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 will achieve the desired outcome.

The Social Threats to Construct Validity. The remaining major threats to construct validity can be distinguished from the ones I discussed previously because they all stem from the social and human nature of the research endeavor. I cover these in the following sections.

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 makes them perform poorly (and not your program conditions), 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. (I guess this would be the social research equivalent to the mid-stream urine sample!) 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. 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 this information in that analysis to attempt to adjust for expectations.

3-2 Reliability

reliability
The degree to which a measure is consistent or dependable; the degree to which it would give you the same result over and over again, assuming the underlying phenomenon is not changing.

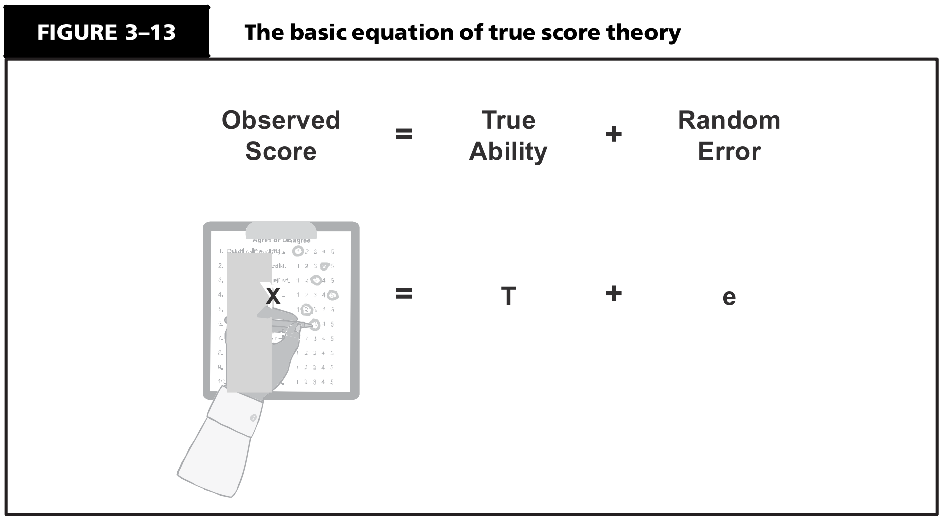
Reliability has to do with the quality of measurement. In its everyday sense, reliability is the consistency or repeatability of your measures. Before I can define reliability precisely, I have to lay the groundwork. First, you have to learn about the foundation of reliability, the true score theory 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 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
A theory that maintains that every measurement is an additive composite of two components: the true ability of the respondent and random error.

True score theory is a theory about measurement. Like all theories, you need to recognize that it is not proved; 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 an additive composite of two

FIGURE 3–13 The basic equation of true score theory



components: true ability (or the true level) of the respondent on that measure and random error. This is illustrated in Figure 3–13. You observe the measurement: a score on the 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. (Only God knows what those values are.) You assume that there are only the two components to the right side of the equal sign in the equation.

The simple equation of $X = T + e$ has a parallel equation at the level of the variance or variability of a measure. That is, across a set of scores, you can assume

$$X \qquad T \qquad e_X$$

In more human terms, this means that the variability of your measure is the sum of the variability due to true score and the variability due to random error. This will have important implications when we consider some of the more advanced models for adjusting for errors in measurement later in Section 14-4a, Nonequivalent Groups Analysis.

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, which will be discussed later in this chapter. A measure that has no random error (is all true score) is perfectly reliable; a measure that has no true score (is all 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

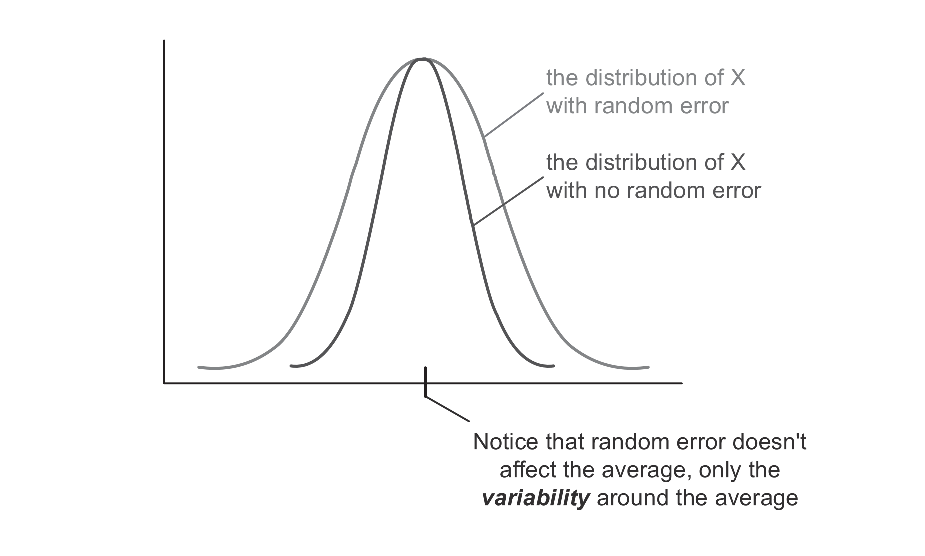
True score theory is a good simple model for measurement, but it may not always be an accurate reflection of reality. In particular, it assumes that any observation is composed of the true value plus some random error value; but is that reasonable? What if all error is not random? Isn't it possible that 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

FIGURE 3–14 **Random and systematic errors in measurement**

$$X = T + e$$

$$X = T + e_r + e_s$$

FIGURE 3–15 **Random error adds variability to a distribution but does not affect central tendency (the average)**

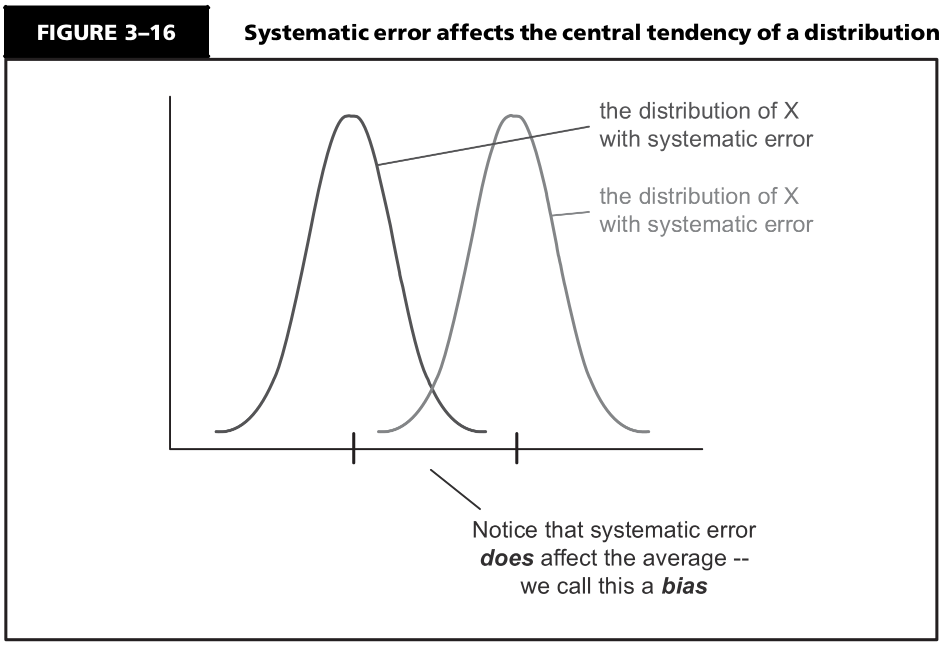


into two subcomponents, random error and systematic error. Figure 3–14 shows these two components of measurement error, what the difference between them is, and how they affect research.

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*. God can see the random errors, but she's not telling us what they are!) The important property of random error is that it adds variability to the data but does not affect average performance for the group (Figure 3–15). 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 loud traffic going by just outside of a classroom where students are taking a test, this noise is liable 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 (Figure 3–16).

FIGURE 3–16 Systematic error affects the central tendency of a distribution

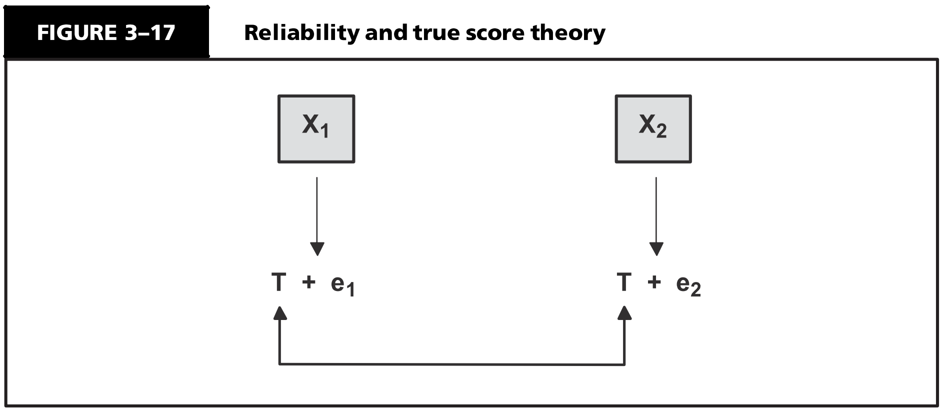


Reducing Measurement Error So, how can you reduce measurement errors, random or systematic? One thing you can do is to pilot test your instruments to get feedback from your respondents regarding how easy or hard the measure was and information about how the testing environment affected their performance. 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-punched and verified. This means that you enter the data twice, the second time having your data-entry machine 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 modeling 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, news people 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). 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.

FIGURE 3–17 **Reliability and true score theory**



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 3–17. 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 (see Section 3-2a, True Score Theory).

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 the students' observed abilities appear lower than their true or actual abilities.

Okay, back to reliability. If your measure, *X*, is reliable, you should find that if you measure or observe it twice on the same persons, the scores should be pretty much the same, result each time you measure in time. Why would they be the same? If you look at Figure 3–17, 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* and *e*) 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 (by definition in this example) the same for both observations. What does this mean? The two observed scores, *X* and *X*■, 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). 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 same on both observations (assuming, of course, that your true ability didn't change between the two measurement occasions).

FIGURE 3–18 **Reliability can be expressed as a simple ratio**

$$\frac{\text{true level on the measure}}{\text{the entire measure}}$$

FIGURE 3–19 **The reliability ratio can be expressed in terms of variances**

$$\frac{\text{the variance of the true score}}{\text{the variance of the measure}}$$

FIGURE 3–20 **The reliability ratio expressed in terms of variances in abbreviated form**

$$\frac{\text{var(T)}}{\text{var(X)}}$$

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 3–18.

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. Remember that the variance is a measure of the spread or distribution of a set of scores. So, I can now state the definition as shown in Figure 3–19.

I might put this into slightly more technical terms by using the abbreviated name for the variance and our variable names (Figure 3–20).

We're getting to the critical part now. If you look at the equation in Figure 3–20, 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 12.) So how do you calculate the variance of the true scores? You can't see the true scores. (You only see *X*!) Only God knows the true score for a specific observation. 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, *X* and *X* ? You assume (using true score theory described earlier in this chapter) 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 *X* and *X* . Figure 3–21 shows a simple formula for the correlation.

FIGURE 3–21 **The formula for estimating reliability**

$$\frac{\text{covariance}(X_1, X_2)}{\text{sd}(X_1) * \text{sd}(X_2)}$$

standard deviation
The spread or variability of the scores around their average in a *single sample*. The standard deviation, often abbreviated sd, is mathematically the square root of the variance. The standard deviation and variance both measure dispersion, but because the standard deviation is measured in the same units as the original measure and the variance is measured in squared units, the standard deviation is usually more directly interpretable and meaningful.

In Figure 3–21, 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* and *X* are the only things the two observations share! So, the top part is essentially an estimate of *var(T)* in this context. In addition, 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 *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 the measure administered twice!

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.

There's only one other issue I want to address here. How big is an estimate of reliability? To figure this out, let's go back to the equation given earlier (Figure 3–22).

Remember, because $X = T + e$, you can substitute in the bottom of the ratio as shown in Figure 3–23.

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, var(*e*) is zero and the equation would reduce to the equation shown in Figure 3–24.

FIGURE 3–22 **The reliability ratio expressed in terms of variances in abbreviated form**

$$\frac{\text{var}(T)}{\text{var}(X)}$$

FIGURE 3–23 **The reliability ratio expressed in terms of variances with the variance of the observed score subdivided according to true score theory**

$$\frac{\text{var}(T)}{\text{var}(T) + \text{var}(e)}$$

FIGURE 3–24 When there is no error in measurement, you have perfect reliability and the reliability estimate is 1.0

$$\frac{\text{var}(T)}{\text{var}(T)}$$

FIGURE 3–25 When there is only error in measurement, you have no reliability and the reliability estimate is 0

$$\frac{0}{\text{var}(e)}$$

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 3–25.

Therefore, the reliability = 0. From this you know that reliability will always range between 0 and 1.

The value of a reliability estimate tells you the proportion of variability in the measure attributable to the true score. A reliability of .5 means that about half of the variance of the observed score is attributable to truth and half is attributable to error. A reliability of .8 means the variability is about 80% true ability and 20% error, and so on.

3-2d Types of Reliability

You learned in Section 3-2c, Theory of Reliability, that it's not possible to calculate reliability exactly. Instead, you have to estimate reliability, and this is always an imperfect endeavor. Here, I want to introduce the major reliability estimators and talk about their strengths and weaknesses.

There are four general classes of reliability estimates, each of which estimates reliability in a different way:

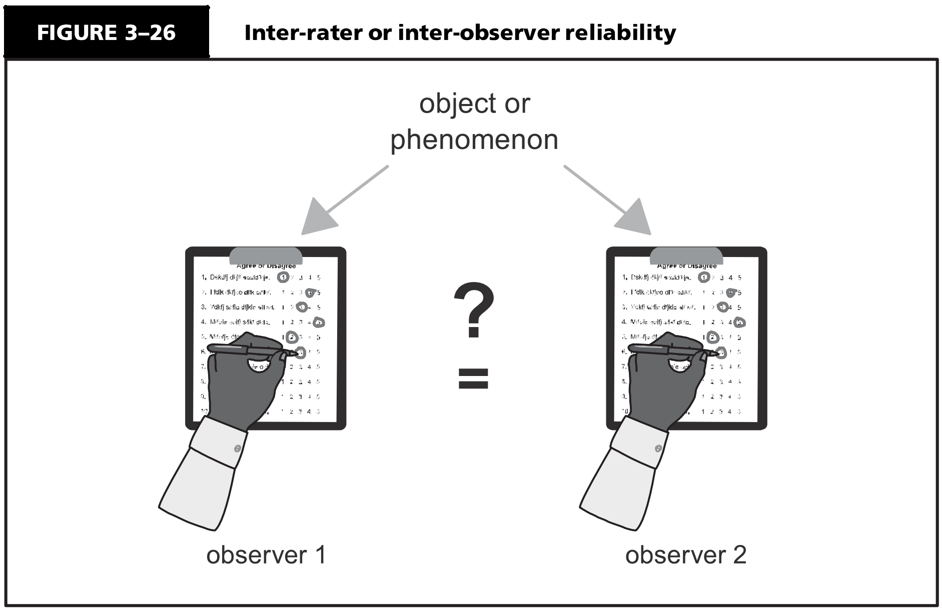
- **Inter-rater or inter-observer reliability** is used to assess the degree to which different raters/observers give consistent estimates of the same phenomenon.
- **Test-retest reliability** is used to assess the consistency of a measure from one time to another.
- **Parallel-forms reliability** is used to assess the consistency of the results of two tests constructed in the same way from the same content domain.
- **Internal consistency reliability** is used to assess the consistency of results across items within a test.

I'll discuss each of these in turn.

Inter-Rater or Inter-Observer Reliability Whenever you use humans as a part of your measurement procedure, you have to worry about whether the results you get are reliable or consistent. People are notorious for their inconsistency. We are easily distractible. We get tired of doing repetitive tasks. We daydream. We misinterpret.

So how do you determine whether two observers are being consistent in their observations? You probably should establish inter-rater reliability outside of the context of the measurement in your study. After all, if you use data from your study to establish reliability, and you find that reliability is low, you're kind of stuck. Probably it's best to do this as a side study or pilot study. If your study continues for a

FIGURE 3–26 **Inter-rater or inter-observer reliability**



long time, you may want to reestablish inter-rater reliability from time to time to ensure that your raters aren't changing.

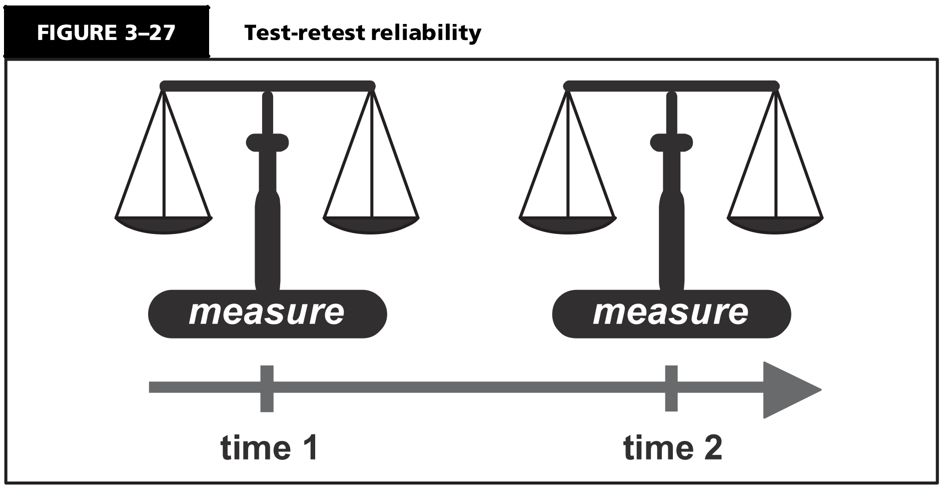
There are two major ways to actually estimate inter-rater reliability. If your measurement consists of categories—the raters are checking off which category each observation falls in—you can calculate the percent of agreement between the raters. For instance, let's say you had 100 observations that were being rated by two raters. For each observation, the rater could check one of three categories. Imagine that on 86 of the 100 observations the raters checked the same category. In this case, the percentage of agreement would be 86%. Okay, it's a crude measure, but it does give an idea of how much agreement exists, and it works no matter how many categories are used for each observation.

The other major way to estimate inter-rater reliability is appropriate when the measure is a continuous one. In such a case, all you need to do is calculate the correlation between the ratings of the two observers. For instance, they might be rating the overall level of activity in a classroom on a 1 to 7 scale. You could have them give their rating at regular time intervals (every 30 seconds). The correlation between these ratings would give you an estimate of the reliability or consistency between the raters (Figure 3–26).

You might think of this type of reliability as calibrating the observers. There are other things you could do to encourage reliability between observers, even if you don't estimate it. For instance, I used to work in a psychiatric unit where every morning a nurse had to do a ten-item rating of each patient on the unit. Of course, we couldn't count on the same nurse being present every day, so we had to find a way to ensure that all the nurses would give comparable ratings. The way we did it was to hold weekly calibration meetings where we would have all of the nurses' ratings for several patients and discuss why they chose the specific values they did. If there were disagreements, the nurses would discuss them and attempt to come up with rules for deciding when they would give a 3 or a 4 for a rating on a specific item. Although this was not an estimate of reliability, it probably went a long way toward improving the reliability between raters.

Test-Retest Reliability You estimate test-retest reliability when you administer the same test to the same (or a similar) sample on two different occasions (Figure 3–27). This approach assumes that there is no substantial change in the construct being measured between the two occasions. The amount of time allowed between measures is critical. You know that if you measure the same thing twice,

FIGURE 3–27 Test-retest reliability



the correlation between the two observations will depend in part on how much time elapses between the two measurement occasions. The shorter the time gap, the higher the correlation; the longer the time gap, the lower the correlation because the two observations are related over time; the closer in time you get, the more similar the factors that contribute to error. Since this correlation is the test-retest estimate of reliability, you can obtain considerably different estimates depending on the interval.

Parallel-Forms Reliability In parallel-forms reliability, you first have to create two parallel forms. One way to accomplish this is to start with a large set of questions that address the same construct and then randomly divide the questions into two sets. You administer both instruments to the same sample of people. The correlation between the two parallel forms is the estimate of reliability. One major problem with this approach is that you have to be able to generate lots of items that reflect the same construct, which is often no easy feat. Furthermore, this approach makes the assumption that the randomly divided halves are parallel or equivalent. Even by chance, this will sometimes not be the case. The parallel-forms approach is similar to the split-half reliability described later. The major difference is that parallel forms are constructed so that the two forms can be used independently of each other and considered equivalent measures. For instance, you might be concerned about a testing threat to internal validity. If you use Form A for the pretest and Form B for the posttest, you minimize that problem. It would 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 develop randomly split halves only for purposes of estimating reliability (Figure 3–28).

Internal-Consistency Reliability In internal-consistency reliability estimation, you use your 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 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 Interitem Correlation. The average interitem 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 Figure 3–29. For example, if you have six items, you will have fifteen different item pairings

FIGURE 3–28 **Parallel-forms reliability**

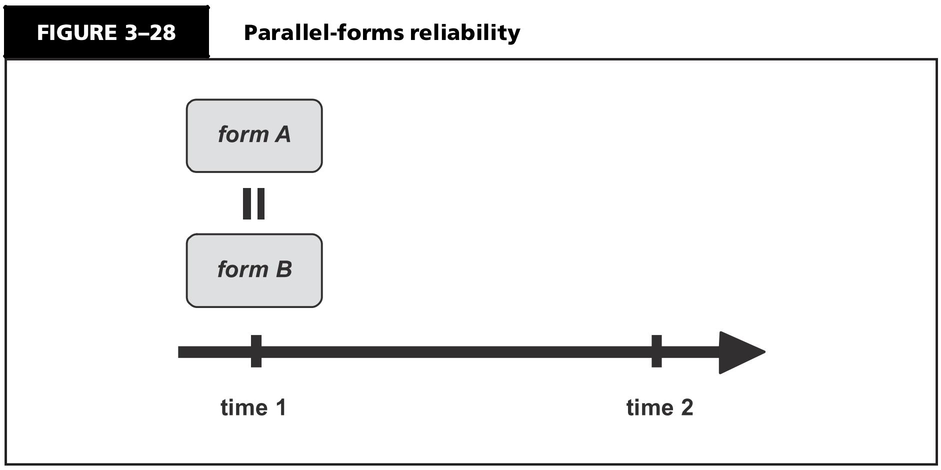
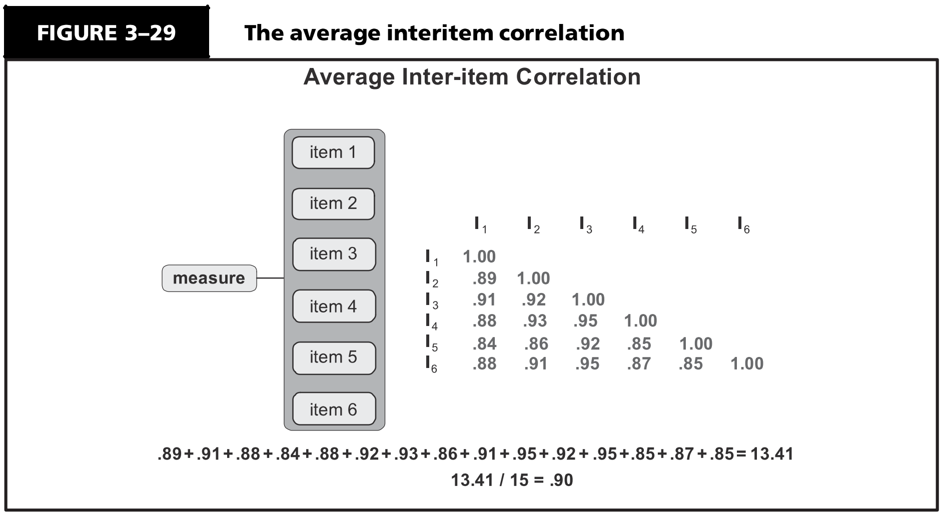


FIGURE 3–29 **The average interitem correlation**



	I_1	I_2	I_3	I_4	I_5	I_6
I_1	1.00					
I_2	.89	1.00				
I_3	.91	.92	1.00			
I_4	.88	.93	.95	1.00		
I_5	.84	.86	.92	.85	1.00	
I_6	.88	.91	.95	.87	.85	1.00

$$.89+.91+.88+.84+.88+.92+.93+.86+.91+.95+.92+.95+.85+.87+.85=13.41$$

$$13.41 / 15 = .90$$

(fifteen correlations). The average interitem correlation is simply the average or mean of all these correlations. In the example, you find an average interitem correlation of .90 with the individual correlations ranging from .84 to .95.

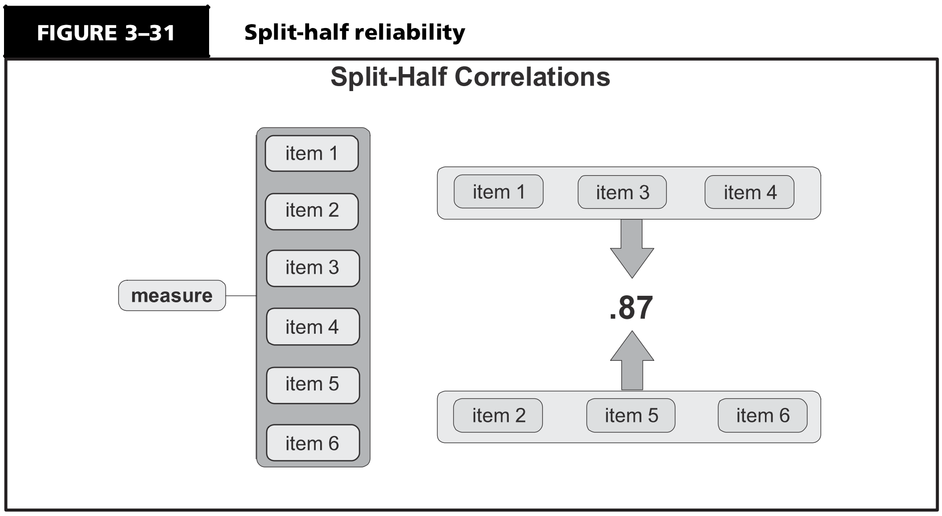
Average Item-Total Correlation. This approach also uses the interitem correlations. In addition, you compute a total score for the six items and use that as a seventh variable in the analysis. Figure 3–30 shows the six item-to-total correlations at the bottom of the correlation matrix. They range from .82 to .88 in this sample analysis, with the average of these at .85.

Split-Half Reliability. In split-half reliability, you randomly divide all items that purport to measure the same construct into two sets. You administer the entire instrument to a sample of people and calculate the total score for each randomly divided half. The split-half reliability estimate, as shown in Figure 3–31, is simply the correlation between these two total scores. In the example, it is .87.

If you think about this in practice, it might occur to you that (1) basically we are pretending we have two equivalent tests, both half as long as the "real" test, and (2) not all tests are the same length. So you might reasonably wonder about the relationship of test length to reliability. In general, longer tests are more reliable, so the split-half procedures systematically underestimate the reliability of a test. A formula has been developed to help us estimate the reliability of the full test based

FIGURE 3–30 **Average item-total correlation**

FIGURE 3–31 **Split-half reliability**



on the split-half estimate. This formula is called the *Spearman-Brown formula* and is shown in Figure 3–32.

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. **Cronbach's alpha** is mathematically equivalent to the average of all possible split-half estimates (although that's not how it's typically computed). Notice that when I say you compute all possible split-half estimates, I don't mean that each time you measure a new sample! That would take forever. Instead, you calculate all split-half estimates from the same sample. Because you measured your entire sample on each of the six items, all you have to do is have the computer analysis do the random subsets of items and compute the resulting correlations. Figure 3–33 shows several of the split-half estimates for our six-item example and lists them as SH with a subscript. Keep in mind that although Cronbach's alpha is equivalent to the average of all possible split-half correlations, you would never actually calculate it that way. Some clever mathematician (Cronbach, I presume!) figured out a way to get the mathematical equivalent a lot more quickly.

Cronbach's alpha
One specific method of estimating the reliability of a measure. Although not calculated in this manner, Cronbach's alpha can be thought of as analogous to the average of all possible split-half correlations.

Comparison of Reliability Estimators. Each of the reliability estimators has certain advantages and disadvantages. Inter-rater reliability is one of the best ways

FIGURE 3–32 **Spearman-Brown formula**

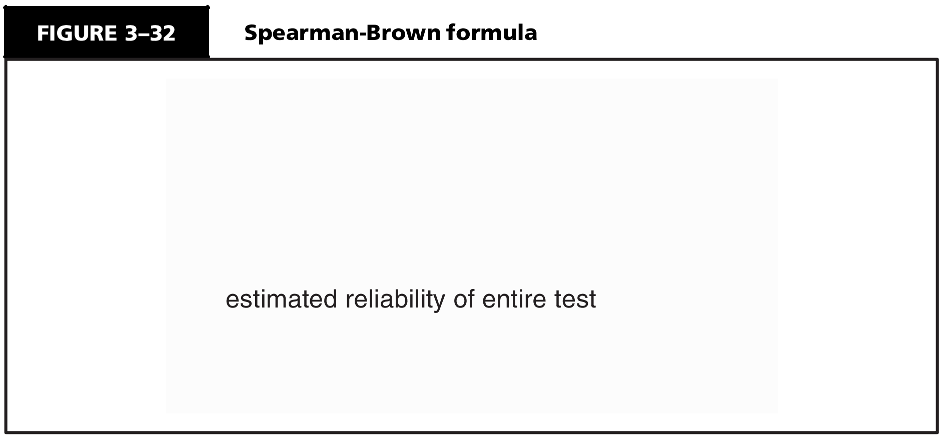
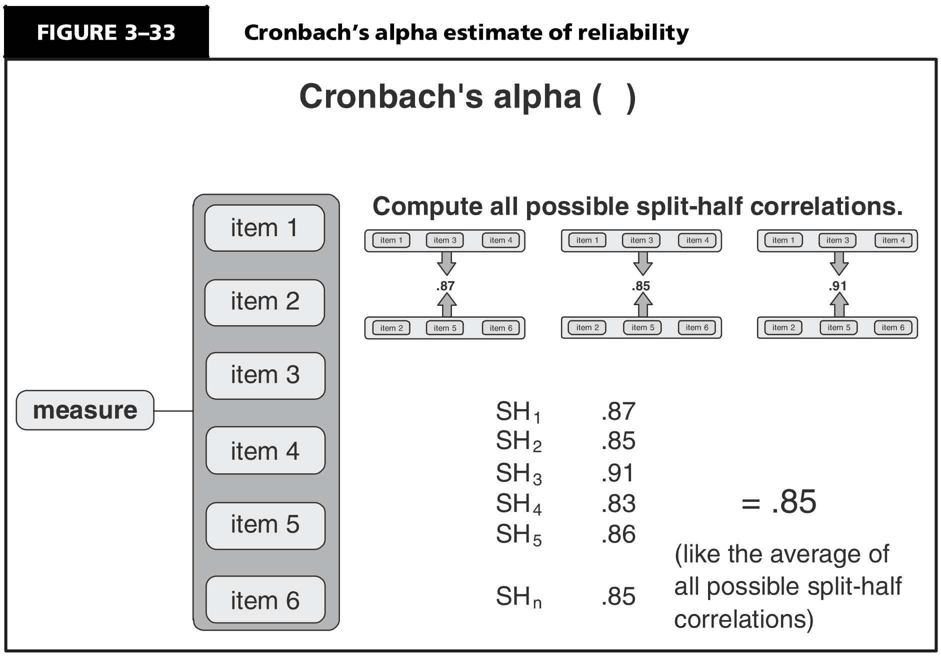


FIGURE 3–33 **Cronbach's alpha estimate of reliability**



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. To estimate test-retest reliability, you could have a single rater code the same videos on two different occasions. 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. You might use the test-retest approach when you have only a single rater and don't want to train any others. On the other hand, in some studies it is reasonable to do both to help establish the reliability of the raters or observers.

You use the parallel-forms estimator only in situations where you intend to use the two forms as alternate measures of the same thing. Both the parallel forms and all of the internal consistency estimators have one major constraint: you have to have lots of items designed to measure the same construct. This is relatively easy to achieve in certain contexts like achievement testing. (It's easy, for instance, to construct many similar addition problems for a math test.) However, for more complex or subjective constructs, this can be a real challenge. With lots of items, Cronbach's alpha tends to be the most frequently used estimate of internal consistency.

The test-retest estimator is especially feasible in most experimental and quasi-experimental designs that use a no-treatment control group. In these designs, you always have a control group that is measured on two occasions (pretest and posttest). The main problem with this approach is that you don't have any information about reliability until you collect the posttest and, if the reliability estimate is low, you're pretty much sunk.

Each of the reliability estimators gives a different value for reliability. In general, the test-retest and inter-rater reliability estimates will be lower in value than the parallel-forms and internal-consistency estimates because they involve measuring at different times or with different raters. Since reliability estimates are often used in statistical analyses of quasi-experimental designs (see Section 10-1, The Nonequivalent-Groups Design), the fact that different estimates can differ considerably makes the analysis even more complex.

3-2e Reliability and Validity

We often think of reliability and validity as separate ideas but, in fact, they're related to each other. Here, I want to show you two ways you can think about their relationship.

One of my favorite metaphors for the relationship between reliability and validity is that of the 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 (Figure 3–34).

Figure 3–34 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 estimate, 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 RobinHood 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.)

FIGURE 3–34 **The shooting-target metaphor for reliability and validity of measurement**

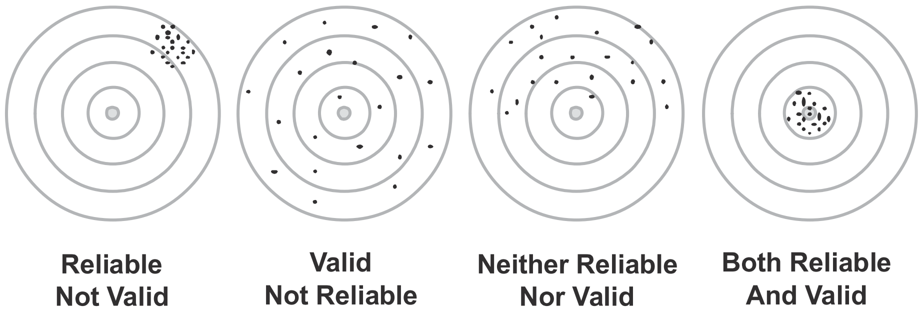
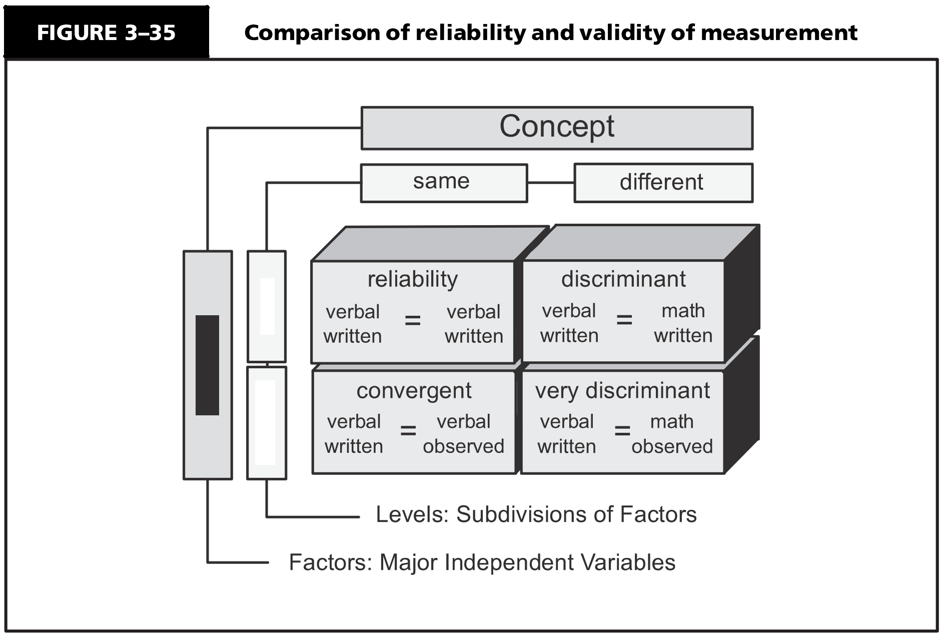


FIGURE 3–35 **Comparison of reliability and validity of measurement**


Concept
same
different
reliability
verbal written = verbal written
discriminant
verbal written = math written
convergent
verbal written = verbal observed
very discriminant
verbal written = math observed
Levels: Subdivisions of Factors
Factors: Major Independent Variables

Another way to think about the relationship between reliability and validity is shown in Figure 3–35, which contains a 2 2 table. The columns of the table indicate whether you are trying to measure the same or different concepts. The rows show whether you are using the same or different methods of measurement. Imagine that you have two concepts you would like to measure: student verbal and math ability. Furthermore, imagine that you can measure each of these in two ways. First, you can use a written, paper-and-pencil examination (much like the SAT or GRE examinations). Second, you can ask the students' classroom teachers to give you a rating of the students' ability based on their own classroom observation.

The first cell on the upper left shows the comparison of the verbal written test score with the verbal written test score; but how can you compare the same measure with itself? You could do this by estimating the reliability of the written test through a test-retest correlation, parallel forms, or an internal consistency measure (see Section 3-2d, Types of Reliability). What you are estimating in this cell is the reliability of the measure.

The cell on the lower left shows a comparison of the verbal written measure with the verbal teacher observation rating. Because you are trying to measure the same concept, you are looking at convergent validity (see Section 3-1a, Measurement Validity Types).

The cell on the upper left shows the comparison of the verbal written examination with the math written examination. Here, you are comparing two different concepts (verbal versus math) and so you would expect the relationship to be lower than a comparison of the same concept with itself (verbal versus verbal or math versus math). Thus, you are trying to discriminate between two concepts and this could be labeled discriminant validity.

Finally, you have the cell on the lower right. Here, you are comparing the verbal written examination with the math teacher observation rating. Like the cell on the upper right, you are also trying to compare two different concepts (verbal versus math), so this is also a discriminant validity estimate. However, here you are also trying to compare two different methods of measurement (written examination versus teacher observation rating). So, I'll call this very discriminant to indicate that you would expect the relationship in this cell to be even lower than in the one above it.

The four cells incorporate the different values that you examine in the MTMM approach to estimating construct validity described earlier in this chapter. When you look at reliability and validity in this way, you see that, rather than being distinct, they actually form a continuum. On one end is the situation where the concepts and methods of measurement are the same (reliability) and on the other is the situation where both concepts and methods of measurement are different (very discriminant validity).

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 (Figure 3–36). 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. Don't assume that higher values mean more of something or lower numbers signify less. Don't assume the value of 2 means that democrats are twice something that republicans are or that republicans are in first place or have the highest priority just because they have the value of 1. 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 a measure is nominal, you know that you would never average the data values or do a *t*-test on the data.

There are typically four levels of measurement that are defined (Figure 3–37):

- *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 3–36 **The level of measurement describes the relationship among the values associated with the attributes of a variable**

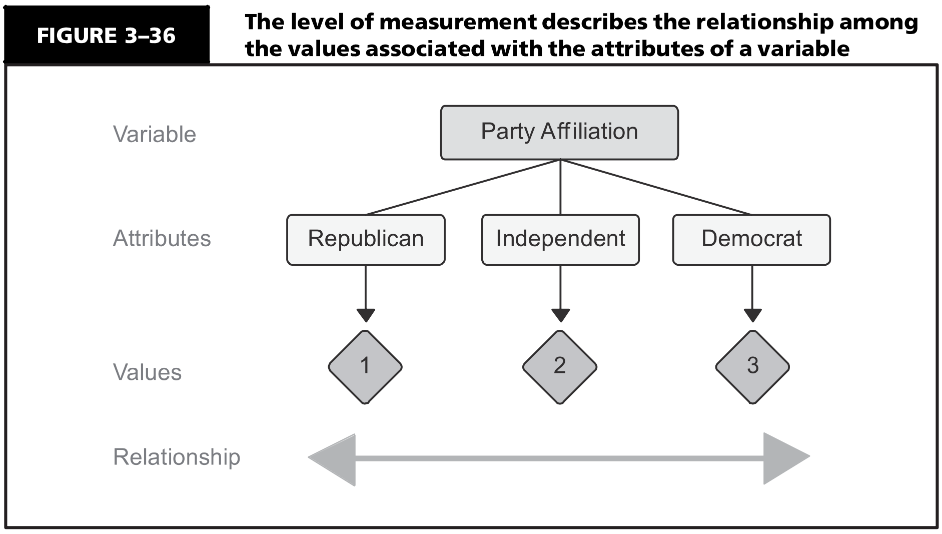
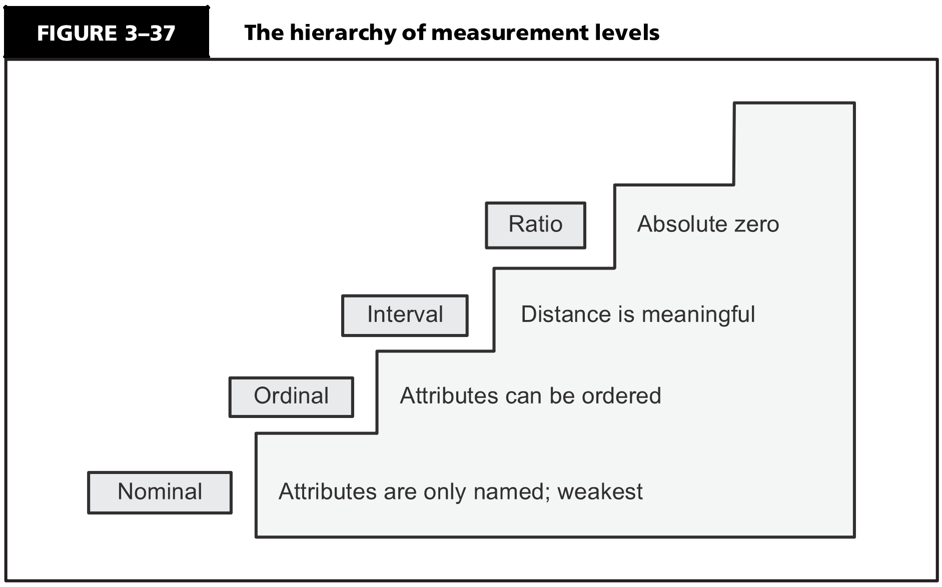


FIGURE 3–37 **The hierarchy of measurement levels**



- *Ordinal.* In ordinal measurement the attributes can be rank-ordered. Here, distances between attributes do not have any meaning. For example, on a survey you might code Educational Attainment as 0 = less than H.S.; 1 = some H.S.; 2 = H.S. degree, 3 = some college, 4 = college degree; 5 = post college. In this measure, higher numbers mean *more* education. Is distance from 0 to 1 the same as 3 to 4? Of course not. The interval between values is not interpretable in an ordinal measure.
- *Interval.* In interval measurement the distance between attributes *does* have meaning. For example, when we measure temperature (in Fahrenheit), the distance from 30 to 40 is same as distance from 70 to 80. The interval between values is interpretable. Because of this, it makes sense to compute an average of an interval variable, where it doesn't make sense to do so for ordinal scales. Note, however, that in interval measurement ratios don't make any sense; 80 degrees is not twice as hot as 40 degrees (although the attribute value is twice as large).
- *Ratio.* In ratio measurement there is always a meaningful absolute zero that is meaningful. This means that you can construct a meaningful fraction (or ratio) with a ratio variable. Weight is a ratio variable. In applied social research most *count* variables are ratio, for example, the number of clients in the past 6 months. Why? Because you can have zero clients and because it is meaningful to say, "We had twice as many clients in the past 6 months as we did in the previous 6 months."

It's important to recognize that there is a hierarchy implied in the level of measurement idea. At lower levels of measurement, assumptions tend to be less restrictive and data analyses tend to be less sensitive. At each level up the hierarchy, the current level includes all of the qualities of the one below it and adds something new. In general, it is desirable to have a higher level of measurement (such as interval or ratio) rather than a lower one (such as nominal or ordinal).

Summary

This chapter laid the foundation for the idea of measurement. Three broad topics were considered. First, *construct validity* refers to the degree to which you are measuring what you intended to measure. Construct validity is divided into translation validity (the degree to which you translated the construct well) and criterion-related validity (the degree to which your measure relates to or predicts other criteria as theoretically predicted). There is a long tradition of methods that attempt to assess construct validity that goes back to the original articulation of the nomological network, through the MTMM matrix and on to pattern-matching approaches. Second, *reliability* refers to the consistency or dependability of your measurement. Reliability is based on true score theory, which holds that any observation can be divided into two—a true score and error component. Reliability is defined as the ratio of the true score variance to the observed variance in a measure. Third, the level of a measure describes the relationship implicit among that measure's values and determines the type of statistical manipulations that are sensible. With these three ideas—construct validity, reliability, and level of measurement—as a foundation, you can now move on to some of the more practical and useful aspects of measurement in the next few chapters.

Login to the Online Edition of your text at www.atomicdog.com to find additional resources located in the Study Guide at the end of each chapter.