4

Validity

> What we know depends on how we know it. – Anonymous

4.1 Overview

According to the *Standards for Educational and Psychological Testing* (American Educational Research Association et al., 2014, p. 11), measurement validity is "the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests." We summarized reliability with three words: repeatability, consistency, and precision. A summary of validity in three words is accuracy, utility, and meaningfulness. Validity is tied to the *interpretation* of a measure's scores for the *proposed uses*, not (just) to the measure itself. The same set of scores can have different degrees of validity for different purposes. For instance, a measure's scores may have stronger validity for making a diagnostic decision than for making a prediction about future behavior. Thus, as *the Standards* indicate, it is incorrect to use the unqualified phrase "the validity of the measure" or "the measure is (in)valid," because these phrases do not specify which scores were used from the test, what the use is (e.g., predicting whether a person will succeed in a given job), and what interpretation was made of the test scores for this purpose.[1]

Below, we prepare the data to provide some validity-related examples throughout the rest of the chapter.

4.2 Getting Started

4.2.1 Load Libraries

```
library("petersenlab")
library("lavaan")
```

[1]For shorthand and to avoid repetition, we sometimes refer to the validity of the measure, but in such instances, we are actually referring to the validity of the interpretation of a measure's scores for a given use.

```
library("semPlot")
library("rockchalk")
library("semTools")
library("semPlot")
library("MASS")
library("psych")
library("simstandard")
library("DT")
library("MOTE")
library("here")
library("tidyverse")
library("tinytex")
library("knitr")
library("kableExtra")
library("rmarkdown")
library("bookdown")
```

4.2.2 Prepare Data

4.2.2.1 Simulate Data

```
sampleSize <- 1000

set.seed(52242)

means <- c(50, 100)
standardDeviations <- c(10, 15)

correlationMatrix <- matrix(.7, nrow = 2, ncol = 2)
diag(correlationMatrix) <- 1
rownames(correlationMatrix) <- colnames(correlationMatrix) <-
  c("predictor","criterion")

covarianceMatrix <- psych::cor2cov(
  correlationMatrix,
  sigma = standardDeviations)

mydataValidity <- as.data.frame(mvrnorm(
  n = sampleSize,
  mu = means,
  Sigma = covarianceMatrix,
  empirical = TRUE))

errorToAddToPredictor <- 3.20
errorToAddToCriterion <- 6.15

mydataValidity$predictorWithMeasurementErrorT1 <-
  mydataValidity$predictor +
```

```
  rnorm(n = sampleSize, mean = 0, sd = errorToAddToPredictor)

mydataValidity$predictorWithMeasurementErrorT2 <-
  mydataValidity$predictor +
  rnorm(n = sampleSize, mean = 0, sd = errorToAddToPredictor)

mydataValidity$criterionWithMeasurementErrorT1 <-
  mydataValidity$criterion +
  rnorm(n = sampleSize, mean = 0, sd = errorToAddToCriterion)

mydataValidity$criterionWithMeasurementErrorT2 <-
  mydataValidity$criterion +
  rnorm(n = sampleSize, mean = 0, sd = errorToAddToCriterion)

mydataValidity$oldpredictor <- mydataValidity$criterion +
  rnorm(n = sampleSize, mean = 0, sd = 7.5)

latentCorrelation <- .8
reliabilityPredictor <- .9
reliabilityCriterion <- .85

mydataValidity$predictorLatentSEM <- rnorm(sampleSize, 0 , 1)

mydataValidity$criterionLatentSEM <- latentCorrelation *
  mydataValidity$predictorLatentSEM + rnorm(
    sampleSize,
    0,
    sqrt(1 - latentCorrelation ^ 2))

mydataValidity$predictorObservedSEM <- reliabilityPredictor *
  mydataValidity$predictorLatentSEM + rnorm(
    sampleSize,
    0,
    sqrt(1 - reliabilityPredictor ^ 2))

mydataValidity$criterionObservedSEM <- reliabilityCriterion *
  mydataValidity$criterionLatentSEM + rnorm(
    sampleSize,
    0,
    sqrt(1 - reliabilityCriterion ^ 2))
```

4.2.2.2 Add Missing Data

Adding missing data to dataframes helps make examples more realistic to real-life data and helps you get in the habit of programming to account for missing data.

```
missingValuesPredictor <- sample(
  1:sampleSize,
  size = 50,
```

```
  replace = FALSE)
missingValuesCriterion <- sample(
  1:sampleSize,
  size = 50,
  replace = FALSE)

mydataValidity$predictor[
  missingValuesPredictor] <- NA
mydataValidity$predictorWithMeasurementErrorT1[
  missingValuesPredictor] <- NA
mydataValidity$predictorWithMeasurementErrorT2[
  missingValuesPredictor] <- NA
mydataValidity$predictorObservedSEM[
  missingValuesPredictor] <- NA

mydataValidity$criterion[
  missingValuesCriterion] <- NA
mydataValidity$criterionWithMeasurementErrorT1[
  missingValuesCriterion] <- NA
mydataValidity$criterionWithMeasurementErrorT2[
  missingValuesCriterion] <- NA
mydataValidity$criterionObservedSEM[
  missingValuesCriterion] <- NA

mydataValidity$oldpredictor[
  missingValuesPredictor] <- NA
```

4.3 Types of Validity

Like reliability, validity is not one thing. There are many types of validity. In this book, we discuss the following types of validity:

- face validity
- content validity
- criterion-related validity
- concurrent validity
- predictive validity
- construct validity
- convergent validity
- discriminant (divergent) validity
- incremental validity
- treatment utility of assessment
- discriminative validity
- elaborative validity
- consequential validity
- representational validity
- factorial (structural) validity

- ecological validity
- process-focused validity
- diagnostic validity
- social validity
- cultural validity
- internal validity
- external validity
- (statistical) conclusion validity

We arrange these types of validity into two broader categories: measurement validity and research design validity.

4.3.1 Measurement Validity

Aspects of *measurement validity* involve the validity of a particular measure, or more specifically, the validity of interpretations of scores from that measure for the proposed uses. Aspects of measurement validity include:

- face validity
- content validity
- criterion-related validity
- concurrent validity
- predictive validity
- construct validity
- convergent validity
- discriminant (divergent) validity
- incremental validity
- treatment utility of assessment
- discriminative validity
- elaborative validity
- consequential validity
- representational validity
- factorial (structural) validity
- ecological validity
- process-focused validity
- diagnostic validity
- social validity
- cultural validity

4.3.1.1 Face Validity

The interpretation of a measure's scores has *face validity* (for a given construct and a given use) if a typical person—a nonexpert—who looks at the content of each item will believe that the item belongs in the scale for this construct and for this use. The measure, and each item, looks "on its face" like it assesses the target construct. There are several advantages of a measure having face validity. First, outside groups will be less likely to be critical of the measure because it is intuitive. Second, use of a face valid measure is rarely objected to on ethical and bias charges for selling the measure to the public or clinicians. Third, face validity can be helpful for dissemination because more people may be receptive to it.

However, face validity also has important disadvantages. First, judgments of face validity are not based on theory. Second, face validity is based on subjective judgment, which can be inaccurate. Third, these subjective judgments are made by laypeople whose judgments may

be inaccurate because of biases and lack of awareness of scientific knowledge. Fourth, a face valid measure may be too simple because anybody can understand the questions and what the questions are intended to assess, so (presumably) respondents can easily fake responses to achieve their goals. Faking of responses may be more of a concern in situations when there is an incentive for the respondent to achieve a particular outcome (e.g., be deemed competent to stand trial, be judged competent to obtain a job or custody of child, be judged to have a disorder to receive accommodations or disability benefits).

It is disputed whether having face validity is good or bad. Whether face validity is important to a given measure depends on the construct that is intended to be assessed, the context in which the assessment will occur, who will be paying for and/or administering the assessment, whether the respondents have incentives to achieve particular scores, and the goals of the assessment (i.e., how the assessment will be used). There is also controversy about whether face validity is a true form of validity; many researchers have argued that it is not a true psychometric form of validity, because the *appearance* of validity is not validity (Royal, 2016).

4.3.1.2 Content Validity

Content validity involves a judgment about whether or not the content (items) of the measure theoretically matches the construct that is intended to be assessed—that is, whether the operationalization accurately reflects the construct. Content validity is developed based on items generated and selected by experts of the construct and based on the subjective determination that the measure adequately assesses and covers the construct of interest. Content validity differs from face validity in that, for face validity, a *layperson* determines whether or not the measure seems to assess the construct of interest. By contrast, for content validity, an *expert* determines whether or not the measure adheres to the construct of interest.

For a measure to have content validity, its items should span the breadth of the construct. For instance, the construct of depression has many facets, such as sleep disturbances, weight/appetite changes, low mood, suicidality, etc., as depicted in Figure 4.1.

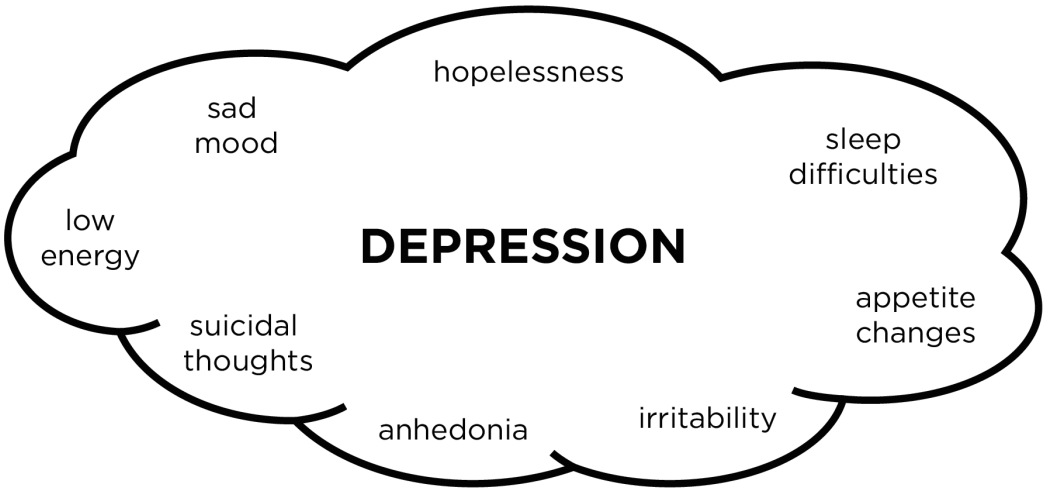

FIGURE 4.1 Content Facets of the Construct of Depression.

For a measure to have content validity, there should be no *gaps*—facets of the construct that are not assessed by the measure—and there should be no *intrusions*—facets of different

constructs that are assessed by the measure. Consider the construct of depression. If theory states the construct includes various facets such as sadness, loss of interest in activities, sleep disturbances, lack of energy, weight/appetite change, and suicidal thoughts, then a content-valid measure should assess all of these facets. If the measure does not assess sleep disturbances (a gap), the measure would lack content validity. If the measure assessed facets of other constructs, such as impulsivity (an intrusion), the measure would lack content validity.

With content validity, it is important to consider the population of interest. The same construct may look different in different populations and may require different content to assess it. For instance, it is important to consider the cultural relativity of constructs. The content of a construct may depend on the culture, such as in the case of culture-bound syndromes. Culture-bound syndromes are syndromes that are limited to particular cultures. An example of a culture-bound syndrome among Korean women is *hwa-byung*, which is the feeling of an uncomfortable abdominal mass in response to emotional distress. Another important dimension to consider is development. Constructs can manifest differently at different points in development, known as heterotypic continuity, which is discussed in Section 22.8 of Chapter 22 on repeated assessments across time. When considering the different dimensions of your population, it can be helpful to remember the acronym *ADDRESSING*, which is described in Section 24.2.1 of Chapter 24 on cultural and individual diversity.

However, like face validity, content validity is based on subjective judgment, which can be inaccurate.

4.3.1.3 Criterion-Related Validity

Criterion-related validity examines whether a measure behaves the way it should given your theory of the construct. This is quantified by the correlation between a measure's scores and some (hopefully universally accepted) criterion we select. For instance, a criterion could be a diagnosis, a child's achievement in school, an employee's performance in a job, etc.

Below, we provide an example of criterion-related validity by examining the Pearson correlation between a predictor and a criterion.

```
cor.test(x = mydataValidity$predictor, y = mydataValidity$criterion)
```

```
    Pearson's product-moment correlation

data:  mydataValidity$predictor and mydataValidity$criterion
t = 30, df = 900, p-value <2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6738 0.7391
sample estimates:
   cor
0.7079
```

In this case, the estimate of criterion-related validity is $r = .71$. There are two types of criterion-related validity: concurrent validity and predictive validity.

4.3.1.3.1 Concurrent validity

Concurrent validity considers the *concurrent* association between the chosen measure and the criterion. That is, both the measure and the criterion are assessed at the same point in time. An example of concurrent validity would be examining self-report of court involvement in relation to current court records.

4.3.1.3.2 Predictive validity

Predictive validity considers the association between the chosen measure and the criterion at a *later* time point. An example of predictive validity would be examining the predictive association between children's scores on an academic achievement test in first grade and their eventual academic outcomes years later.

4.3.1.3.3 Empiricism and theory

Criterion-related validity arose out of a movement known as *radical operationalism*. Radical operationalism was a pushback against psychoanalysis. Psychoanalysis focused on grand theoretical accounts for how constructs relate. The goal of radical operationalism was to clarify concepts from a behavioristic perspective to allow predicting and changing behavior more successfully. An "operation" in radical operationalism refers to a fully described measurement.

Proponents of radical operationalism argued that all constructs in psychology that could not be operationally defined should be excluded from the field as "nonscientific." They asserted that operations should be well-defined enough to be able to replicate the findings. So, constructs had to be defined precisely according to this perspective, but how precisely? You could go on forever trying to more precisely describe a behavior in terms of its form, frequency, duration, intensity, situation, antecedents, consequences, biological substrates, etc. So, radical operationalists asserted that we should use theory of the construct to determine what is essential and what is not.

Radical operationalism was also related to *radical behavioralism*, which was espoused by B.F. Skinner. Skinner famously used a box (the "Skinner Box") to more directly control, describe, and assess behaviors. Skinner noted the major role that the environment played in influencing behavior. Skinner proposed a theory of implicit learning about a behavior or stimulus based on its consequences, known as operant conditioning. According to operant conditioning, something that increases the frequency of a given behavior is called a reinforcer (e.g., praise), and something that decreases the frequency of a behavior is called a punisher (e.g., loss of a privilege). Through this work, Skinner came to view everything an organism does (e.g., action, thought, feeling) as a behavior.

Related to these historical perspectives was a perspective known as *dustbowl empiricism*. Dustbowl empiricism focused on the empirical connections between things—how things were associated using data. It was a completely atheoretical perspective in which interpretation was entirely data driven. An example of dustbowl empiricism is the approach that was used to develop the first version of the Minnesota Multiphasic Personality Inventory (MMPI). The MMPI was developed using an approach known as empirical-criterion keying, where items were selected for the scale for no reason other than the items demonstrate an association with the criterion. That is, an item was selected if it showed a strong ability to discriminate (differentiate) between clinical and control groups. Using this method with hundreds of items (and thousands of inter-item correlations), the MMPI developed 10 clinical scales, which involved operational rules based on previously collected empirical evidence.

But what do you know with this abundance of correlations? You can use data reduction methods to reduce the many variables, based on their inter-correlations, down to a more parsimonious set of factors. But how do you name each factor, which is composed of many items? The developers originally numbered the MMPI clinical scales from 1 to 10. But numbered scales are not useful for other people, so the factors were eventually given labels (e.g., Paranoia). And if a client received an elevated score on a factor, many people would label the clients as _____ [the name of the factor], such as "paranoid." The MMPI is discussed in further detail in Chapter 17 on objective personality testing.

The idea of dustbowl empiricism was to develop a strong empirical base that would provide a strong foundation to help build up to a broad understanding that was integrated, coherent, and systematic. However, this process was unclear when there was only a table of correlations. Radical operationalists were opposed to content validity because it allows intrusion of our flawed thinking. According to operationalists, there are no experts. According to this perspective, the content does not matter; we just need enough data to bootstrap ourselves to a better understanding of the constructs.

Although the atheoretical approach can perform reasonably well, it can be improved by making better use of theory. An empirical result (e.g., a correlation) might not necessarily have a lot of meaning associated with it. As the maxim goes, correlation does not imply causation.

4.3.1.3.3.1 Correlation does not imply causation

Just because X is associated with Y does not mean that X causes Y. Consider that you find an association between variables X and Y, consistent with your hypothesis, as depicted in Figure 4.2.

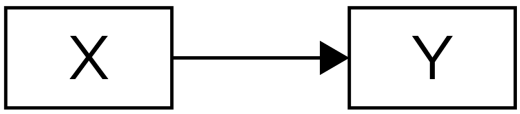

FIGURE 4.2 Hypothesized Causal Effect Based on an Observed Association Between X and Y, Such That X Causes Y.

There are three primary reasons that an observed association between X and Y does not necessarily mean that X causes Y. First, the association could reflect the opposite direction of effect, where Y actually causes X, as depicted in Figure 4.3.

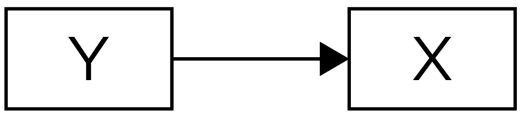

FIGURE 4.3 Reverse (Opposite) Direction of Effect from the Hypothesized Effect, Where Y Causes X.

Second, the association could reflect the influence of a third variable. If a third variable is a common cause of each and accounts for their association, it is a *confound*. An observed association between X and Y could reflect a confound—i.e., a cause (Z) that influences both X and Y, which explains why X and Y are correlated even though they are not causally related. A third variable confound that is a common cause of both X and Y is depicted in Figure 4.4.

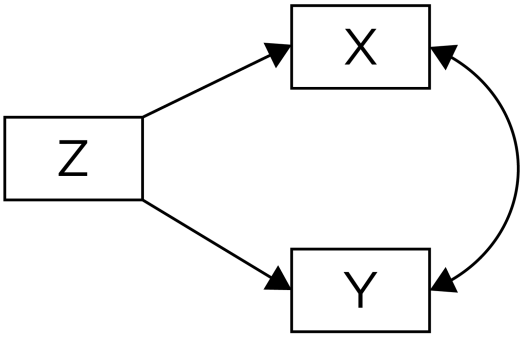

FIGURE 4.4 Confounded Association Between X and Y due to a Common Cause, Z.

Third, the association might be spurious. It might just reflect random variation (i.e., chance), and that when tested on an independent sample, what appeared as an association may not hold when testing whether the association generalizes.

However, even if the association between X and Y reflects a causal effect and that X causes Y, it does not necessarily mean that the effect is clinically actionable or useful. An association may reflect a static or unmodifiable predictor that is not practically useful as a treatment target.

4.3.1.3.3.2 Understanding the causal system

As Silver (2012) notes, "The numbers have no way of speaking for themselves. We speak for them. We imbue them with meaning" (p. 9). If we *understand* the variables in the system and how they influence each other, we can predict things more accurately than predicting for the sake of predicting. For instance, we have made great strides in the last decades when it comes to more accurate weather forecasts, including extreme weather events like hurricanes. These great strides have more to do with a better causal understanding of the weather system and the ability to conduct simulations of the atmosphere than merely because of big data (Silver, 2012). By contrast, other events are still incredibly difficult to predict, including earthquakes, in large part because we do not have a strong understanding of the system (and because we do not have ways of precisely measuring those causes because they occur at a depth below which we are realistically able to drill) (Silver, 2012).

4.3.1.3.3.3 Model over-fitting

Statistical models applied to big data (i.e., lots of variables and lots of samples) can *over-fit* the data, which means that the statistical model accounts for error variance (an overly specific prediction), which will not generalize to future samples. So, even though an over-fitting statistical model appears to be accurate, it is not actually that accurate—it will predict new data less accurately than how accurately it accounts for the data with which the model was built.

In Figure 4.5, the black line represents the true distribution of the data, and the gray line is an over-fitting model.

4.3.1.3.3.4 Criterion contamination

An important issue in predictive validity is the criterion problem—finding the right criterion. It is important to avoid *criterion contamination*, which is artificial commonality between

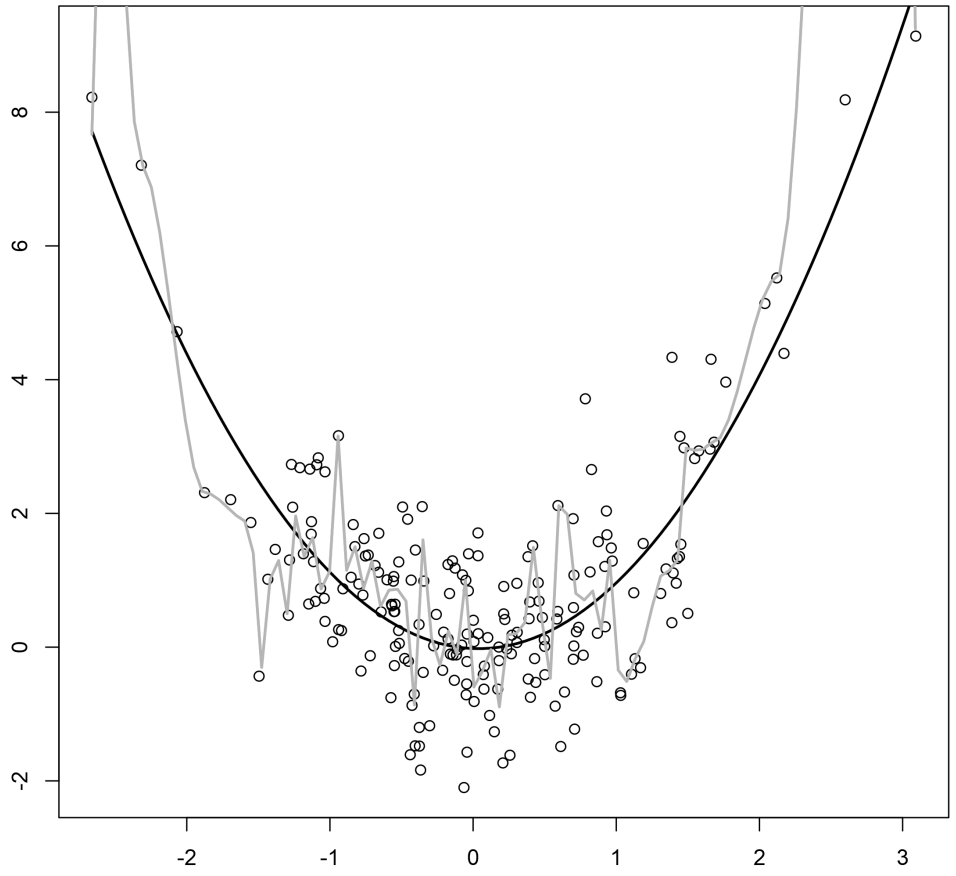

FIGURE 4.5 Over-fitting Model in Gray Relative to the True Distribution of the Data in Black.

the measure and the criterion. The criterion is not always a well-measured clear criterion to predict (like predicting death in the medical field). And you may not have access to a predictive criterion until a long time from now. So, what researchers often do is adopt intermediate assessments, which are not actually what they are interested in, but it is related to the criterion of interest, and it is in a window of time that allows for some meaningful prediction. For instance, intermediate graduate school markers of whether a graduate student will go on to have a successful career could include their grades in graduate school, whether they completed the dissertation, their performance in comprehensive/qualifying exams, etc. However, these intermediate assessments do not always indicate whether or not a student will go on to have a successful career (i.e., they are not always correlated with the real criterion of interest).

4.3.1.3.3.5 Using theory as a guide

So, empiricism is often not enough. It is important to use theory to guide selection of an intermediate criterion that will relate to the real criterion of interest. In psychology, even our long-term criteria are not well defined relative to other sciences. In clinical psychology, for example, we are often predicting a diagnosis, which is not that much more valid compared to our measure/predictor.

At the same time, in psychology, our theories of the causal processes that influence outcomes are not yet very strong. Indeed, we have misgivings calling them theories because they do not meet the traditional scientific standard for a theory. A scientific theory is an explanation of the natural world that is testable and falsifiable, and that has withstood rigorous scientific testing and scrutiny. In psychology, our "theories" are more like conceptual frameworks. And these conceptual frameworks are often vague, do not make specific predictions of effects *and* noneffects, and do not hold up consistently when rigorously tested. As described by Meehl (1978):

> I consider it unnecessary to persuade you that most so-called "theories" in the soft areas of psychology (clinical, counseling, social, personality, community, and school psychology) are scientifically unimpressive and technologically worthless ... Perhaps the easiest way to convince yourself is by scanning the literature of soft psychology over the last 30 years and noticing what happens to theories. Most of them suffer the fate that General MacArthur ascribed to old generals—They never die, they just slowly fade away. In the developed sciences, theories tend either to become widely accepted and built into the larger edifice of well-tested human knowledge or else they suffer destruction in the face of recalcitrant facts and are abandoned, perhaps regretfully as a "nice try." But in fields like personology and social psychology, this seems not to happen. There is a period of enthusiasm about a new theory, a period of attempted application to several fact domains, a period of disillusionment as the negative data come in, a growing bafflement about inconsistent and unreplicable empirical results, multiple resort to ad hoc excuses, and then finally people just sort of lose interest in the thing and pursue other endeavors. (pp. 806–807).

Even if we had strong theoretical understanding of the causal system that influences behavior, we would likely still have difficulty making accurate predictions because the field has largely relied on relatively crude instruments. According to one philosophical perspective known as LaPlace's demon, if we were able to know the exact conditions of everything in the universe, we would be able to know how the conditions would be in the future. This is an example of scientific determinism, where if you know the initial conditions, you also know the future. Other perspectives, such as quantum mechanics and chaos theory, would say that, even if we knew the initial conditions with 100% certainty, there would still be uncertainty in our understanding of the future. But assume, for a moment, that LaPlace's demon is true. The challenge in psychology is that we have a relatively poor understanding of the initial conditions of the universe. Thus, our predictions would necessarily be probabilistic, similar to weather forecasts. Despite having a strong understanding of how weather systems behave,

we have imperfect understanding of the initial conditions (e.g., the position and movement of all molecules) (Silver, 2012).

4.3.1.3.3.6 Psychoanalysis versus empiricism

We can consider the difference between psychoanalysts and empiricists in cultural references. If we consider a scene from *The Hitchhiker's Guide to the Galaxy* where a man extrapolates a simulation of the universe from a single piece of cake, we find similarities to how psychoanalysts connect everything to everything else through grand theories. Psychoanalysts try to reconstruct the entire universe from a sparse bit of data with supposedly strong theoretical understanding (when, in reality, our theories are not so strong). Their "theories" make grand conceptual claims.

Let us contrast psychoanalysts with empiricists/radical operationalism. Figure 4.6 presents a depiction of empiricism. Empiricists evaluate how an observed predictor relates to an observed outcome. The rectangles in the figure represent entities that can be observed. For instance, an empiricist might examine the extent to which a person's blood pressure is related to the number of words they speak per minute.

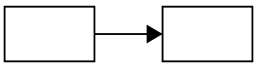

FIGURE 4.6 Conceptual Depiction of Empiricism.

Contrast the empirical approach with psychoanalysis, as depicted in Figure 4.7.

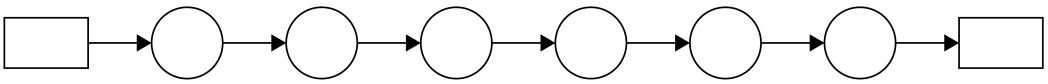

FIGURE 4.7 Conceptual Depiction of Psychoanalysis.

Circles represent unobserved, latent entities. For instance a psychoanalyst may make a conceptual claim that one observed variable influences another observed variable through a complex chain of intervening processes that are unobservable.

There is a classic and hilarious Monty Python video[2] that is an absurd example akin to radical operationalism taken to the extreme. In the video, the researchers make lots of measurements and predictions. The researchers identify interspecies differences in intelligence where humans show better performance on the English-based intelligence test (who got an IQ score of 100) than the penguins, who got an IQ score of 2. But the penguins did not speak English and were unable to provide answers to the English-based intelligence test. So, the researchers also assessed a group of non-English speaking humans as an attempt to control for language ability. They found that the penguins' scores were equal to the scores of the non-English speaking humans. They argued that, based on their smaller brain and equal IQ when controlling for language ability, that penguins are smarter than humans. However, the researchers clearly made mistakes about confounding variables. And inclusion of a control group of non-English speaking humans does not solve the problem of validity or bias; it just obscures the problem. In summary, radical operationalism provides rich lower-level information, but lacks the broader picture. So, it seems, that we need both theory *and* empiricism. Theory and empiricism can—and should—inform each other.

[2] https://osf.io/gc79d

4.3.1.4 Construct Validity

Construct validity is the extent to which a measure accurately assesses a target construct (Cronbach & Meehl, 1955). Construct validity is not quantified by a single index but rather consists of a "web of evidence" (the totality of evidence), which reflect the sum of inferences from multiple aspects of validity. That is, construct validity is the extent to which a measure assesses a target construct. Construct validity deals with the association between a measure and an unobservable criterion, i.e., a latent construct. By contrast, criterion-related validity deals with an observable criterion.

Construct validity encompasses all other types of measurement validity (e.g., content and criterion-related validity), in addition to

- scores on the measure show homogeneity—i.e., scores on the measure assess a single construct
- scores on the measure show theoretically expected developmental changes
- scores on the measure show theoretically expected group differences
- scores on the measure show theoretically expected intervention effects
- establishing the *nomological network* of a construct

4.3.1.4.1 Nomological network

A *nomological network* is the interlocking system of laws that constitute a theory. It describes how the concepts (constructs) of interest are causally linked, including their observable manifestations and the causal relations among and between them. An example of a nomological network is depicted in Figure 4.8.

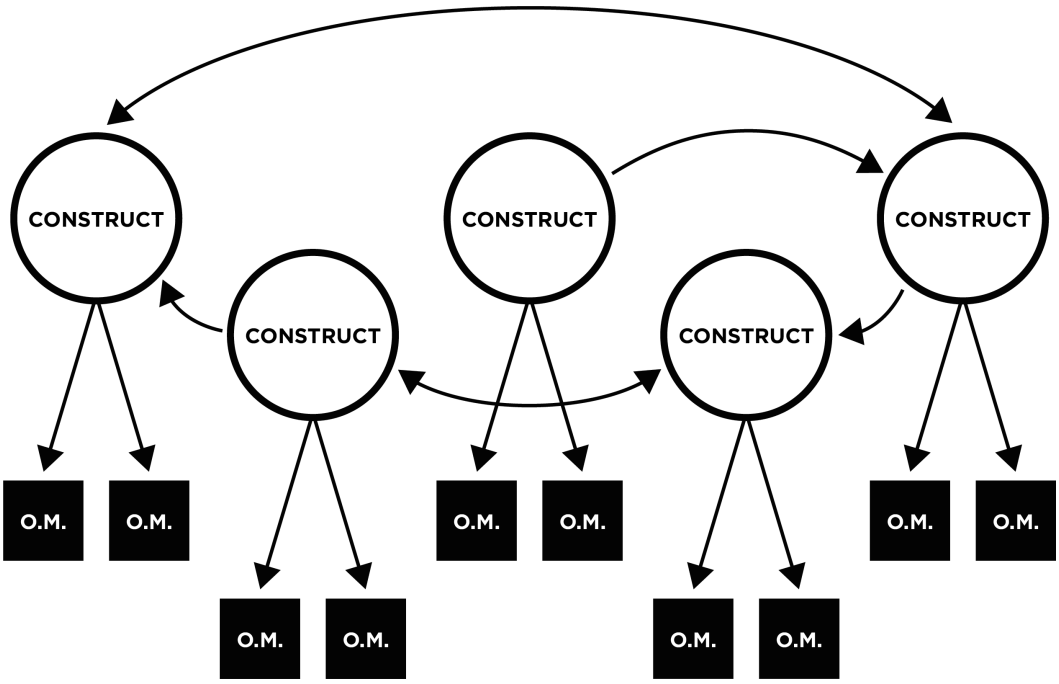

FIGURE 4.8 Example of a Nomological Network. O.M. = Observable Manifestation.

With construct validity, we can judge the quality of a measure by how well or how sensibly

it fits in a nomological network. Latent constructs and observed measures improve each other step by step. But there is no established way to evaluate the process.

Historically, construct validity became a way for some researchers to skip out on other types of validity. People found a correlation between a measure's scores and some group membership and argued, therefore, that the measure has construct validity because there was a theoretically expected correlation between some measure and _____ (insert whatever measure). People started finding a correlation of a measure with other measures, asserting that it provides evidence of construct validity, and saying "that's my nomological network." But a theoretically expected association is not enough! For example, consider a measure of how quickly someone can count backward by seven. Performance is impaired in those with schizophrenia, anxiety (due to greater distractibility), and depression (due to concentration difficulties and cognitive slowing). Therefore, it is a low-quality claim that counting backward is part of the phenomenology of these disorders because it lacks differential deficit or discriminant validity (D. T. Campbell & Fiske, 1959). This is related to the "glop problem," which asserts that every bad thing is associated with every other bad thing—there is high comorbidity. Therefore, researchers needed some way to distinguish method variance from construct variance. This led to the development of the multitrait-multimethod matrix.

4.3.1.4.2 Multitrait-multimethod matrix (MTMM)

The *multitrait-multimethod matrix* (MTMM), as proposed by Campbell and Fiske (1959), is a concrete way to evaluate the validity of a measure. The MTMM allows you to split the variance of measures' scores into variance that is due to the method (i.e., method variance or method bias) and variance that is due to the construct (trait) of interest (i.e., construct variance). To create an MTMM, you need at least two methods and at least two constructs. For example, an MTMM could include self-report and observation of depression and introversion. You would then examine the correlations across combinations of construct and method.

For an example of an MTMM, see Figure 4.9. Several aspects of psychometrics can be evaluated with an MTMM, including reliability, convergent validity, and discriminant validity. The reliability diagonal of an MTMM is the correlation of a variable with itself, i.e., the test–retest reliability or monotrait-monomethod correlations. The reliability coefficients should be the highest values in the matrix because each measure should be more correlated with itself than with anything else.

A multitrait-multimethod matrix can be organized by method and then by construct or vice versa. An MTMM organized by method then by construct is depicted in Figure 4.10.

An MTMM organized by construct then by method is depicted in Figure 4.11.

4.3.1.4.2.1 Convergent validity

Convergent validity is the extent to which a measure is associated with other measures of the same target construct. In an MTMM, convergent validity evaluates whether measures targeting the same construct, but using different methods, converge upon the same construct. These are observed in the validity diagonals, also known as the convergent correlations or the monotrait-heteromethod correlations. For strong convergent validity, we would expect the values in the validity diagonals to be significant and high-ish.

A Synthetic Multitrait-Multimethod Matrix

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

Note.—The validity diagonals are the three sets of italicized values. The reliability diagonals are the three sets of values in parentheses. Each heterotrait-monomethod triangle is enclosed by a solid line. Each heterotrait-heteromethod triangle is enclosed by a broken line.

FIGURE 4.9 Multitrait-Multimethod Matrix. (Figure reprinted from Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*, 81-105. https://doi.org/10.1037/h0046016. The content is in the public domain.)

4.3.1.4.2.2 Discriminant (divergent) validity

You can also evaluate the *discriminant validity*, also called divergent validity, of a measure in the context of an MTMM. Discriminant validity is the extent to which a measure is *not* associated (or less strongly associated) with measures of different constructs that are not theoretically expected to be related (compared to associations with measures of the same construct). In an MTMM, discriminant validity determines the extent to which a measure does not correlate with measures that share a method but assess different constructs. For strong discriminant validity, we would expect the discriminant correlations (heterotrait monomethod) to be low.

According to Campbell and Fiske (1959), discriminant validity of measures can be established when three criteria are met:

1. The convergent (monotrait-heteromethod) correlations are stronger than heterotrait-heteromethod correlations. This provides the weakest evidence for discriminant validity. That is, the convergent correlations are higher than the values in the same column or row in the same heteromethod block. This can be evaluated using the heterotrait-monotrait ratio (described below).
2. The convergent (monotrait-heteromethod) correlations are stronger than discriminant correlations (monomethod-heterotrait). This provides stronger evidence of discriminant validity.

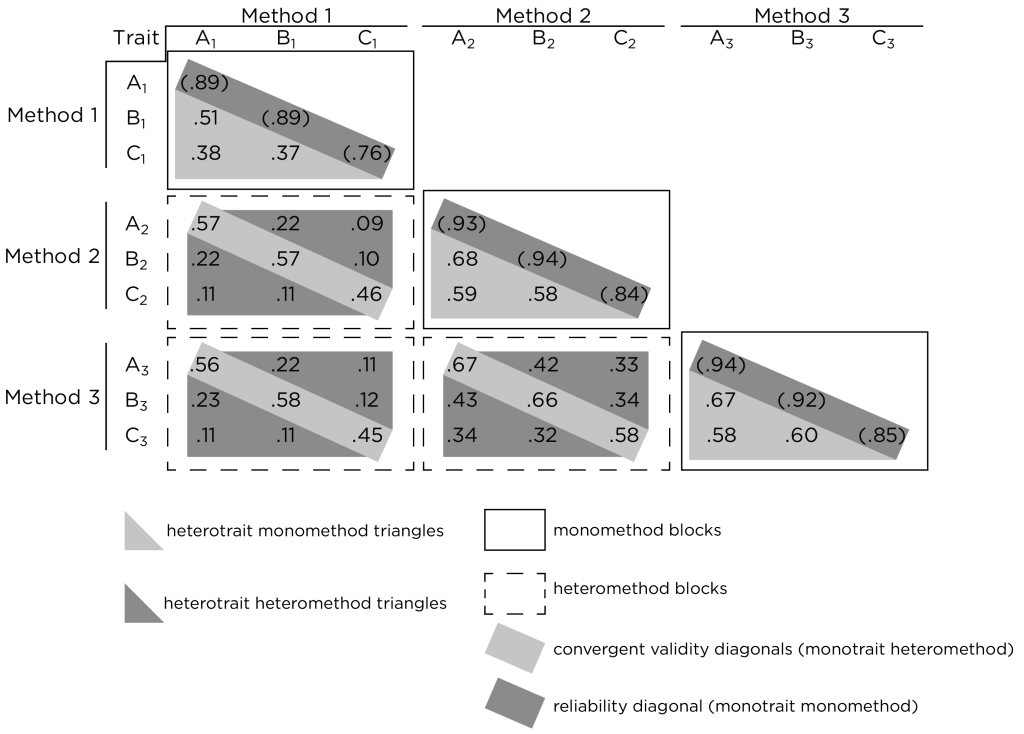

FIGURE 4.10 Multitrait-Multimethod Matrix Organized by Method Then by Construct.

3. The patterns of intercorrelations between constructs are the same, regardless of which measurement method is used. That is, the pattern of inter-trait associations is the same in all triangles. For example, if extraversion and anxiety are moderately correlated with each other but uncorrelated with achievement, we would expect this pattern of interrelations between constructs would hold, regardless of which method was used to assess the construct.

You can estimate a measure's degree of discriminant validity based on the *heterotrait-monotrait ratio* [HTMT; Henseler et al. (2015); Roemer et al. (2021)]. HTMT is the average of the heterotrait-heteromethod correlations (i.e., the correlations of measures from different measurement methods that assess different constructs), relative to the average of the monotrait-heteromethod correlations (i.e., the correlations of measures from different measurement methods that assess the same construct). As described here (http://www.henseler.com/htmt.html; archived at https://perma.cc/A9DU-WZWQ) based on evidence from Voorhees et al. (2016),

---

> If the HTMT is clearly smaller than one, discriminant validity can be regarded as established. In many practical situations, a threshold of 0.85 reliably distinguishes between those pairs of latent variables that are discriminant valid and those that are not.

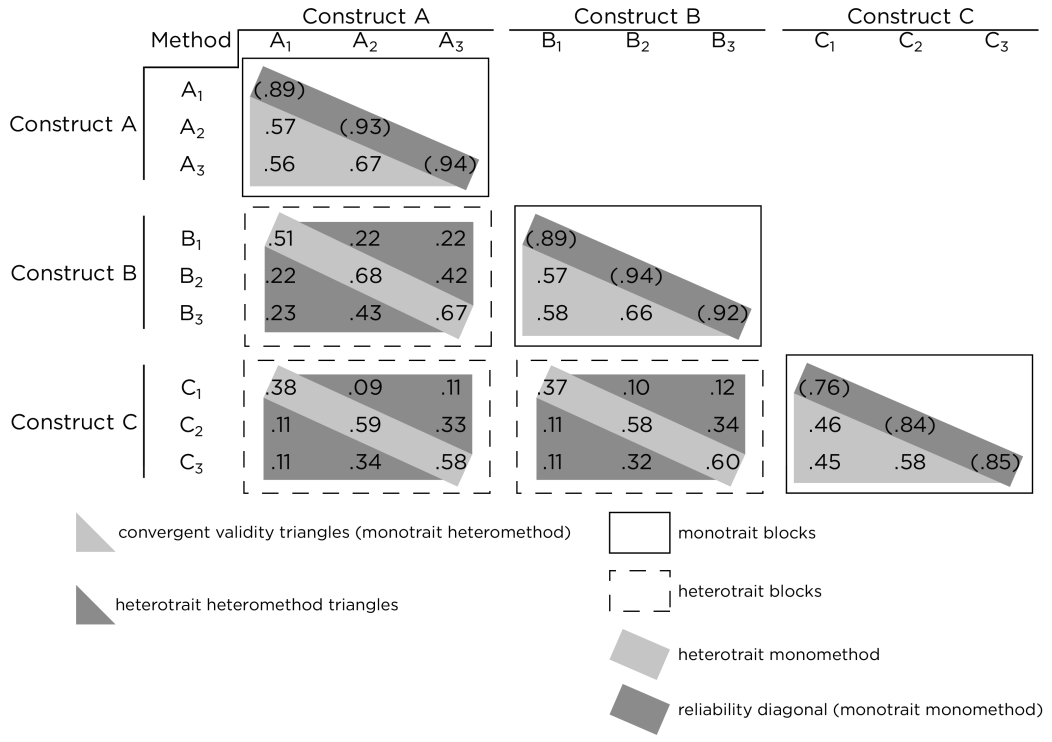

FIGURE 4.11 Multitrait-Multimethod Matrix Organized by Construct Then by Method.

The authors updated the HTMT to use the geometric mean of the correlations rather than the arithmetic mean of the correlations to relax the assumption of tau-equivalence (Roemer et al., 2021). They called the updated index HTMT2. HTMT2 was calculated below using the `semTools` package (Jorgensen et al., 2021). In this case, the HTMT values are less than .85, providing evidence of discriminant validity.

```
modelCFA <- '
  visual  =~ x1 + x2 + x3
  textual =~ x4 + x5 + x6
  speed   =~ x7 + x8 + x9
'

htmt(
  modelCFA,
  data = HolzingerSwineford1939,
  missing = "ml")
```

```
        visual textul speed
visual   1.000
textual  0.384  1.000
```

```
speed    0.387  0.280 1.000
```

Other indexes of discriminant validity include the CI_{CFA}(sys) (Rönkkö & Cho, 2020), χ^2(sys) (Rönkkö & Cho, 2020), and the Fornell-Larcker Ratio [FLR; Fornell & Larcker (1981)]. The FLR is based on the idea that a construct should be more strongly associated with its indicators than with other constructs, and is computed as the ratio of (a) the highest squared inter-correlation the construct has with other constructs to (b) the average variance extracted [AVE] by the construct. A criterion of FLR < 1 evaluates whether the AVE of a factor is greater than the highest squared inter-correlation between the factor and other factors in the model (Fornell & Larcker, 1981). Another index estimates the degree of correspondence of convergent and discriminant validity correlations with a priori hypotheses as an index of construct validity (Furr & Heuckeroth, 2019).

Using an MTMM, a researcher can learn a lot about the quality of a measure and build a nomological network. For instance, we can estimate the extent of method variance, or variance that is attributable to the measurement method rather than the construct of interest. Method variance is estimated by the difference between monomethod versus heteromethod blocks.

A poor example of an MTMM would be having two constructs such as height and depression and two methods such as Likert and true/false. This would be a poor MTMM for two reasons: (1) there are trivial differences in the methods, which would lead to obvious convergence, and (2) the differences between the constructs are not important—they are obviously discriminant. Using such an MTMM would find strong convergent associations because the methods are maximally similar and weak discriminant associations because the constructs are maximally different. It would be better to use maximally different measurement methods (e.g., self-report and performance-based measures) and to use constructs that are important to distinguish (e.g., depression and anxiey).

The paper by Campbell and Fiske (1959) that introduced the MTMM is one of the most widely cited papers in psychology of all time. *Psychological Bulletin* published a more recent paper by Fiske and Campbell (1992), entitled "Citations Do Not Solve Problems." They noted how their original paper was the most widely cited paper in the history of *Psychological Bulletin*, but they argued that nothing came of their paper. The MTMM matrices published today show little-to-no improvement compared to the ones they published in 1959. In part, this may be because we need better measures (i.e., a higher ratio of construct to method variance).

4.3.1.4.2.3 Multitrait-multimethod correlation matrix

This example is courtesy of W. Joel Schneider. First, we simulate data for an MTMM model using a fixed model with population parameters using the `simstandard` package (Schneider, 2021):

```
model_fixed <- '
 Verbal =~ .5*VO1 + .6*VO2 + .7*VO3 +
           .7*VW1 + .6*VW2 + .5*VW3 +
           .6*VM1 + .7*VM2 + .5*VM3
 Spatial =~ .7*SO1 + .7*SO2 + .6*SO3 +
            .6*SW1 + .7*SW2 + .5*SW3 +
            .7*SM1 + .5*SM2 + .7*SM3
 Quant =~ .5*QO1 + .7*QO2 + .5*QO3 +
```

TABLE 4.1 Multitrait-Multimethod Correlation Matrix.

	VO1	VO2	VO3	VW1	VW2	VW3	VM1	VM2	VM3	SO1	SO2	SO3	SW1	SW2	SW3	SM1	SM2	SM3	QO1	QO2	QO3	QW1	QW2	QW3	QM1	QM2	QM3
VO1	1.00																										
VO2	.50	1.00																									
VO3	.47	.57	1.00																								
VW1	.35	.43	.50	1.00																							
VW2	.30	.37	.43	.66	1.00																						
VW3	.27	.31	.37	.54	.43	1.00																					
VM1	.31	.37	.44	.43	.36	.33	1.00																				
VM2	.36	.42	.50	.50	.43	.37	.67	1.00																			
VM3	.25	.31	.35	.35	.30	.25	.46	.49	1.00																		
SO1	.36	.44	.43	.36	.30	.26	.30	.34	.24	1.00																	
SO2	.35	.44	.43	.35	.30	.26	.30	.34	.24	.58	1.00																
SO3	.42	.50	.45	.29	.26	.22	.25	.28	.22	.57	.57	1.00															
SW1	.20	.25	.29	.65	.50	.38	.26	.29	.21	.42	.43	.34	1.00														
SW2	.24	.29	.35	.65	.50	.42	.31	.35	.25	.49	.49	.41	.72	1.00													
SW3	.18	.22	.25	.54	.41	.33	.22	.25	.18	.35	.36	.30	.59	.61	1.00												
SM1	.25	.30	.36	.36	.31	.26	.54	.59	.39	.50	.50	.42	.43	.51	.36	1.00											
SM2	.18	.21	.26	.25	.21	.18	.46	.49	.32	.34	.34	.30	.29	.35	.26	.59	1.00										
SM3	.26	.31	.36	.36	.31	.27	.60	.64	.42	.50	.50	.42	.43	.50	.36	.79	.65	1.00									
QO1	.39	.47	.39	.22	.18	.16	.18	.22	.15	.35	.35	.44	.14	.17	.13	.17	.12	.18	1.00								
QO2	.32	.39	.39	.28	.25	.23	.25	.29	.21	.33	.33	.37	.19	.24	.17	.24	.17	.25	.53	1.00							
QO3	.30	.37	.33	.20	.17	.15	.18	.21	.15	.29	.29	.34	.13	.17	.12	.17	.12	.18	.49	.47	1.00						
QW1	.15	.19	.22	.45	.34	.28	.19	.21	.16	.18	.18	.15	.39	.38	.32	.18	.12	.18	.26	.35	.26	1.00					
QW2	.17	.20	.25	.49	.38	.31	.22	.25	.19	.21	.21	.17	.41	.42	.35	.23	.15	.22	.29	.41	.29	.46	1.00				
QW3	.19	.24	.29	.59	.45	.37	.25	.29	.20	.24	.25	.20	.50	.50	.43	.25	.16	.25	.34	.48	.35	.55	.61	1.00			
QM1	.14	.18	.22	.21	.18	.17	.39	.41	.27	.18	.17	.16	.14	.18	.14	.38	.32	.42	.25	.34	.24	.26	.30	.35	1.00		
QM2	.17	.21	.26	.25	.21	.20	.37	.40	.25	.20	.21	.17	.18	.20	.15	.36	.30	.39	.30	.42	.29	.30	.35	.41	.42	1.00	
QM3	.21	.25	.31	.31	.26	.24	.41	.45	.30	.24	.24	.21	.21	.26	.19	.39	.32	.43	.35	.49	.35	.35	.43	.49	.48	.50	1.00

```
           .5*QW1 + .6*QW2 + .7*QW3 +
           .5*QM1 + .6*QM2 + .7*QM3
 Oral =~ .4*VO1 + .5*VO2 + .3*VO3 +
         .3*SO1 + .3*SO2 + .5*SO3 +
         .6*QO1 + .3*QO2 + .4*QO3
 Written =~ .6*VW1 + .4*VW2 + .3*VW3 +
            .6*SW1 + .5*SW2 + .5*SW3 +
            .4*QW1 + .4*QW2 + .5*QW3
 Manipulative =~ .5*VM1 + .5*VM2 + .3*VM3 +
                 .5*SM1 + .5*SM2 + .6*SM3 +
                 .4*QM1 + .3*QM2 + .3*QM3
 Verbal ~~ .7*Spatial + .6*Quant
 Spatial ~~ .5*Quant
'

MTMM_data <- sim_standardized(
  model_fixed,
  n = 10000,
  observed = TRUE,
  latent = FALSE,
  errors = FALSE)
```

A multitrait-multimethod matrix correlation matrix is presented in Table 4.1.

4.3.1.4.2.4 Construct validity beyond Campbell and Fiske

There are a number of other approaches that can be helpful for establishing construct validity in ways that go beyond the approaches proposed by Campbell and Fiske (1959). One way is known as triangulation. Triangulation is conceptually depicted in Figure 4.12. Triangulation involves testing a hypothesis with multiple measures and/or methods to see if the findings are consistent—that is, whether the findings triangulate.

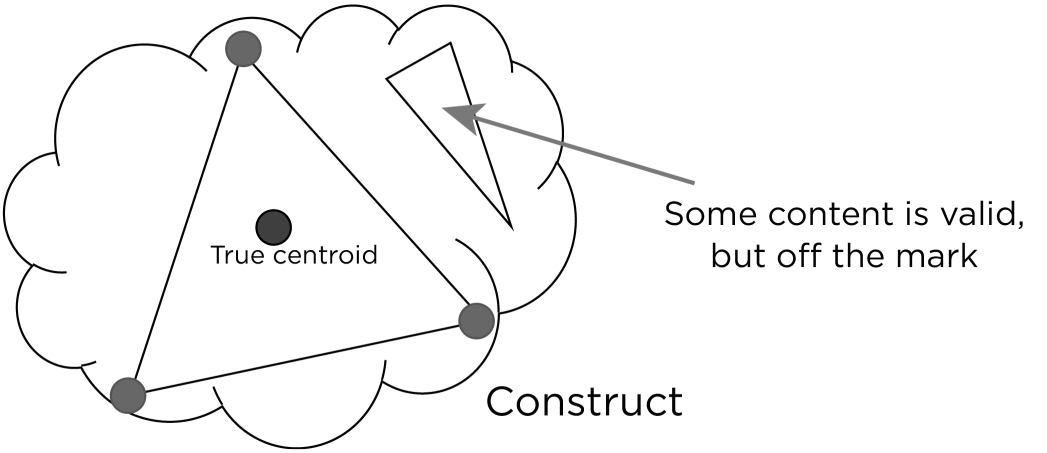

FIGURE 4.12 Using Triangulation to Arrive at a Closer Estimate of the Construct Using Multiple Measures and/or Methods.

A contemporary approach to multitrait-multimethod modeling uses confirmatory factor analysis, as described below.

4.3.1.4.2.5 MTMM in confirmatory factor analysis

Using modern modeling approaches, there are even more advanced ways of examining an MTMM. For instance, you can use structural equation modeling (SEM) or confirmatory factor analysis (CFA) to derive a latent variable of a construct from multiple methods to be free from method-related error variance to generate purer assessments of the construct to see how it relates to other constructs. For an example of an MTMM in confirmatory factor analysis, see Figure 4.13 and Section 6.4.2.12 in Chapter 6 on factor analysis.

4.3.1.5 Incremental Validity

Accuracy is not enough for a measure to be useful. Proposed by Sechrest (1963), measures should also be judged by the extent to which they provide an increment in predictive efficiency over the information otherwise easily and cheaply available. *Incremental validity* deals with the incremental value or utility over a measure beyond other sources of information. It must be demonstrated that the addition of a measure will produce better predictions than are made on the basis of information ordinarily available. It is not enough to show that the measure is better than chance, and the measure should not just be capitalizing on shared method variance with the criterion or on increased reliability of the measure. That is, the measure should explain truly unique variance—variance that was not explained before. Incremental validity demonstrates added value, unless the other measure is cheaper or less time-consuming. Incremental validity is a specific kind of criterion-related validity: significantly increased R^2 in hierarchical regression. The incremental validity of a measure can be evaluated by examining whether the measure explains significant unique variance in the criterion when accounting for other information, such as easily accessible information, traditionally available measures, or the current gold-standard measure. The extent of incremental validity of a measure can be quantified with the change in the coefficient of determination (ΔR^2) that compares (a) the model that includes the old predictor(s) to (b) the model that includes old predictors and the new predictor (measure).

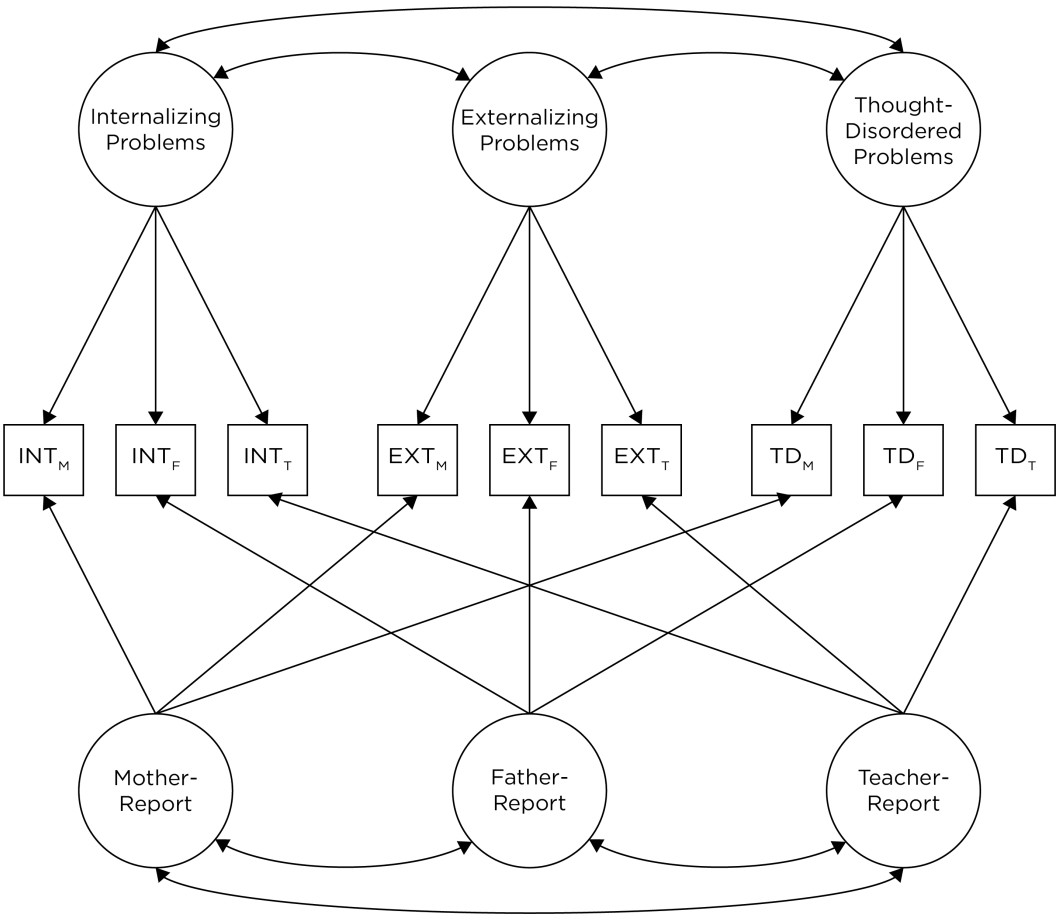

FIGURE 4.13 Multitrait-Multimethod Model in Confirmatory Factor Analysis with Three Constructs (Internalizing, Externalizing, and Thought-Disordered Problems) and Three Methods (Mother-, Father-, and Teacher-report).

```
model1 <- lm(
  criterion ~ oldpredictor,
  data = na.omit(mydataValidity))

model2 <- lm(
  criterion ~ oldpredictor + predictor,
  data = na.omit(mydataValidity))

model1Rsquare <- summary(model1)$r.squared
model2Rsquare <- summary(model2)$r.squared

deltaRsquare <- model2Rsquare - model1Rsquare

deltaRsquare
```

```
[1] 0.03628
```

```
anova(model1, model2)
```

```
Analysis of Variance Table

Model 1: criterion ~ oldpredictor
Model 2: criterion ~ oldpredictor + predictor
  Res.Df   RSS Df Sum of Sq   F Pr(>F)
1    900 44295
2    899 36886  1      7409 181 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The predictor shows significant incremental validity above the old predictor in predicting the criterion. `Model 1` explains 78.31% of the variance ($R^2 = 0.78$), and `Model 2` explains 81.94% of the variance ($R^2 = 0.82$). Thus, the predictor explains only 3.63% additional variance in the criterion above the variance explained by the old predictor (ΔR^2).

4.3.1.6 Treatment Utility of Assessment

S. C. Hayes et al. (1987) argued that it is not enough for measures to be reliable and valid. The authors raised another important consideration for the validity of a measure: its treatment utility or usefulness. They asked the question: What goal do assessments accomplish in clinical psychology? In clinical psychology, the goal of assessment is to lead to better treatment outcomes. Therefore, the *treatment utility* of a measure is the extent to which a measure is shown to contribute to beneficial treatment outcomes. That is, if a clinician has the information/results from having administered this measure, do the clients have better outcomes? The treatment utility of assessment is a specific kind of criterion-related validity, with the idea that "all criteria are not created equal." And the criterion that is most important to optimize, from this perspective, when developing and selecting measures is a client's treatment outcomes.

S. C. Hayes et al. (1987) described different approaches to evaluating the extent to which a measure shows treatment utility. These are a priori group comparison approaches that examine whether a specific difference in the assessment approach relates to treatment outcome. They distinguished between (a) manipulated assessment and (b) manipulated use. Manipulated assessment and manipulated use are depicted in Figure 4.14.

In *manipulated assessment*, a single group of subjects is randomly divided into two subgroups, and either the collection or availability of assessment data is varied systematically. Therapists then design or implement treatment in accord with the data available. As an example, the measure of interest may be administered in one condition, and the standard measure may be administered in the other condition. Then the treatment outcomes would be compared across the two conditions. An advantage of manipulated assessment is that this type of design is more realistic than manipulated use. Making the assessment data available but not assigning a certain treatment based on the assessment outcomes better simulates a realistic clinical environment. Also, because the control group has no access to the data, it might serve as a stronger control. A disadvantage of manipulated assessment is that it depends on whether and how clinicians use the measure, which depends on how positively the measure was received by the clinicians.

In *manipulated use*, the same assessment information is available for all subjects, but the researcher manipulates the way in which the assessment information is used. For example,

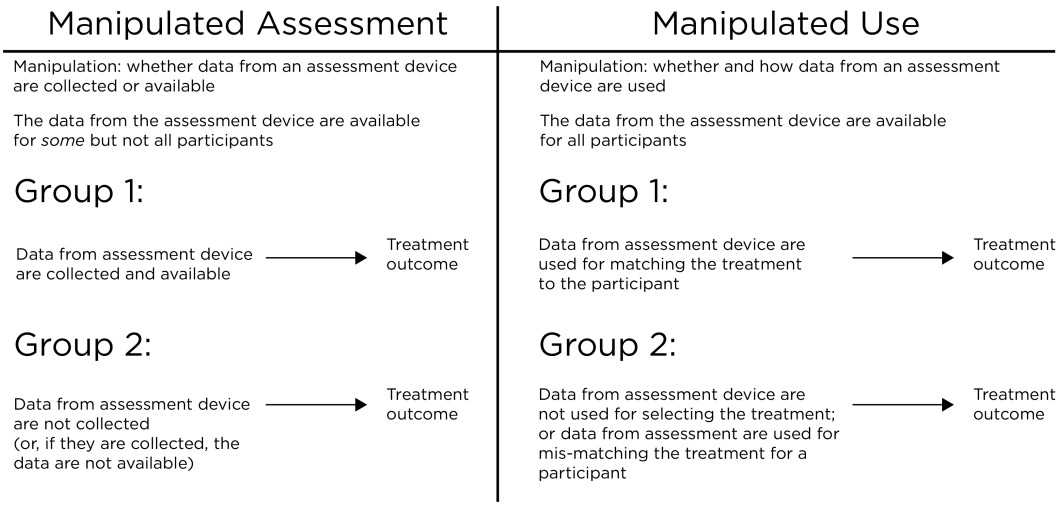

FIGURE 4.14 Research Designs That Evaluate the Treatment Utility of Assessment.

one group gets a treatment matched to assessment outcomes, and the other group gets a cross-matched treatment that does not target the problem identified by the assessment. So, in one group, the assessment information is used to match the treatment to the client based on their assessment results, whereas the other group receives a standard treatment regardless of their assessment results. An advantage of this design is that you can be certain that the treatment decisions are explicitly informed by the assessment results because the researcher ensures this, whereas the decision of how the assessment information is used is up to the clinician in manipulated assessment.

Relatively few measures show evidence of treatment utility of assessment. A review on the treatment utility of assessment is provided by Nelson-Gray (2003).

4.3.1.7 Discriminative Validity

Discriminative validity is the degree to which a measure accurately identifies persons placed into groups on the basis of another measure. Discriminative validity is not to be confused with discriminant (divergent) validity and discriminant analysis. A measure shows discriminant validity if it does not correlate with things that it would not be theoretically expected to correlate with. A measure shows discriminative validity, by contrast, if the measure is able to accurately differentiate things (e.g., two groups such as men and women). Discriminant analysis is a model that combines predictors to differentiate between multiple groups.

4.3.1.8 Elaborative Validity

Elaborative validity involves the extent to which a measure increases our theoretical understanding of the target construct or of neighboring constructs. It deals with a measure's meaningfulness. Elaborative validity is a type of incremental theoretical validity. It is a combination of criterion-related validity and construct validity that examines how much a given measure increases our understanding of a construct's nomological network. However, I am unaware of strong examples of measures that show strong elaborative validity in psychology.

4.3.1.9 Consequential Validity

Consequential validity is a form of validity that differs from evidential validity, or a test's potential to provide accurate, useful information based on research. Consequential validity takes a more macro-view and deals with the community, sociological, and public policy perspective. *Consequential validity* evaluates the consequences of our measures beyond the circumstances of their development, based on their impact. Measures can have positive, negative, or mixed effects on society. An example of consequential validity would be asking what the impacts of aptitude testing are on society—how aptitude testing affects society in the long run. Some would argue that, even in cases where the aptitude tests have some degree of evidential validity (i.e., they accurately assess to some degree what they tend to assess), they are consequentially invalid—that is, they have had a net negative effect on society and, due to their low value to society, are invalid. The tests themselves may be fine in terms of their accuracy, but consequential validity says that their validity depends on what we do with the test, i.e., how the test is used.

Another example of consequential validity is when the validity of a measure changes over time due to changes in people's or society's response, as depicted in Figure 4.15. Judgments and predictions can change the way society reacts and the way people behave so that the predictions become either more or less accurate. A prediction that becomes more accurate as a result of people's response to a prediction is a self-fulfilling prediction or prophecy. For instance, if school staff make a prediction that a child will be held back a grade, the teacher may not provide adequate opportunities for the child to learn, which may lead to the child being more likely to be held back.

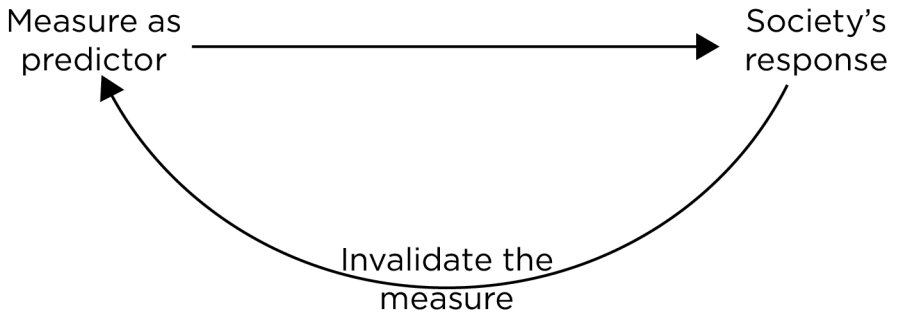

FIGURE 4.15 Invalidation of a Measure Due to Society's Response to the Use of the Measure.

A prediction that becomes less accurate as a result of people's response to a prediction is a self-canceling prediction. The most effective prediction about a flu outbreak might be one that leads people to safer behavior; therefore lowering flu rates, which does not correspond to the initial prediction (Silver, 2012). Society's response to a measure can invalidate the measure. For example, consider that an organization rates cities based on quality-of-life measures. If quality-of-life indices include the percent of solved cases by the city's police (i.e., the clearance rate), cities may try to improve their ratings of "quality-of-life" by increasing their clearance rate either by increasing the number of cases they mark as solved (such as by marking cases falsely as resolved) or by decreasing the number of cases (such as by investigating fewer cases). That is, cities may "game" the system based on the quality-of-life indices used by various organizations. In this case, the clearance rate becomes a less accurate measure of a city's quality of life.

Another example of a measure that becomes less valid due to society's response is the use of alumni donations as an indicator of the strength of a university that is used for generating university rankings. Such an indicator could lead schools to admit wealthier students who give the most donations, and students whose parents were alumni and provide lavish donations to the university. Yet another example could be standardized testing, where instructors may "teach to the test" such that better performance might not necessarily reflect better underlying competence.

4.3.1.10 Representational Validity

Representational validity examines the extent to which the items or content of a measure "flesh out" and mirror the true nature and mechanisms of the construct. There are many different types of validity, but many of them are overlapping and can be considered types of other forms of validity. For instance, representational validity is a type of elaborative validity and content validity.

4.3.1.11 Factorial (Structural) Validity

Factorial validity is examined in Section 6.1.4.2.2 on factor analysis. *Factorial validity* considers whether the factor structure of the measure(s) is consistent with the construct. According to factorial validity, if you claim that the measure is unidimensional with four items, factor analysis should support the unidimensionality. Thus, it involves testing empirically whether the measure has the same structure as would be expected based on theory. It is a type of construct validity.

4.3.1.12 Ecological Validity

Ecological validity examines the extent to which a measure provides scores that are indicative of the behavior of a person in the natural environment.

4.3.1.13 Process-Focused Validity

Process-focused validity attempts to get closer to the underlying mechanisms. *Process-focused validity* examines the degree to which respondents engage in a predictable set of psychological processes (which are specified a priori) when completing the measure (Bornstein, 2011; Furr, 2017). These psychological processes include effects of the instrument (e.g., observer versus third-party report versus projective test) as well as effects of the context (e.g., assessment setting, assessment instructions, affect state of the participant, etc.). To determine whether a test is valid in the process-focused technique, one experimentally manipulates variables that moderate the test score–criterion association—to better understand the underlying processes. The ideal outcome is a measure that both (1) has an adequate outcome (correlations where expected) as well as (2) adequate process validity—the psychological processes that one engages in are well hypothesized.

The idea of process-focused validity is that if a measure does not inform us about process or mechanisms, it is not worth doing. Process-focused validity is a type of elaborative validity and construct validity. For instance, consider the common finding that low socioeconomic status is associated with poorer outcomes. To make an impact, process-focused validity would argue that we need to know the mechanisms that underlie this association, and it involves how a measure helps us understand process.

4.3.1.14 Diagnostic Validity

Diagnostic validity is the extent to which the diagnostic category accurately captures the abnormal phenomenon of interest. It is a form of construct validity for diagnoses.

4.3.1.15 Social Validity

Social validity involves the extent to which the proposed procedures (assessment, intervention, etc.) will be well-liked and acceptable by those who receive and implement them.

4.3.1.16 Cultural Validity

Cultural validity refers to "the effectiveness of a measure or the accuracy of a clinical diagnosis to address the existence and importance of essential cultural factors" (Leong & Kalibatseva, 2016, p. 58). Essential cultural factors may include values, beliefs, experiences, communication patterns, and approaches to knowledge (epistemologies).

4.3.2 Research Design (Experimental) Validity

Aspects of *research design validity*, also called experimental validity, involve the validity of a research design for making various interpretations. Research design validity includes internal validity, external validity, and conclusion validity.

4.3.2.1 Internal Validity

Internal validity is the extent to which causal inference is justified from the research design. This encompasses multiple features including

- temporal precedence—does the (purported) cause occur before the (purported) effect?
- covariation of cause and effect—correlation is necessary (even if insufficient) for causality
- no plausible alternative explanations—such as third variables that could influence both variables and explain their covariation

There are number of potential threats to internal validity, which are important to consider when designing studies and interpreting findings. Examples of potential threats to internal validity include history, maturation, testing, instrumentation, regression, selection, experimental mortality, and an interaction of threats (Slack & Draugalis, 2001).

Research designs differ in the extent to which they show internal validity versus external validity. An experiment is a research design in which one or more variables (independent variables) are manipulated to observe how the manipulation influences the dependent variable. In an experiment, the researcher has greater control over the variables and attempts to hold everything else constant (e.g., by standardization and random assignment). In correlational designs, however, the researcher has less control over the variables. They may be able to statistically account for potential confounds using covariates or for the reverse direction of effect using longitudinal designs. Nevertheless, we can have greater confidence about whether a variable influences another variable in an experiment. Thus, experiments tend to have higher internal validity than correlational designs.

4.3.2.2 External Validity

External validity is the extent to which the findings can be generalized to the broader population and the real world. External validity is crucial to consider for studies that intend to make inferences to people outside of those who were assessed. For instance, norm-referenced assessments attempt to identify the distribution of scores for a given population from a

sample within that population. The validity of the norms of a norm-referenced assessment are important to consider (Achenbach, 2001). The validity of norms and external validity, more broadly, can depend highly on how representative the sample is of the target population and how appropriate this population is to a given participant. Some measures are known to have poor norms, including the Exner Comprehensive System (Exner, 1974; Exner & Erdberg, 2005) for administering and scoring the Rorschach Inkblot Test, which has been known to over-pathologize (Wood, Teresa, et al., 2001; Wood, Nezworski, et al., 2001). The Rorschach is discussed in greater detail in Chapter 18.

4.3.2.2.1 Tradeoffs of internal validity and external validity

It is important to note that there is a tradeoff between internal and external validity—a single research design cannot have both high internal and high external validity. Some research designs are better suited for making causal inferences, whereas other designs tend to be better suited for making inferences that generalize to the real world. The research design that is best suited to making causal inferences is an experiment, where the researcher manipulates one variable (the independent variable) and holds all other variables constant to see how a change in the independent variable influences the outcome (dependent variable). Thus, experiments tend to have higher internal validity than other research designs. However, by manipulating one variable and holding everything else constant, the research takes place in a very standardized fashion that can become like studying a process in a vacuum. So, even if a process is theoretically causal in a vacuum, it may act very differently in the real world when it interacts with other processes.

Observational designs have greater capacity for external validity than experimental designs because people can be observed in their natural environments to see how variables are related in the real world. However, the greater external validity comes at a cost of lower internal validity. Observational designs are not well-positioned to make causal inferences because they have multiple threats to internal validity, including issues of temporal precedence in cross-sectional observational designs, and there are numerous potential third variables (i.e., confounds) that could act as a common cause of both the predictor and outcome variables. Thus, just because two variables are associated does not necessarily mean that they are causally related.

As the internal validity of a study's design increases, its external validity tends to decrease. The greater control we have over variables (and, therefore, have greater confidence about causal inferences), the lower the likelihood that the findings reflect what happens in the real world because it is studying things in a metaphorical vacuum. Because no single research design can have both high internal and external validity, scientific inquiry needs a combination of many different research designs so we can be more confident in our inferences—experimental designs for making causal inferences and observational designs for making inferences that are more likely to reflect the real world.

Case studies, because they have smaller sample sizes, tend to have lower external validity than both experimental and observational studies. Case studies also tend to have lower internal validity because they have less potential to control for threats to internal validity, such as potential confounds or temporal precedence. Nevertheless, case studies can still be useful for generating hypotheses that can then be tested empirically with a larger sample in experimental or observational studies.

4.3.2.3 (Statistical) Conclusion Validity

Conclusion validity, also called statistical conclusion validity, considers the extent to which conclusions are reasonable about the association among variables based on the data. That is, were the correct statistical analyses performed, and are the interpretations of the findings from those analyses correct?

4.3.3 Putting It All Together: An Organizational Framework

There are many types of measurement validity, but the central psychometric aspect of measurement validity is construct validity. That is, whether the measure accurately assesses the target construct is the most important consideration of measurement validity. As discussed earlier, construct validity includes the nomological network of the construct. Construct validity also subsumes other key types of measurement validity, including

- Convergent validity
- Discriminant (divergent) validity
- Criterion-related validity
  - Concurrent validity
  - Predictive validity
- Content validity

The organization of types of measurement validity that are subsumed by construct validity is depicted in Figure 4.16.

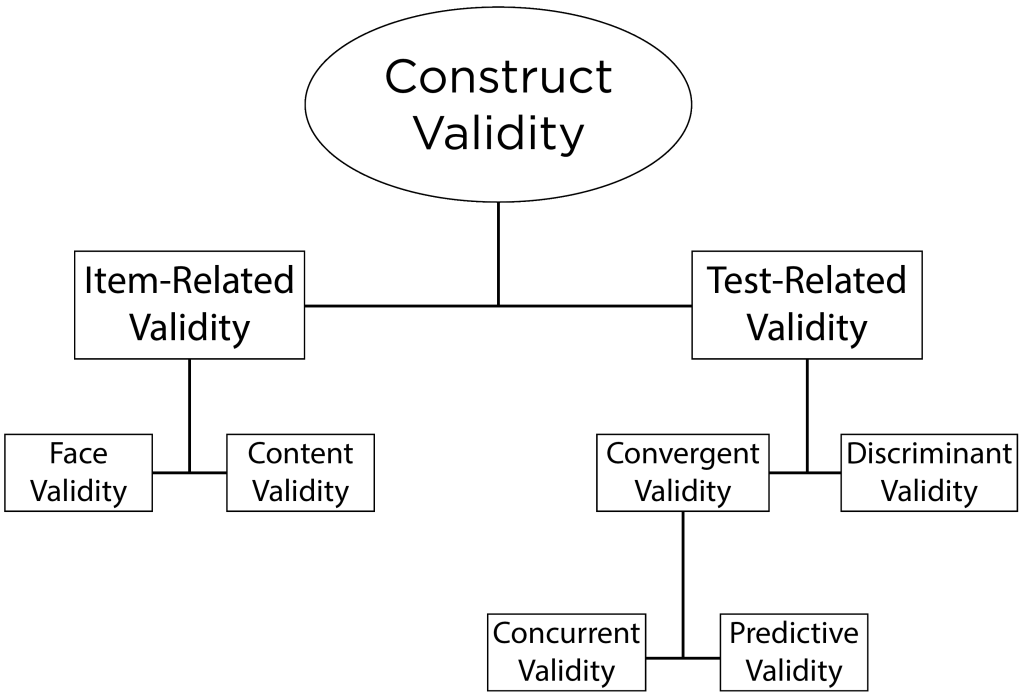

FIGURE 4.16 Organization of Types of Measurement Validity That are Subsumed by Construct Validity.

Moreover, many different types of reliability and validity can be viewed through the lens of construct validity:

- Internal consistency, which can be estimated as the coefficient of internal consistency, where the criterion for criterion-related validity is other items on the same measure
- Test–retest reliability, which can be estimated as the coefficient of stability, where the criterion for criterion-related validity is the same measure at another time point
- Parallel-forms reliability or convergent validity, which can be estimated as the coefficient of equivalence, where the criterion for criterion-related validity is the parallel form

4.4 Validity Is a Process, Not an Outcome

Validity (and validation) is a continual process, not an outcome. Validation is never complete. When evaluating the validity of a measure, we must ask: Validity for what and to what degree? We would not just say that a measure is or is not valid. We would express the strength of evidence on a measure's validity across the different types of validity for a particular purpose, with a particular population, in a particular context (consistent with generalizability theory).

4.5 Reliability Versus Validity

Reliability and validity are not the same thing. Reliability deals with consistency, whereas validity deals with accuracy. A measure can be consistent but not accurate (see Figure 4.17). That is, a measure can be reliable but not valid. However, a measure cannot be accurate if it is inconsistent; that is, a measure cannot be valid and unreliable.

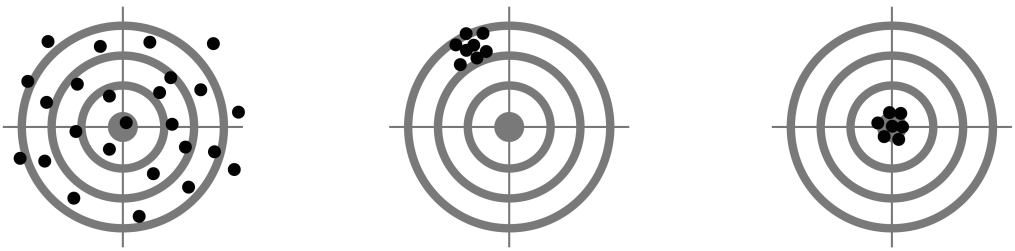

FIGURE 4.17 Traditional Depiction of Reliability (Consistency) Versus Validity (Accuracy).

The typical way of depicting the distinction between reliability and validity is in Figure 4.17, in which a measure can either have (a) low reliability and low validity, (b) high reliability and low validity, or (c) high reliability and high validity. However, it can be worth thinking about validity in terms of accuracy at the person level versus group level. When we distinguish between person- versus group-level accuracy, we can distinguish four general combinations of reliability and validity, as depicted in Figure 4.18: (a) low reliability, low accuracy at the person level, and low accuracy at the group level, (b) low reliability, low accuracy at the

person level, and high accuracy at the group level, (c) high reliability, low accuracy at the person level, and low accuracy at the group level, and (d) high reliability, high accuracy at the person level, and high accuracy at the group level. However, as discussed before, reliability and validity are not binary states of low versus high—they exist to varying degrees.

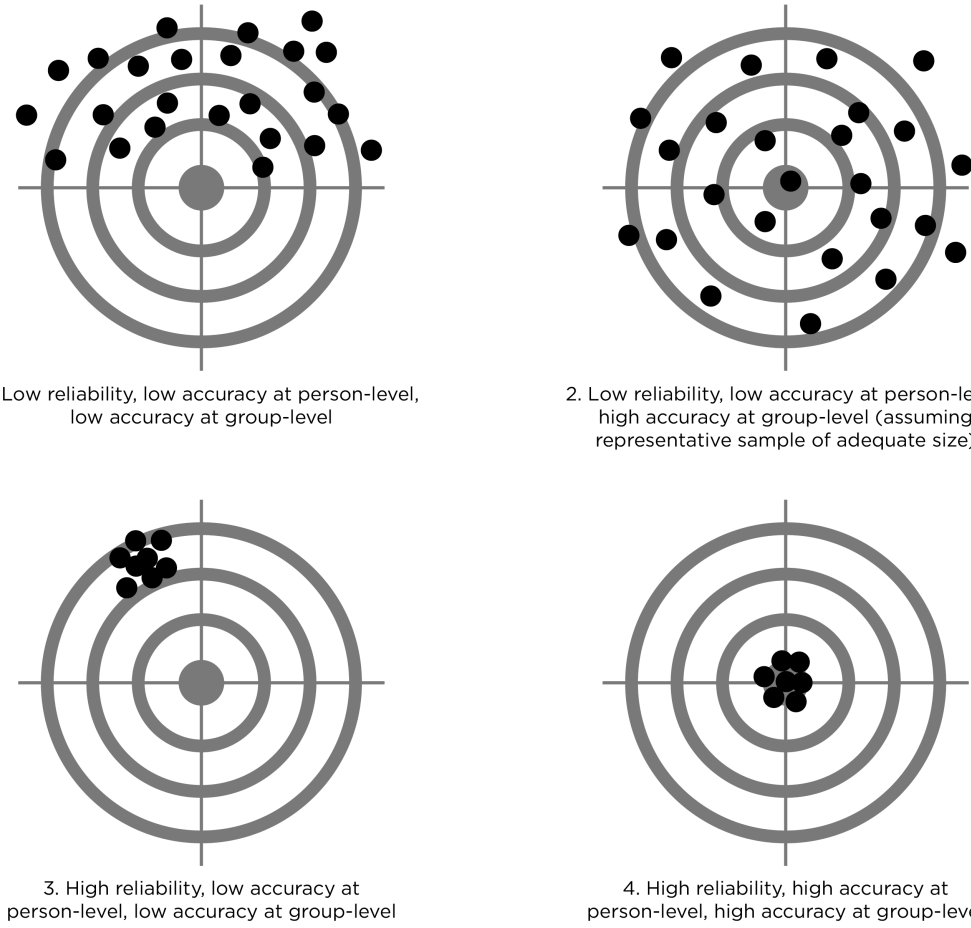

FIGURE 4.18 Depiction of Reliability Versus Validity, While Distinguishing Between Validity (Accuracy) at the Person Versus Group Level.

Even though reliability and validity are not the same thing, there is a relation between reliability and validity. Validity depends on reliability. Reliability is necessary but insufficient for validity. However, measurement error (unreliability) can be systematic or random. If measurement error is systematic, it reduces the validity of the measure. If measurement error is random, it reduces the precision of an individual's score on a measure, but the measure could still be a valid measure of the construct at the group level. However, random error would make it more difficult to make an accurate prediction for an individual person.

Reliability places the upper bound on validity because a measure can be no more valid than it is reliable. In other words, a measure should not correlate more highly with another variable than it correlates with itself. Therefore, the maximum validity coefficient is the square root of the product of the reliability of each measure, as in Equation (4.1):

$$\begin{aligned} r_{xy_{\max}} &= \sqrt{r_{xx} r_{yy}} \\ \text{maximum association between } x \text{ and } y &= \sqrt{\text{reliability of } x \text{ and } y} \end{aligned} \tag{4.1}$$

So, the maximum validity coefficient is based on the reliability of each measure. To the extent that one of the measures is unreliable, the validity coefficient will be attenuated relative to the true validity (i.e., the true strength of association of the constructs), as we describe below.

4.6 Effect of Measurement Error on Associations

Figure 4.19 depicts the classical test theory approach to understanding the validity of a measure, i.e., its association with another measure, which is the validity coefficient (r_{xy}).

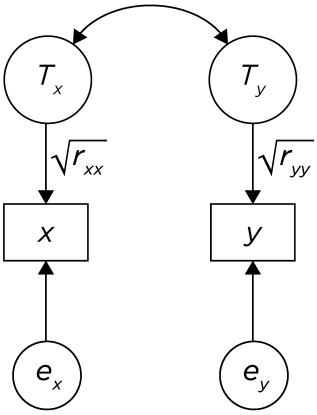

FIGURE 4.19 The Criterion-Related Validity of a Measure, i.e., Its Association with Another Measure, as Depicted in a Path Diagram.

As described above, (random) measurement error weakens (or attenuates) the association between variables (Goodwin & Leech, 2006; Schmidt & Hunter, 1996). The greater the random measurement error, the weaker the association. Thus, the correlation between x and y depends on both the true correlation of x and y ($r_{x_t y_t}$) and the reliabilities of x (r_{xx}) and y (r_{yy}). So, measurement error in x and y can reduce the observed correlation below the true correlation. This is known as the *attenuation formula* (Equation (4.2)):

$$\begin{aligned} r_{xy} &= r_{x_t y_t} \sqrt{r_{xx} r_{yy}} \\ \text{observed association between } x \text{ and } y &= (\text{true association of constructs}) \\ &\quad \times \sqrt{\text{reliability of } x \text{ and } y} \end{aligned} \tag{4.2}$$

The lower the reliability, the greater the attenuation of the validity coefficient relative to the true association between the constructs.

All of these r values (excluding the true correlation) are just estimates unless the sample size is infinite, so the observed association is an imperfect estimate. Hence, we need a correction for this attenuation (Schmidt & Hunter, 1996). This correction for the attenuation of an association due to measurement error (unreliability) is known as the disattenuation of a correlation, i.e., correction for the attenuation of an association due to measurement error to get a more accurate estimate of the true association between constructs. Rearranging the terms from the attenuation formula, the formula for disattenuation of a correlation (i.e., the *disattenuation formula*) is in Equation (4.3):

$$\begin{aligned} r_{x_t y_t} &= \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \\ \text{true association of constructs} &= \frac{\text{observed association between } x \text{ and } y}{\sqrt{\text{reliability of } x \text{ and } y}} \end{aligned} \tag{4.3}$$

All of this is implied in the path diagram (see Figure 4.19). The attenuation and disattenuation formulas are based on classical test theory, and therefore assume that all measurement error is random, that errors are uncorrelated, etc. Nevertheless, the attenuation formula can be informative for understanding how imperiled your research is when your measures have low reliability (i.e., when there is instability in the measure). Researchers recommend accounting for measurement reliability, to better estimate the association between constructs, either with the disattenuation formula (Schmidt & Hunter, 1996) or with structural equation modeling, as described in the next chapter.

4.6.1 Attenuation of True Correlation Due to Measurement Error

The attenuation formula is presented in Equation (4.2). We extend it to a specific example in Equation (4.4):

$$\begin{aligned} r_{xy} &= r_{x_t y_t} \sqrt{r_{xx} r_{yy}} \\ \text{observed correlation between } x \text{ and } y &= (\text{true association between construct } A \text{ and construct } B) \times \\ &\quad \sqrt{\text{reliability of } x \times \text{reliability of } y} \end{aligned} \tag{4.4}$$

where x = measure of construct A; y = measure of construct B

Below is an example of how to find the observed correlation between the predictor and criterion if the true association between the constructs (i.e., correlation between true scores of constructs) is .70, the reliability of the predictor is .90, and the reliability of the criterion is .85. The `petersenlab` package (Petersen, 2024) contains the `attenuationCorrelation()` function that estimates the observed association given the true association and the reliability of the predictor and criterion:

```
attenuationCorrelation(
  trueAssociation = 0.7,
  reliabilityOfPredictor = 0.9,
  reliabilityOfCriterion = 0.85)
```

```
[1] 0.6122
```

The observed association (r = .61) is attenuated relative to the true association (r = .70).

4.6.2 Disattenuation of Observed Correlation Due to Measurement Error

The disattenuation formula is presented in Equation (4.3). We extend it to a specific example in Equation (4.5):

$$\begin{aligned} r_{x_t y_t} &= \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}} \\ \text{true association between construct } A \text{ and construct } B &= \frac{\text{observed correlation between } x \text{ and } y}{\sqrt{\text{reliability of } x \times \text{reliability of } y}} \end{aligned} \tag{4.5}$$

where x = measure of construct A; y = measure of construct B.

Find the true association between the construct assessed by the predictor and the construct assessed by the criterion given an observed association if the reliability of the predictor is .9, and the reliability of the criterion is .85: The `petersenlab` package (Petersen, 2024) contains the `disattenuationCorrelation()` function that estimates the observed association given the true association and the reliability of the predictor and criterion:

```
observedAssociation <- cor.test(
  x = mydataValidity$predictor,
  y = mydataValidity$criterion)$estimate

observedAssociation
```

```
   cor
0.7079
```

```
reliabilityOfPredictor <- 0.9
reliabilityOfCriterion <- 0.85

disattenuationCorrelation(
  observedAssociation = observedAssociation,
  reliabilityOfPredictor = reliabilityOfPredictor,
  reliabilityOfCriterion = reliabilityOfCriterion)
```

```
   cor
0.8094
```

The disattenuation of an observed correlation due to measurement error can be demonstrated using structural equation modeling. For instance, consider the following observed association:

```
cor(
  x = mydataValidity$predictorObservedSEM,
  y = mydataValidity$criterionObservedSEM,
  use = "pairwise.complete.obs")
```

```
[1] 0.6118
```

The observed association can be estimated in structural equation modeling in the `lavaan` package (Rosseel et al., 2022) using the following syntax:

TABLE 4.2 Parameter Estimates of Observed Association in Structural Equation Model.

lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper	std.lv	std.all	std.nox
criterionObservedSEM		predictorObservedSEM	0.633	0.027	23.366	0.000	0.580	0.686	0.633	0.611	0.611
predictorObservedSEM		predictorObservedSEM	0.971	0.044	21.932	0.000	0.885	1.058	0.971	1.000	1.000
criterionObservedSEM		criterionObservedSEM	0.654	0.030	21.518	0.000	0.595	0.714	0.654	0.627	0.627
criterionObservedSEM	1		−0.012	0.027	−0.459	0.646	−0.064	0.040	−0.012	−0.012	−0.012
predictorObservedSEM	1		0.012	0.032	0.373	0.709	−0.050	0.074	0.012	0.012	0.012

```
observedSEM_syntax <- '
  criterionObservedSEM ~ predictorObservedSEM

  # Specify residual errors (measurement error)
  predictorObservedSEM ~~ predictorObservedSEM
  criterionObservedSEM ~~ criterionObservedSEM
'

observedSEM_fit <- sem(
  observedSEM_syntax,
  data = mydataValidity,
  missing = "ML")
```

Parameter estimates from the model are shown in Table 4.2.

Now consider when we account for the degree of unreliability of each measure. We can account for the (un)reliability of each measure by specifying the residual errors as 1 − reliability, as below:

```
disattenuationSEM_syntax <-
  paste(
    '
    # Factor loadings
    predictorLatent =~ 1*predictorObservedSEM
    criterionLatent =~ 1*criterionObservedSEM

    # Factor correlation
    criterionLatent ~ predictorLatent

    # Specify residual errors (measurement error)
    predictorObservedSEM ~~ (1 - ',
      reliabilityOfPredictor, ')*predictorObservedSEM
    criterionObservedSEM ~~ (1 - ',
      reliabilityOfCriterion, ')*criterionObservedSEM
    ',
    sep = "")

disattenuationSEM_fit <- sem(
  disattenuationSEM_syntax,
  data = mydataValidity,
  missing = "ML")
```

TABLE 4.3 Parameter Estimates of Disattenuated Association in Structural Equation Model.

lhs	op	rhs	est	se	z	pvalue	ci.lower	ci.upper	std.lv	std.all	std.nox
predictorLatent	=	predictorObservedSEM	1.000	0.000	NA	NA	1.000	1.000	0.933	0.947	0.947
criterionLatent	=	criterionObservedSEM	1.000	0.000	NA	NA	1.000	1.000	0.946	0.925	0.925
criterionLatent		predictorLatent	0.706	0.030	23.152	0.000	0.646	0.766	0.697	0.697	0.697
predictorObservedSEM		predictorObservedSEM	0.100	0.000	NA	NA	0.100	0.100	0.100	0.103	0.103
criterionObservedSEM		criterionObservedSEM	0.150	0.000	NA	NA	0.150	0.150	0.150	0.144	0.144
predictorLatent		predictorLatent	0.871	0.044	19.674	0.000	0.785	0.958	1.000	1.000	1.000
criterionLatent		criterionLatent	0.460	0.031	14.960	0.000	0.399	0.520	0.514	0.514	0.514
predictorObservedSEM	1		0.012	0.032	0.373	0.709	−0.050	0.074	0.012	0.012	0.012
criterionObservedSEM	1		−0.005	0.033	−0.143	0.887	−0.069	0.060	−0.005	−0.005	−0.005
predictorLatent	1		0.000	0.000	NA	NA	0.000	0.000	0.000	0.000	0.000
criterionLatent	1		0.000	0.000	NA	NA	0.000	0.000	0.000	0.000	0.000

Parameter estimates from the model are presented in Table 4.3.

The observed association ($\beta = 0.61$) becomes $\beta = 0.70$ when it is disattenuated for measurement error.

4.7 Generalizability Theory

Generalizability theory (G-theory) is discussed in greater detail in the chapter on reliability in Section 3.11 and in the chapter on generalizability theory. As a brief reminder, G-theory is a measurement theory that, unlike classical test theory, does not treat all measurement differences across time, rater, or situation as "error" but rather as a phenomenon of interest. G-theory can simultaneously consider multiple aspects of reliability and validity in the same model, something that classical test theory cannot achieve.

4.8 Ways to Increase Validity

Here are potential ways to increase the validity of the interpretation of a measure's scores for a particular purpose:

- Make sure the measure's scores are reliable. For potential ways to increase the reliability of measurement, see Section 3.14. But do not switch to a less valid measure or to items that are less valid merely because they are more reliable.
- Use multiple measures and multiple methods to remedy the effects of method bias.
- Design the measure with a particular population and purpose in mind. When describing the measure in papers or in public spheres, make it clear to others what the population and intended purposes are and what they are not.
- Make sure each item's scores are valid, based on theory and empiricism, for the particular population and purpose. For instance, the items should show content validity—the items should assess facets of the target construct for the target population as defined by experts, without item intrusions from other constructs. The items' scores should show convergent validity—the items' scores should be related to other measures of the construct, within the population of interest. The items' scores should show discriminant validity—the items'

scores should be more strongly related to measures that are intended to assess the same construct than they are to measures that are intended to assess other constructs.
- Obtain samples that are as representative of the population as possible, paying attention to including people who are traditionally under-represented in research (if such groups are part of the target population).
- Make sure that people in the population can understand, interpret, and respond to each item in a meaningful and comparable way.
- Make sure the measure and its items are not biased against any subgroup within the population of interest. Test bias is discussed in Chapter 15.
- Be careful to administer the measure to the population of interest under the conditions in which it is designed. If the measure must be administered to people from a different population or under different conditions from which it was designed, be careful to (a) note that the measure was not designed to be administered for these other populations or conditions, and (b) note that individuals' scores may not accurately reflect their level on the construct. If interpretations are made based on these scores, make them cautiously and say how the differences in population or condition may have influenced the scores.
- Continue to monitor the validity of the measure's scores for the given population and purpose. The validity of measures' scores can change over time for a number of reasons. Cohort effects can lead items to become obsolete over time. If people or organizations change their behavior in response to a measure, this can invalidate a measure's scores for the intended purpose, as described in Section 4.3.1.9 when discussing consequential validity.

4.9 Conclusion

Validity is how much accuracy, utility, and meaningfulness the interpretation of a measure's scores have for a particular purpose. Like reliability, validity is not one thing. There are multiple aspects of validity. Validity is also not a characteristic that resides in a test. The validity of a measure's scores reflect an interaction of the properties of the test with the population for whom it is designed and the sample and context in which it is administered. Thus, when reporting validity in papers, it is important to adequately describe the aspects of validity that have been considered and the population, sample, and context in which the measure is assessed.

4.10 Suggested Readings

Cronbach & Meehl (1955); L. A. Clark & Watson (2019)

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com